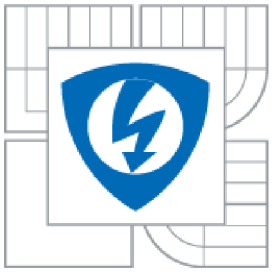
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF OF BIOMEDICAL ENGINEERING

POROVNÁVÁNÍ MITOCHONDRIÁLNÍ DNA PRO IDENTIFIKACI DRUHŮ

COMPARISON OF MITOCHONDRIAL DNA FOR SPECIES IDENTIFICATION

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE — RENÉ LABOUNEK
AUTHOR

VEDOUCÍ PRÁCE — ING. DENISA MADĚRÁNKOVÁ
SUPERVISOR

BRNO 2010

VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Student: René Labounek
Ročník: 3

ID: 106135
Akademický rok: 2009/10

NÁZEV TÉMATU:

Porovnávání mitochondriální DNA pro identifikaci druhů

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s metodou identifikace druhů pomocí porovnání mitochondriální DNA sekvence genu COI a vypracujte rešerši této metody. V Matlabu vytvořte s využitím grafického rozhraní program pro stažení požadované sekvence druhu z databáze NCBI a srovnání se stejným genem pro člověka. Dále tento program rozšiřte pro vytvoření databáze stažených sekvencí COI. Program bude také umět identifikovat druh organismu podle uživatelem zadané sekvence genu COI, bude umožňovat převod sekvencí do alespoň dvou různých numerických formátů a bude umožňovat analýzu podobnosti sekvencí pomocí distanční matice. Program vyzkoušejte na několika vybraných sekvencích a výsledky zhodnoťte.

DOPORUČENÁ LITERATURA:

[1] Stoeckle M.Y., Herbert, P.D.: Čárový kód života. Scientific American České vydání, 11/2008
[2] Chase, M. W., Fay, M. F.: Barcoding of Plants and Fungi. Science 325, pp. 682-683, 2009

Termín zadání: 8.2.2010 ***Termín odevzdání:*** 31.5.2010

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti bakalářské práce:

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

UPOZORNĚNÍ:

Abstrakt

Práce se zabývá metodou rozeznávání živočišných druhů na základě analýzy úseku mitochondriální DNA. K této analýze a zařazení se využívá úsek genu CO1 v literaturách nazýván jako čárový kód života. V úvodu práce je rozebrána teorie mitochondriální dědičnosti a podmínek tvorby čárového kódu. Z této teorie nadálé vychází její praktické využití při tvorbě knihovny vytvořených čárových kódů jednotlivých živočišných druhů. Data pro tvorbu knihovny jsou čerpány z veřejných databází NCBI a BOLD Systems. Další část práce se zabývá metodami porovnávání jednotlivých čárových kódů mezi sebou a hlavně s čárovým kódem člověka. K těmto analýzám byly použity tři hlavní výpočetní postupy. Byly to Needleman-Wunschův algoritmus, Smith-Watermanův algoritmus a porovnávání podobností pomocí distanční matice. S metodou porovnávání pomocí distanční matice je úzce spjat převod sekvencí molekuly DNA ze znakové podoby do numerických formátů, kterým se tato práce také zabývá. K usnadnění práce s daty byly vytvořeny vyhledávací algoritmy čárových kódu podle zadaného názvu druhu a naopak.

Abstract

The work deals with the method of recognizing species on the analysis of mitochondrial DNA segment. This analysis and classification using segment gene called CO1 in literatures such as barcode of life. In the beginning of work is analyzed the mitochondrial theory of heredity and conditions of formation of barcode. Practical use is based on this theory in creating database of barcodes generated to different animal species. Data used for creating the library are drawn from public databases NCBI and BOLD Systems. The next part of this work concerns about methods of comparison of the individual barcodes to the others and especially to the barcode of human. Three main computing methods were used tore these analyses: Needleman-Wunsch algorithm, Smith-Waterman algorithm and comparison of similarities using distance matrix. This work also concerns about transformation of DNA molecule sequences from symbols to numeric formats, which is required for the distance matrix comparison method. Algorithms for searching for a barcode of a species and vice versa were created to ease the work with data.

Klíčová slova

čárový kód, vlajka čárového kódu, mitochondrie, mitochondriální DNA, genom, gen CO1, NCBI, druh, jedinec, data, databáze, BOLD Systems, Needleman-Wusnchův algoritmus, Smith-Watermanův algoritmus, distanční matice, numerický formát

Keywords

barcode, barcode flag, mitochondria, mitochondrial DNA, genome, gene CO1, NCBI, species, specimen, data, database, BOLD Systems, Needleman-Wunsh algorithm, Smith-Waterman algorithm, distance matrix, numerical format

Bibliografická citace

LABOUNEK, R. *Porovnávání mitochondriální DNA pro identifikaci druhů*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2010. 54 s. Vedoucí bakalářské práce Ing. Denisa Maděránková.

Prohlášení

Prohlašuji, že svoji bakalářskou práci na téma Porovnávání mitochondriální DNA pro identifikaci druhů jsem vypracoval samostatně pod vedením vedoucí bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

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

V Brně dne 31. května 2010 .....

podpis autora

Poděkování

Děkuji vedoucí bakalářské práce Ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce. Zároveň bych také chtěl poděkovat svým rodičům za podporu při studiích.

V Brně dne 31. května 2010 .....

podpis autora

Obsah

Seznam ilustrací

1. Úvod

Porovnávání živočišných druhů na základě analýzy CO1 genu mitochondriální DNA je nová metoda katalogizace života na Zemi. Od dob vzniku klasické systematické biologie rozdělující druhy na základě morfologické stavby těla je tato metoda první, která začíná živočišné druhy systematicky uspořádávat na základě molekulární genetické predispozice. Jelikož se vhodný úsek DNA pro tuto metodu vyskytuje na mitochondriálním genomu, na počátku této práce je rozebrána mitochondrie, stručně jsou popsány její vlastnosti, funkce a je uvedena zmínka se o důvodu a příčině existence vlastní mitochondriální DNA. V další části práce je řečeno, co je to tvorba čárového kódu života, kdo tuto metodu vytvořil, jaké jsou na ni kladeny požadavky a jestli je tato metoda přesnější, než klasický morfologický popis.

V další části práce je vytvořen program získávající sekvence CO1 genu z veřejných databází NCBI a CBOL, tyto data se ukládají do strukturované databáze. Těchto sekvencí bude nashromážděno co možná nejvíce, od velkého počtu živočišných druhů. Dále budou uvedeny 2 možné příklady zarovnání stažených sekvencí se sekvencí člověka CO1 genu, k tomuto zarovnání bude použito Needleman-Wunschova a Smith-Watermanova algoritmu. Vytvoří se vyhledávač sekvencí, podle uživatelem zadaného druhu a naopak, převaděč sekvencí do dvou numerických formátů, které budou následně využity pro výpočet podobnosti dvou různých sekvencí pomocí tzv. distanční matice.

Po naprogramování veškerých potřebných výpočtových funkcí se začnou využívat v grafickém prostředí GUI a vytvoří se uživatelsky příjemné prostředí pro provádění výše popsaných analýz. Důvod tvorby tohoto prostředí je zřejmý, nejčastějšími uživateli těchto analýz jsou biologové, kteří odmítají přijmout do svého podvědomí cokoliv, co má v názvu příkazová řádka.

V závěru této práce vybereme několik živočišných druhů, na kterých se provede analýza všech vytvořených algoritmů a v závěru dojde k jejich zhodnocení případné přesnosti či nepřesnosti.

2. Historie záznamu života

Již od pradávných dob pravěku byli tamější lidé a naši genetičtí předkové nerozlučně spjati s přírodou a jejími zákony. Už v těchto dobách si uvědomovali, že nejsou jedinými obyvateli prostředí, ve kterém žili. Tak jako ostatní zvířata, hnáni pocitem hladu a pudem sebezáchovy, lovili naši předkové ostatní živočišné druhy a sbírali rostlinné plody s kořínky. A přes to vše, co měli se zbytkem své lovené kořisti společné, byli to právě tito lidé, kteří položili základy naší společné zaznamenané historie. I přes neznalost písma a čtení objevil homo sapiens sapiens (člověk dnešního typu) první rukou zaznamenaný dokument, malbu. A není se čemu divit, že většina záznamů, od člověka této doby, byly malby zvířat.

Obr. 1: Pravěká malba v jeskyni Lasaux na jihozápadě Francie (http://www.lovecpokladu.cz/home/lascaux-sixtinska-kaple-prehistorie-3300)

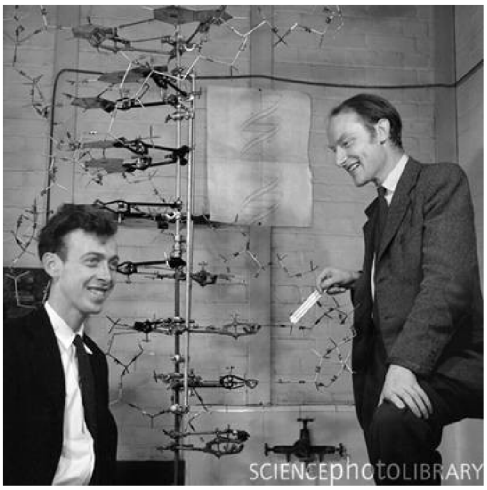
Poté následovalo období Starověku, který do lidských dějin vnáší vynález písma, a tudíž i nový pohled na svět. Ze starověkých análů víme například, že starověkými Egypťany byla kočka uctívané zvíře, že kráva je pro hinduisty posvátná již z těchto dob. Co do existence systematické biologie se ovšem nacházíme v období pravěku. Platón (427-347 před n. l.), jako jeden z prvních filozofů, formuloval v obecné podobě myšlenku nadřazenosti a podřazenosti tříd. Aristoteles (384-322 před n. l.) předpokládal, že drobní živočichové (roztoči, červi aj.) vznikají samoplozením. Tuto myšlenku vyvrátil až L. Pasteur (1882-1895).

Absolutní převrat ve studiu života, přinesl Carl Linné (1707-1778), který položil základy systematické biologii, tak jak ji známe dnes. Zavedl pojem druh a rozdělil organismy na říši živočišnou a říši rostlinnou. Zabýval se hlavně botanikou, napsal mnoho dodnes uznávaných děl. Snad nejznámější je Species plantarum (1753). Nutno podotknout, že Linné vytvořil systém řazení rostlinných a živočišných druhů, o kterém tvrdil, že je pevně daný a neměnný.

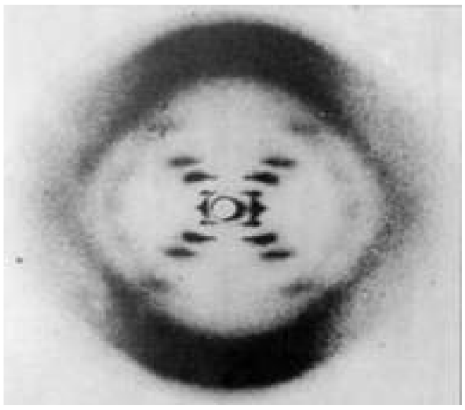
Tyto myšlenky byly postupem času vyvráceny. Nejprve J. B. Lamarckem (1744-1832), který řekl, že druhy byly stvořeny, ale mění se v závislosti na podmínkách prostředí. Po Lamarkovi následoval Charles Darwin se svojí teorií o evoluci a vývoji druhů.

I když se v tomto období biologové přeli o stálosti nebo proměnlivosti živočišné a rostlinné říše, měli obě tyto skupiny badatelů společný systém zařazování a popis druhů. Hlavním nástrojem pro katalogizaci každého druhu byl vnější popis jedince; např. počet končetin jedince, barva těla, velikost...

Doslova revoluce, z pohledu porozumění historii života na Zemi, přišla v polovině 20. století. Již od počátku tohoto století se vědci ptali, z jakého materiálu je vytvořena genetická informace. Ve 40. letech se prováděly výzkumy na jednoduchých houbách a zjistilo se, že genetická informace, nese instrukce převážně pro tvorbu proteinů. V tomto desetiletí byla také deoxyribonukleová kyselina označena za pravděpodobného nosiče genetické informace. Průlom v tomto bádání zaznamenali v roce 1953 pánové James Watson a Francis Crick, kteří, za pomoci rentgenové difrakce na krystalu molekuly DNA, vytvořili trojrozměrný model dvojšroubovice DNA.



Obr. 2: J. Watson, F. Crick a jejich 3D model dvojšroubovice DNA (http://apendix.bf.jcu.cz/Dolezal/vyuka/dna/DNA.htm)



Obr. 3: Obraz RTG difrakce na krystalu DNA (http://apendix.bf.jcu.cz/Dolezal/vyuka/dna/DNA.htm)

Dnes stojíme na počátku 21. století a náš pohled na systematickou biologii se mění od základů. Nemá smysl zastírat, že úvahy o zařazení živočichů a rostlin podle sekvence jejich jedinečné molekuly DNA nevznikly ihned po rozluštění její struktury. Trvalo to ovšem ještě dalších 50 let, než se tato myšlenka mohla začít uskutečňovat. V roce 2008 v říjnovém čísle časopisu Scientific American vyšel článek Čárový kód života (Barcode of Life) . Tento článek je od dvou autorů Marka Y. Stoecklea a Paula D. N. Herberta.

3. Mitochondrie a její dědičnost

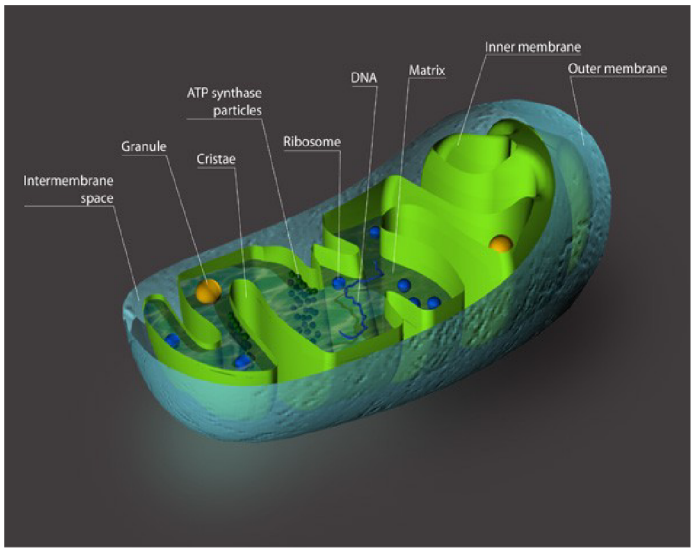
3.1 Mitochondrie

Jelikož metoda rozpoznávání živočišných druhů je úzce spjata s touto organelou, uvedeme o ní několik slov i v této práci. Její morfologie a funkce není pro naši práci moc podstatná, proto ji popisujeme jen ve zkratce.

Každá mitochondrie se skládá ze dvou membrán (vnější a vnitřní). Tyto membrány vymezují vnitřní prostor mitochondrie zvaný matrix a vnější mezimembránový prostor. Tyto membrány hrají ústřední roli v tvorbě ATP, která je pro heterotrofní organismus (živočichové a houby) jediným zdrojem energie. Jinými slovy mitochondrii můžeme chápat jako energetickou elektrárnu každé eukaryotní buňky.

Vnější membrána je tvořena dvojitou vrstvou lipidů (=> hydrofilní vlastnosti membrány), mezi nimž jsou proteiny zvané porin, které slouží jako vodné kanály. Díky těmto proteinům je membrána propustná pro všechny molekuly do velikosti 5000 daltonů.

Na vnitřní membráně dochází k průchodu elektronů uvolněných z NADH za vzniku NAD+. Tyto elektrony se pohybují podél elektrotransportního řetězce. Během tohoto pohybu se uvolní energie, která umožní přečerpání protonu H+. Vznikne protonový gradient na membráně, který při zpětném přečerpání, na základě koncentračního spádu, pohání oxidační fosforylaci (ADP + Pi vznikne ATP) Elektron, který se pohyboval membránou doputuje až k volnému kyslíku, kde společně s protonem H+ dá vznik molekule vody. Vnitřní plocha vnitřní membrány ohraničuje prostor zvaný matrix mitochondrie. Zde se nachází velice koncentrovaná směs enzymů, například enzymy účastnící se citrátového cyklu. Dále se zde nachází **několik identických kopií mitochondriální DNA**, mitochondriální ribosomy, tRNA a enzymy potřebné k translaci mitochondriální mRNA.

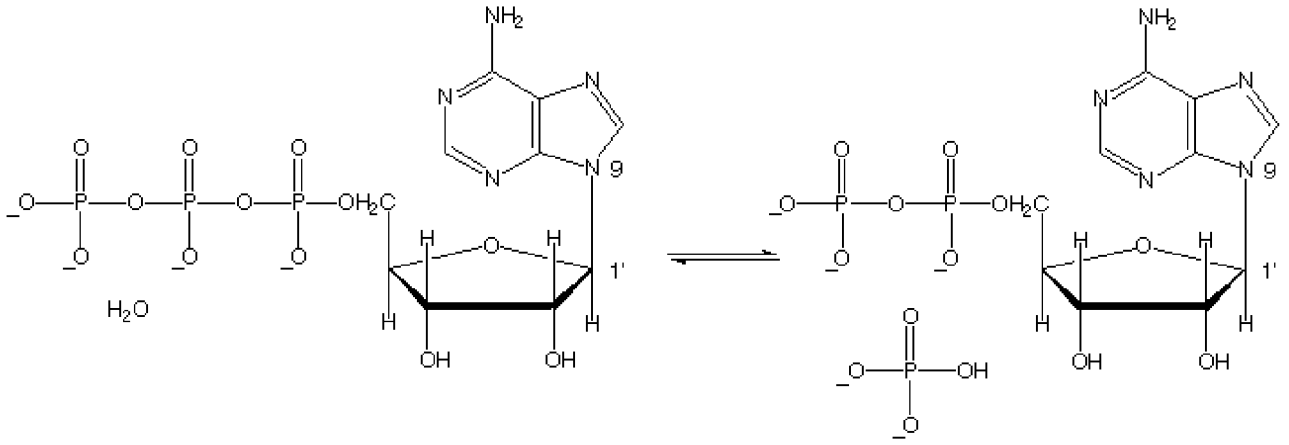


Obr. 4: Stavba mitochondrie
(http://mitolab.lf1.cuni.cz/homepage.php?stranka=MitoPor)

Na schématickém snímku je vidět, že struktura vnitřní membrány je velmi členitá. Důvodem, proč se tato membrána zvlnila a utvořila útvary zvané kristy, bylo zvětšení celkové plochy membrány. Jinými slovy zvětšení plochy, na kterém dochází k syntéze organismy využívané formy energie.

Mezimembránový prostor obsahuje enzymy využívající ATP vypuzenou z matrix mitochondrie k fosforylaci dalších nukleotidů. Fosforylace je enzymaticky řízená reakce, která zajišťuje vznik energeticky bohatých makromolekul v buňkách. Nejznámějším druhem této reakce je změna adenosindifosfátu na adenosintrifosfát (ADP ---> ATP), tato reakce nemusí ale probíhat pouze s adeninovou bází. Stejný mechanismus řídí i fosforylaci GDP na GTP, atd. Při těchto fosforylacích

dochází vždy je vzniku (ADP ---> ATP) nebo zániku (ATP ---> ADP) esterové vazby mezi fosfátovými skupinami.



Obr. 5: Reverzibilní defosforylace ATP na ADP
(http://tonga.usip.edu/gmoyna/biochem341/lecture36.html)

3.2 Mitochondriální dědičnost

Jak již z předchozího textu vyplývá, mitochondrie je, co do dědičnosti buňky částečně samostatná jednotka. Existence uložení genetické informace buňky mimo jádro byla předpokládána již na počátku 20. století. Mitochondriální DNA a její následná exprese byla prokázána v 60. letech 20. století. Tento objev potvrdil tzv. teorii endosymbiózy. Tato teorie předpokládá symbiotické splynutí předchůdců dnešních eukaryotních buněk s buňkami prokaryotickými. Původně předpokládanými samostatně žijícími prokaryotními organismy mají být mitochondrie, chloroplasty a ostatní plastidy. U všech těchto organel jsme zaznamenali existenci samostatné DNA. Konkrétně endosymbiotická teorie v případě mitochondrie předpokládá vniknutí původně aerobní bakterie do bakterie anaerobní (předchůdce eukaryot). Tato aerobní bakterie dala za vznik organele zvané protomitochondrie.

Platnost této teorie dokazuje i fakt, že mitochondrie připomíná prokaryotní organismus, jak svojí stavbou, tak i strukturou vlastní molekuly DNA. Zatímco jaderná DNA eukaryotních organismů je lineární dvou vláknová šroubovice se dvěma konci. Prokaryotní organismy, mitochondrie i plastidy mají molekulu DNA kruhovou. Tudíž tato molekula nemá žádný konec. Existují však i organismy, které mají mtDNA lineární, patří mezi ně: *Chlamydomonas reinhardtii* a *Paramecium aurelia*.

Tato molekula DNA podléhá genetické expresi stejně jako molekula jaderné DNA s několika málo rozdíly. Mitochondrie mají vlastní ribozomy, na kterých dochází k translaci mRNA na primární strukturu proteinů. Zatímco každá buňka má v jádře uloženou pouze jednu molekulu DNA, v jedné mitochondrii je molekul mtDNA hned několik. Zároveň nesmíme opomenout, že každá buňka obsahuje několik mitochondrií. Literatura udává, že jaterní buňky obratlovců obsahují řádově 1000-1500 mitochondrií na buňku. U člověka jsou tyto počty ještě mnohem vyšší řádově od 1000 do 10000 mitochondrií na buňku. Nejdůležitějším rozdílem mezi oběma procesy exprese genetické informace je hlavně rozdílné kódování některých aminokyselin, nebo-li všechny triplety kodonů nekódují stejnou aminokyselinu u obou druhů molekul.

První změnu zaznamenáváme u stop-kodonů. Zatímco v jaderné DNA triplety AGG a AGA nesou zakódovanou informaci pro translaci argininu, v mitochondriální DNA tyto kodony zastupují stop-kodony. Naproti tomu stop-kodon TGA u jaderné DNA, představuje tryptofan u mitochondriální. ATA je isoleucinem v jaderné DNA, zatímco v jazyce mitochondriální exprese tento triplet označuje start-kodon kódující methionin. Start-kodonem u jaderné DNA je triplet ATG, který je transkripován na AUG při přepisu do mRNA.

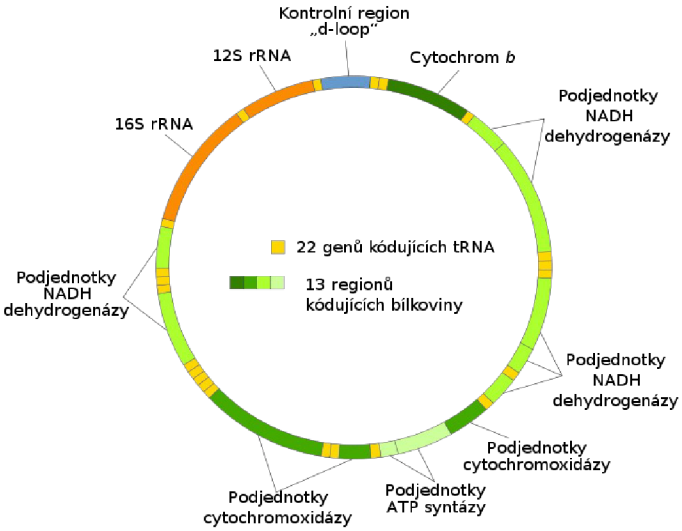
I když byly mitochondrie původně naprosto samostatnými organismy. V podobě, jak je známe dnes by již nebyly schopny samostatné existence. Obě molekuly DNA jsou velmi funkčně provázány a propojeny. Hlavním důvodem potřeby jaderné DNA pro mitochondrii je fakt, že v průběhu evoluce došlo ke značnému zkrácení mtDNA a uložení některých genů přímo do sekvence jaderné DNA. Naproti tomu i některé proteiny uložené v mtDNA řídí některé životně důležité pochody v cytoplazmě. Fylogeneze dvojité membrány vychází ze splynutí dvou původně samostatných organismů. Předpokládá se, že vnější membrána je pozůstatkem cytoplazmatické membrány původního anaerobního organismu. Zatímco vnitřní membrána byla vnější membránou předchůdce dnešní mitochondrie, který původně také žil jako samostatný jedinec.

3.3 Mitochondriální genom

První důležitou zmínkou o mitochondriálním genomu je jeho rozmanitost délky, velikost mitochondriálního genomu se velmi liší jak mezi jednotlivými kmeny, tak hlavně v rámci jednotlivých říší. Například průměrná délka mitochondriálního genomu u obratlovců je 16-17 kbp, naproti tomu u kvetoucích rostlin může obsahovat až 2500 kbp.

Živočišný mitochondriální genom je typický absencí intronů. Jeho sekvence obsahuje pouze úseky potřebné pro transkripci a translaci. Je to jeden z hlavních důvodů, proč je tak krátký. U obratlovců tento genom obsahuje 37 genů.

Lidský mitochondriální genom se skládá ze 16 659 párů bází obsahujících 37 genů. Ve dvou genech je uložena ribosomální RNA (= stavební jednotka mitochondriálních ribosomů). 22 genů kóduje strukturu 22 druhů tRNA a dalších 13 genů kóduje proteiny účastnících se oxidační fosforylace.



Obr. 6: Lidský mitochondriální genom (http://cs.wikipedia.org/wiki/Mitochondriální_DNA)

Délka mitochondriálních genomů bezobratlých organismů je přibližně stejná jako u obratlovců, ale má pár rozdílů v genetickém uspořádání. Tyto rozdíly jsou způsobeny rozdílnou strukturou uspořádání genu v rámci kruhové molekuly mtDNA.

Mitochondriální genom hub bývá delší než u zvířat, řádově se jeho délka pohybuje okolo 78 kbp. Tyto molekuly v sobě mají uložen přepis 33 genů (2 kódující ribosolmální RNA, 23-25 pro tRNA, 1 ukládá ribosomální protein, 7 popisuje proteiny účastnící se oxidativní fosforylace). Důvod, proč je tento genom delší než u zvířat, je přítomnost intronů. Introny jsou úseky sekvence DNA, které nenesou informaci o žádném genu.

Jako poslední zůstává **mitochondriální genom rostlin**. Rostliny mají nejdelší genom ze tří přírodních říší. Tato oblast genomů je ještě poměrně málo probádána. Nemůžeme zde hovořit o průměrné délce rostlinného mitochondriálního genomu, neboť např. mtDNA marchantie polymorphy (porostnice mnohotvárné) je 186 kbp dlouhá. Genom vodního melounu má 300 kbp a jak již bylo uvedeno na začátku podkapitoly, kvetoucí rostliny mohou dosahovat délek genomu až 2500 kbp. Největší část těchto sekvencí zabírají úseky nekódující žádný gen. Tyto oblasti DNA jsou prozatím velká neznámá. Jak u jaderné, tak u mitochondriální DNA, se zatím nepodařilo vysvětlit funkci intronů. Některé práce říkají, že nemají žádnou funkci, na druhé straně existuje velká skupina vědců a populace, kteří nevěří faktu, že by molekula DNA obsahovala nějaké zbytečné informace. Hlavně je-li vzata v úvahu vysoká přesnost, s jakou se DNA replikuje. Na markantní délce rostlinné mtDNA se pravděpodobně nejvíce podepsala schopnost rostlin křížit se mezi sebou i v rámci více druhů. Tuto vlastnost u žádné jiné říše nepozorujeme. Jak se ukáže dále, tento fakt nám bude činit obrovské potíže při genetické klasifikaci těchto druhů.

4. Čárový kód života (Barcode of Life)

4.1 Co je to Čárový kód života?

Čárový kód života je specifický úsek DNA jedinečný pro daný druh. Přečíst genom jakéhokoliv druhu dnešní technologie umožňují už několik let. Čtení celého genomu je nepraktické, neboť zpracování takového množství dat je časově velmi náročné, zdlouhavé a prozatím to zvládne jen specializovaný odborník.

Proto si badatel Paul D. N. Herbert položil otázku, zda by se dal použít i krátký úsek DNA. Po tomto úseku byly požadovány 3 základní vlastnosti:

- musí to být stejná část stejného genu pro všechny živočišné druhy
- musí být dostatečně dlouhý, aby byl pro každý druh jedinečný
- musí být dostatečně krátký, aby umožnil rychlé čtení

Pro tyto účely se ukázala jako ideální část genu uloženého v mitochondriích. Celý tento gen kóduje enzym 1 cytochrom-c-oxidázu, zkráceně CO1. Z tohoto genu je použito 648 bází ve směru jakoby od 5' konce.

Bohužel tento gen se ukázal být jako referenční pouze u živočichů. Při pokusu aplikace tohoto postupu na rostlinnou říši a na houby se ukázalo, že rozdílný vývoj genomů znemožňuje tuto analýzu. Druh je u zvířat a hub chápán jako skupina jedinců, kteří se mezi sebou mohou pářit. U rostlin je známo, že i některé rozdílné druhy se mezi sebou mohou křížit. Z tohoto důvodu genetické hranice mezi těmito druhy splývají a i jejich jednoznačné určení tudíž není možné.

Vrátíme-li se ale zpět k živočichům. Přednost mitochondriální DNA před jadernou měla 2 důvody:

- mtDNA má větší rozdíly v sekvenci párů bází mezi jednotlivými druhy
- mitochondrií je v buňce nespočet, tudíž i zastoupení mtDNA je v buňce větší než jaderné DNA => je snazší získat mtDNA než jadernou DNA

Z tohoto textu tedy vyplývá, že vytvoření čárového kódu života, je snaha přiřadit každému žijícímu živočišnému druhu na Zemi specifickou a unikátní značku, podle které by mohl být daný jedinec identifikován snadno a rychle kdekoliv v terénu.

4.2 Přesnost metody užití čárového kódu

Jelikož se prozatím zabýváme pouze teorií této metody a nemáme prozatím v rukou žádná vlastní naměřená data, ze kterých bychom mohli přímo vycházet, jediným důkazem vysoké specificity této metody jsou prozatím jen závěry přímo od autorů článku Čárového kódu. Z tohoto důvodu jsou přiloženy k této práci 2 následující citace přímo z článku publikovaného v časopise Scientific American v říjnu 2008 v anglickém vydaní a v listopadu 2008 v českém vydání.

„U primátů má například každá buňka asi 3,5 miliardy párů bází. Čárový kód CO1 představuje úsek o délce pouhých 648 párů bází, ale příklady od lidí, šimpanzů a dalších velkých opic nesou dost rozdílů pro spolehlivé rozlišení jednotlivých skupin. Lidé se v úseku čárového kódu navzájem liší jedním nebo dvěma páry bází, ale od svých nejbližších příbuzných, šimpanzů, se lišíme zhruba v 60 místech a od goril asi v 70 místech.“

Druhá citace popisuje shrnutí výsledků, ve kterých autoři dokazovali přesnost a neomylnost této metody.

„Abychom dokázali, že tato malá DNA-značka může skutečně identifikovat druh, testovali jsme spolu s kolegy účinnost čárového kódu CO1 v různých živočišných skupinách ze souše i z moře, od pólů k tropům. Zjistili jsme, že samy CO1 čárové kódy rozlišují asi 98% druhů poznaných předchozím taxonomickým studiem. Ve zbytku zužují určení na páry nebo malé skupiny příbuzných druhů, obecně linie, které se teprve nedávno rozdělily, nebo druhy, které se pravidelně kříží.“

4.3 DNA Barcode Data

Jak bylo uvedeno na konci kapitoly 4.1, snaha vědců je vytvořit prostředek, pomocí kterého budou moci lidé rozeznávat živočišné druhy přímo v terénu a ne jen v laboratoři. Finální představa spočívá v sestrojení přenosného přístroje, který bude schopen přečíst DNA barcode (neboli DNA čárový kód) z nepatrného kousku získané tkáně. K docílení tohoto výsledku nám ovšem stále stojí v cestě několik překážek.

V první řadě doposud neexistuje kompletní databáze čárových kódů všech známých živočišných druhů. Od chvíle, kdy čárový kód života spatřil světlo světa, začala spousta institucí a lidí zabývajících se informatikou a biologií spoluvytvářet **celosvětovou referenční knihovnu** DNA čárových kódů. Touto společnou prací vzniká v tomto oboru nezměrné množství dat, která musí být naměřena, zpracována a uložena. Aby tyto data mohly sloužit k dalšímu zpracování, utvořil se soubor pravidel pro formát, který tyto data musí před uložením do veřejných databází splňovat, tzv. Barcode Data Standards.

4.3.1 Barcode Data Standards

The Consortium for the Barcode of Life (CBOL) uvtvořilo Database Working Group (DWG). Uspořádali zahajovací shromáždění v květnu roku 2004. Zde projednávali vytvoření referenční knihovny DNA sekvencí čárového kódu, která má být začleněna do souboru ostatních znalostí biologické různorodosti (např. s databází vzorků, druhem,biogeografickými informacemi). DWG doporučila ukládání těchto sekvencí na veřejně přístupnou doménu, upřednostnila International Nucleotide Sequence Database Collaboration (INSDC). DWG také doporučila propojení uložených sekvencí s voucher jedinci (vzorky z muzeí) a jejich platnými druhovými jmény.

V září roku 2004 DWG svolala další shromáždění, tentokrát pořádané skupinou GenBank a National Center for Biotechnology Information (NCBI) . Tyto skupiny předložily INSDC nové požadavky na standard pro tyto data, které by měly být použity. V dubnu roku 2005 DWG konzultovala všechny předložené návrhy s hlavními taxonomickými společnostmi a protřídila návrh standardu podle jejich požadavků a připomínek. V květnu roku 2005 GenBank prezentovala výsledný návrh standardu na každoroční konferenci INSDC, kde byl návrh přijat všemi stranami se všeobecným ohlasem. DWG se následně sešla s představiteli hlavních muzeí, projednat realizaci Barcode Data Standardu, i zde byla přijata bez výhrad.

Utvořený standard má tři části:

1. Vytvoření vyhrazeného klíčového slova: **BARCODE** (čárový kód). NCBI a jejich spolupracovníci přidají značku Barcode flag (vlajka), nově odevzdané sekvenci, která bude splňovat předepsané podmínky konzultované s CBOLem. Data, která splňují tato kritéria, budou známa jako BARCODE records v INDSC (= BRIs; nahrávky čárových kódů v INDSC)

2. **Požadované prvky dat musí:**

a) obsahovat odkaz na voucher jedince za použití strukturovaného pole, definovaného skupinami CBOL a NCBI, a na metadata spojená s jedincem, obsažená ve veřejné databázi úložiště jedince (úložiště – muzeum, voucher jedinec)

b) obsahovat odkaz na doložený název druhu, založeného v jednom ze zdrojů specifikovaného skupinami CBOL a NCBI.

c) obsahovat znak země, používající kontrolovanou slovní zásobu užívanou GenBankou

d) původ oblasti genu schválenou CBOLem jako skutečný čárový kód. Nejprve je pouze 1-cytochrom-c-oxidáza schválena jako oblast čárového kódu. Tato oblast je definována poměrně k myšímu mitochondriálnímu genomu jako 648 párů bází, které začínají na 58. a končí na 705. pozici.

e) obsahovat alespoň 500 přiléhajících jednoznačných párů bází v obousměrném řazení uvnitř schválené oblasti čárového kódu. Pokud GenBank zažádá o udělení Barcode flag (= uznání sekvence za čárový kód) u kratší sekvence než 500 kb, následné směrnice odkazují na uznání CBOL.

f) neobsahovat více než 1% nejednoznačných míst pro úplnou předloženou sekvenci

g) obsahovat jméno použité genové oblasti

h) být spojeny s trace files předložené NCBI Trace Archivu nebo Ensembl Trace Serveru

i) obsahovat sekvence všech použitých forward a reverse primerů. Pro nahrávky, ve kterých byly sousedící sekvence sestaveny více než jednou **amplicon** nebo kdy byla použita směs mnohonásobných primerů pro amplification (zesílení), musí být zaznamenány všechny mnohonásobné sety párů primerů. V závěru, zaznamenání názvů forward i reverse primerů spolu se sekvencemi primerů je silně doporučováno.

Silně doporučované prvky uložených dat:

(následující prvky dat byly přidány do INSDC na žádost CBOlu pro ověření správnosti **voucher** jedince, jsou sice silně doporučována, ale ne požadována)

j) zeměpisná šířka a zeměpisná délka

k) jméno osoby, která objevila příslušný druh

l) jméno osoby, která poskládala informace k udělení čárového kódu

m) datum uložení sbírky dat

3. **Řídící pravidla**

ISDNC poskytuje archiv nahrávek, které mohou být měněny pouze autorem. Pro označení či odznačení nahrávky značkou Barcode flag (flag = vlajka) platí následující pravidla:

a) CBOL definuje okolnosti, za kterých bude moci být značkou Barcode flag označena sekvence kratší než 500 kb. Může se jednat o sekvence vzorových jedinců nebo vyhynulých či velmi vzácných jedinců daného druhu.

b) CBOL je zodpovědný za proces založení, provedení a nabízení úseku genu, jímž skupiny badatelů navrhnou a ospravedlní za oblast čárového kódu jiný úsek genu než oblast na CO1 genu, tím pádem dostane tato část genu značku Barcode flag.

c) Sekvence, které splňují podmínky nahrávky čárový kód a byly předloženy z University of Guelph's Barcode of Life Database (BoLD) databázi GenBank, budou považovány GenBankou za předložené společně od jednotlivého badatele a od BoLDu zároveň. Tyto nahrávky mohou být upravovány každou stranou.

d) Sekvence, které splňují podmínky nahrávky čárového kódu a byly předloženy GenBance od jednotlivých výzkumníků, mohou být upravovány pouze osobou, která je zde uložila. Na doporučení CBOLu Genbanka odebere všechny Barcode flags přesně těm sekvencím, u kterých to CBOL doporučí. Tyto nahrávky budou v GenBance označeny jako non-Barcode nahrávky

e) DWG a NCBI vyvinou pro CBOL návrh protokolu pro připojování komentářů jejich kritiky a návrhů oprav sekvencí s Barcode flag od třetí strany uživatelů.

Barcode Data Standards, ze kterých jsme v této práci čerpali byly vydány 19. prosince roku 2005.

Při seznamování se s databázemi obsahující Barcode data zjistíme, že ve většině případů tyto záznamy nejen že obsahují všechny výše popsané prvky, ale také je k záznamu přidána ještě fotografie jedince ve vysokém rozlišení.

4.3.2 Získávání dat

Tvorba kompletní databáze života na Zemi na základě analýzy čtyř opakujících se chemických sloučenin byla rozdělena na několik etap. Jako první bylo zahájení tvorby tzv. celosvětové referenční knihovny. V muzeích po celém světě jsou uloženy sta miliony popsaných jedinců (v anglické literatuře označovaných jako **voucher specimens**). Tyto vzorky jsou analyzovány a sestavuje se z nich již výše uvedená referenční knihovna vzorků. I přes, na první pohled, jednoduchost této prvotní práce, i tuto etapu provází 2 zásadní úskalí. Prvním z nich je fakt, že spousta uložených vzorků byla před uložením zpracována tak, že získat z nich dnes sekvenci mtDNA není možné. Jinak řečeno v době uložení se s DNA analýzou nepočítalo, popřípadě DNA nebyla v té době vůbec známa. Druhým problémem zůstává obrovské stáří některých vzorků. Ve většině případů z těchto vzorků dnes již není možné získat kompletní sekvenci ať už DNA, tak i jen úsek CO1 mtDNA. Z tohoto důvodu museli CBOL a ostatní výše uvedené společnosti ustanovit výjimky, které povolují získání značky Barcode flag i sekvencím kratším než 500 bp. V běžné praxi tyto vzorky mají obvykle pouze jen 100-200 bp. Jelikož musíme počítat i s mírnou rozmanitostí sekvencí v rámci jednoho druhu, měl by každý referenční jedinec obsahovat alespoň 10 záznamů od 10 jedinců stejného druhu.

V momentě, kdy je referenční knihovna kompletní, může začít proces analýzy živých živočichů. K této analýze nám bude postačovat část tkáně jedince, kterého chceme zkoumat (vlas, chlup, část kůže, svalové tkáně, orgánu...). Z tkáně, kterou jsme získali musíme nejprve izolovat mtDNA a poté namnožit požadovaný úsek genu pomocí PCR amplifikace.

Tato analýza může poskytnout dva různé výsledky. V prvním případě získáme sekvenci CO1 mtDNA, která se shoduje s některou ze sekvencí jedince v referenční knihovně. To znamená, že jsme zjistili výskyt již známého druhu v lokalitě, kde jsme náš analyzovaný vzorek získali. Referenční knihovnu tvoří tři celosvětové databáze (GenBank, EMBL a DDBJ) . Je zde uložena analyzovaná sekvence, informace o referenčním (voucher) jedinci a jméno zkoumaného druhu.

Při této analýze může také nastat situace, kdy se námi analyzovaný vzorek nebude shodovat se žádnou ze sekvencí uloženou v referenční knihovně. V tomto případě můžeme hovořit o chvíli, kdy jsme objevili nový doposud nezařazený živočišný druh.

Pozn.: Proces analýzy nového vzorku a porovnání jej s referenční knihovnou dnes trvá několik hodin a stojí méně než dva dolary. V následujících letech se předpokládá, že by se tento proces měl zrychlit na pár minut a cena analýzy by měla být pouze několik penny.

4.4 Morfologický nebo molekulární popis?

Během posledních 250 let, kdy Carl Linné položil základy klasické systematické biologie, se podařilo na základě morfologie popsat přibližně 1,7 milionu druhů rostlin, živočichů a mikrobů. Proč bychom tedy měli zavádět novou formu popisu druhů? Když toto zavedení nebude bezesporu nic levného a snadného.

Důvodů je hned několik. První komplikace, která nastává při klasickém morfologickém popisu, se objevuje při výzkumu velmi podobných druhů, jejich odlišení od sebe je tak obtížné, že specialistu biologa zaměstná na celý život pouze studium těchto příbuzných druhů. Ještě obtížnější určení druhu nastává ve chvíli, máme-li v rukou pouze nevylíhnuté vajíčko, popřípadě část těla jedince.

Dále z morfologického hlediska na jedné straně sice poznáme na základě počtu končetin, barvy povrchu těla, velikosti... o jaký druh se jedná. Na straně druhé nám tento popis ve většině případů neříká v podstatě nic o historii tohoto druhu, jeho předcích ani zda náhodou neexistuje nový druh, který se z námi zkoumaného už mohl vyvinout. Tímto studiem se zabývá mnoho biologů a genetiků, kteří se snaží sestavit tzv. pomyslný strom života.

Studia na základě analýzy mtDNA již dnes odhalila spousty skrytých druhů mezi zástupci jedinců považovaných za jeden jediný druh.

Největší urychlení by ale bezesporu metoda studia, živočišných druhů na základě rychlé analýzy mtDNA zaznamenala při hledání ještě neobjevených druhů. I když se prozatím podařilo popsat téměř 2 miliony druhů, předpokládá se, že na naší planetě žije přibližně okolo 10 milionů druhů.

Po celém světě vzniklo již několik aktivních studií, v rámci zaměření na jednotlivou živočišnou třídu (např. All Birds Barcoding Initiative (ABBI), Fish Barcode of Life (FISH-BOL), Mosquito Barcoding Initiative (MBI) a spousta dalších). Každá z těchto studií má za cíl utvořit kompletní databázi dané živočišné třídy a na této třídě provést řadu studií. Například od MBI se očekává kromě kompletní databáze komárů také databáze všech nemocí, které tito živočichové přenášejí. Díky této databázi by také mohlo být možné navrhnout velmi efektivní insekticid, proti těmto živočichům.

Další praktické využití může čárový kód přinést i do běžné populace konzumentů. V momentě, kdy budeme schopni sestrojit rychlou, skladnou a levnou čtečku čárových kódů, nebudeme už nikdy váhat, že si kupujeme právě ten druh masa, který chceme. Další informaci, kterou se budeme schopni o tomto mase dozvědět, bude, zda je zdravé nebo či je domovem zdraví ohrožujících škůdců.

5. MATLAB - Naprogramované funkce

V této kapitole budou uvedeny a popsány jednotlivé funkce, které bylo nutno naprogramovat ke kompletnímu splnění zadání bakalářské práce. Od získávání dat ze dvou veřejných databází, až po srovnání druhů různými metodami zpracování.

Na úvod bychom měli podotknout, že veškeré kódy byly napsány na verzi programovacího jazyku MATLAB 7.8.0 (R2009a), který byl nainstalován na distribuci operačního systému Ubuntu 9.10 - Karmic Koala, což je jedna z mnoha distribucí operačního systému Linux. Pokud budete chtít, aby náš program pro stahování a analýzu čárových kódů pracoval na vašem počítači plnohodnotně, budete muset mít také nainstalovaný bioinformatický toolbox, který naše funkce využívají.

5.1 Získávání čárových kódů z veřejné databáze NCBI

Jedním z výstupů této práce by měla být alespoň částečná knihovna obsahující sekvence mitochondriální DNA genu CO1, která by měla splňovat alespoň část podmínek pro udělení statusu čárového kódu. Existuje několik cest, odkud tyto sekvence můžeme čerpat. Například z přímé analýzy části tkáně živočicha, z analýzy vzorků živočichů uchovaných v muzeích a nebo jako v našem případě z veřejných databází uchovávajících již analyzované genomy živočišné říše. Jednou z těchto databází je databáze NCBI, se kterou umí programovací prostředí Matlab poměrně dobře a jednoduše pracovat.

Naše řešení se bude skládat ze tří základních částí. První částí bude funkce `barcode.m`, která z databáze NCBI stáhne, uloží a podle našich potřeb zpracuje kompletní mitochondriální genom živočicha, kterého budeme chtít zkoumat. Druhou částí bude funkce `knihovna.m`, kterou jsou získané informace zpracovány, roztříděny a uloženy do navržené knihovny. Třetí a poslední částí tohoto bloku programu bude samotná knihovna, do které se budou ukládat libovolná množství živočišných druhů, společně s jejich úseky genu CO1 a dalšími blíže konkretizujícími informacemi příslušící každé uložené sekvenci a druhu.

5.1.1 Barcode.m

```
function vystup = barcode (nazev_sekvence,druh)
```

Do této funkce vstupují 2 vstupy. Prvním z nich je `nazev_sekvence`, jedná se o zkratkovité označení jednoho konkrétního mitochondriálního genomu druhu, pod kterým je genom uložen v databázi NCBI. Druhým vstupem funkce je proměnná druh, ta označuje název analyzovaného druhu. Pro názornost, uvádíme následující příklad:

Analyzovali bychom živočišný druh, uložený na této adrese: http://www.ncbi.nlm.nih.gov/nuccore/GU320192.1?ordinalpos=1&itool=EntrezSystem2.PEntrez.Sequence.Sequence_ResultsPanel.Sequence_RVDocSum

Do příkazové řádky bychom jako proměnnou nazev_sekvence uvedli 'GU320192' a druh by byl 'homo_sapiens'. Data, která jsou u tohoto genomu v databázi NCBI uložená se poté stáhnou do textového souboru: homo_sapiens_ GU320192.txt.

Jak je ze zdrojového kódu této funkce vidět, stažení informací z NCBI na pevný disk není nijak složité. Matlab ve svém bioinformatickém toolboxu obsahuje jednoduchou funkci `getgenbank`, která tuto práci udělá za uživatele. Stažená data se uloží do struktury, ze které následně dolujeme a ukládáme informace o druhu do výstupu funkce.

Po stažení mitochondriálního genomu na pevný disk, se v uložených informacích najde gen CO1. Postupný seznam genů uložených v mitochondriální DNA zkoumaného druhu je uložen v matici,

která se ve struktuře nachází pod proměnnou `data.CDS(1,b).gene`. Nejrychlejší a nejúspornější postup hledání genu CO1 v této matici je pomocí cyklu `while`, který porovnává hledaný gen s genem proměnné. V momentě, kdy objeví požadovaný textový řetězec, ukončí vyhledávání a již nepokračuje dál. Gen CO1 má na NCBI různá označení (např. CO1, COX1,...), proto se muselo ve vyhledávacím cyklu uvést více podmínek pro hledaný textový řetězec. V momentě, kdy je známa pozice genu CO1 v matici `data.CDS(1,b).gene`, je známa i pozice vektoru pořadí počáteční a koncové báze nesoucí tento gen v genomu. Pozice první a poslední báze genu CO1 v genomu je uložena pod proměnnou `data.CDS(1,b).indices`. Tento vektor si funkce ukládá do proměnné `poloha`. Rozdílem těchto dvou hodnost dostaneme délku genu CO1.

Když už je známa pozice genu v genomu, i jeho délka, stačí jeho sekvenci přizpůsobit podmínkám, které definují čárový kód v kapitole 4.3.1. Říkají, že čárovým kódem může být sekvence genu CO1, která začíná na 58. pozici tohoto genu a končí na 705. pozici. Dále se v této kapitole hovoří o podmínkách udělení statusu čárového kódu i sekvencím kratším než 500 párů bází. V tomto případě se ovšem musí žádat CBOL o udělení výjimky a v našem programu je na takovouto sekvenci upozorněno souvětím, které se zavolá funkcí `disp` a zobrazí se v příkazové řádce ve chvíli zpracovávání této informace počítačem.

Některé genomy mají prohozenou počáteční a koncovou souřadnici genu u proměnných `data.CDS.indices`. V tomto případě vychází délka genu záporná a musela být z tohoto důvodu rozšířena podmínka při tvorbě sekvence čárového kódu.

V tuto chvíli je již vytvořena sekvence čárového kódu a v posledním kroku funkce se vytváří struktura do proměnné výstup, kde se uloží informace o jednom zkoumaném druhu, která potom bude uložena v knihovně všech analyzovaných živočišných druhů. Konkrétně to tedy jsou název druhu, jeho sekvence čárového kódu, taxonomická klasifikace druhu a NCBI kód, což je defakto vstup funkce `nazev_sekvence`, pomocí kterého je každá uložená sekvence čárového kódu jedinečně specifikována a popřípadě i zpětně dohledávána v databázi NCBI.

5.1.2 Knihovna.m

Tato funkce má dva vstupy a tři výstupy:

```
function [struktura,jedinec,poznamka] = knihovna(druh,q)
```

Proměnná `druh` je textový řetězec, který vyjadřuje název hledaného živočišného druhu a proměnná `q` je NCBI kód, který jedinečně specifikuje danou nahrávku druhu a jeho genomu ve veřejné databázi NCBI. Ve výstupu `struktura` je uložena aktuální databáze živočišných druhů i s jejich čárovými kódy a daty druhy blíže specifikující. `Jedinec` nese informace z databáze pouze o jednom konkrétním druhu, kterého jsme zadali ke zkoumání na vstupu, popřípadě nese informaci o faktu, že žádný nový druh nebyl do databáze přidán. `Poznamka` nese binární informaci pro následující grafickou nástavbu a určuje jí, co má v dané situaci dělat.

I když se zdá systém být plně automatický, přesto uživatel musí do provozu vytváření databáze zapojit vlastní iniciativu a to vyhledáním vstupních dat o druhu. Chce-li přidat do databáze sekvenci konkrétního druhu, musí na stránkách databáze NCBI (http://www.ncbi.nlm.nih.gov/) zadat do vyhledávače latinský název hledaného druhu, nejlépe ve spojení s hesly mitochondrion + complete + genome, v kolonce Search změnit z All Databases na Nucleotide a dát vyhledávat. Vyhledávač mu zjistí, zda se daný druh v databázi nachází a zda databáze obsahuje nahrávku jeho mitochondriálního genomu. Pokud ano, objeví se nám hned na druhé řádce pod názvem druhu a názvem dané sekvence NCBI kód, který zadáme za vstupní proměnnou `q`.

Princip funkce `knihovna.m` je prostý. Nejprve zazálohuje stávající uloženou strukturu s již načerpanými daty. Poté stáhne a zpracuje data příslušného druhu pomocí funkce `barcode.m` popsané v předchozí podkapitole.

Má-li funkce `barcode.m` stažena a zpracována všechna data, nastává vyhledávací a porovnávací algoritmus staré databáze s právě staženými daty. Zjištění existence totožné sekvence u totožného druhu ve staré databázi je realizováno paralelním spojením dvou `while` cyklů a sérií podmínek mezi nimi.

První podmínka ošetřuje situaci, kdy máme prázdnou kostru knihovny, počet druhů v ní je nula a je tedy zapotřebí uložit všechna data, která byly pomocí funkce `barcode.m` stažena.

Druhá podmínka porovnává délky názvů druhů, nejsou-li délky stejné, nejsou ani názvy stejné a zkoumaný druh ze staré databáze se tedy neshoduje s druhem staženým. Cyklus tedy jen přeskočí na druh ze staré databáze s indexem o jedna větším. V následujících 2 podmínkách se délka názvů druhů už bude vždy shodovat.

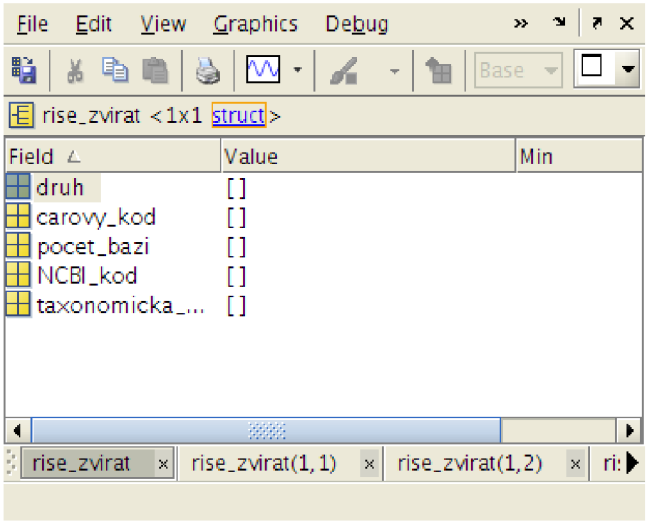
V případě, kdy se shoduje i název, spustí se druhý cyklus `while`, který prohledává uložené NCBI kódy ve staré databázi a porovnává je se staženým NCBI kódem. Shodují-li se v některém případě, znamená to, že právě stažená sekvence se v databázi již nachází, ukončí se oba cykly a skript uloží starou nepozměněnou databázi. Neshodují-li se, zvětší se proměnná `m` o jedničku a porovnává se další NCBI kód. Pokud se ani poslední NCBI kód neshoduje, hodnota `m` naroste na hodnotu `e+1`, ukončí se cyklus, do databáze se uloží nová sekvence s novým NCBI kódem, počtem párů bází u nově uložené sekvence a ukončí se i nadřazený `while` cyklus.

Pokud se neshodují stejně dlouhé vektory názvů druhů, zvětší se hodnota n o jedna a začne se porovnávat další druh uložený ve staré databázi se staženým druhem. V případě, že se neshoduje ani poslední název druhu s právě staženým druhem, stáhli jsme druh, který se v databázi ještě nevyskytuje. V tomto případě platí, že `n = a+1`. Struktura databáze o rozměrech `1xa` se o zvětší na strukturu o rozměrech `1x(a+1)` a na pozici `1x(a+1)` se uloží všechny informace, které v databázi u každého druhu požadujeme.

Na konci celé funkce se nová nebo nepozměněná databáze uloží do souboru `carovy_kod.mat`. V tomto souboru je uložena databáze čárových kódů živočišných druhů, která je uložena pod proměnnou `rise_zvirat`.

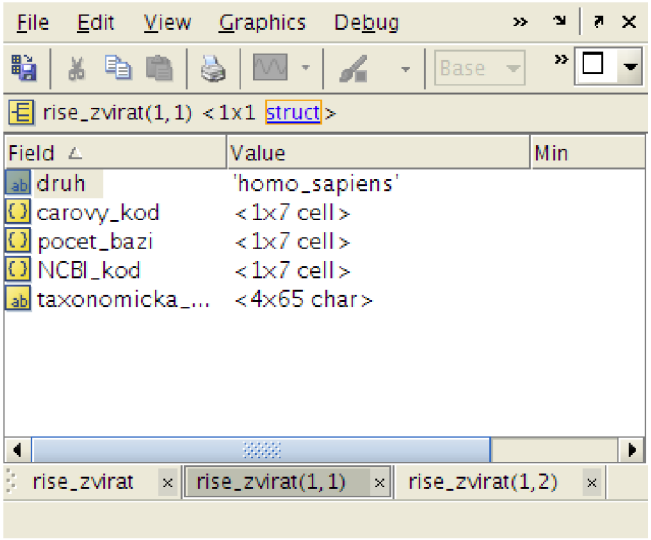
5.1.3 Carovy_kod.mat

Tento .mat soubor je vytvořená knihovna stažených živočišných druhů, sekvencí čárových kódů a dalších informací blíže specifikující daný druh. Než se začnou do knihovny ukládat data, musela být navržena základní struktura databáze. V programovacím prostředí matlab bylo využito příkazů `struct` a `cell`. Prázdná knihovna vypadala následovně:



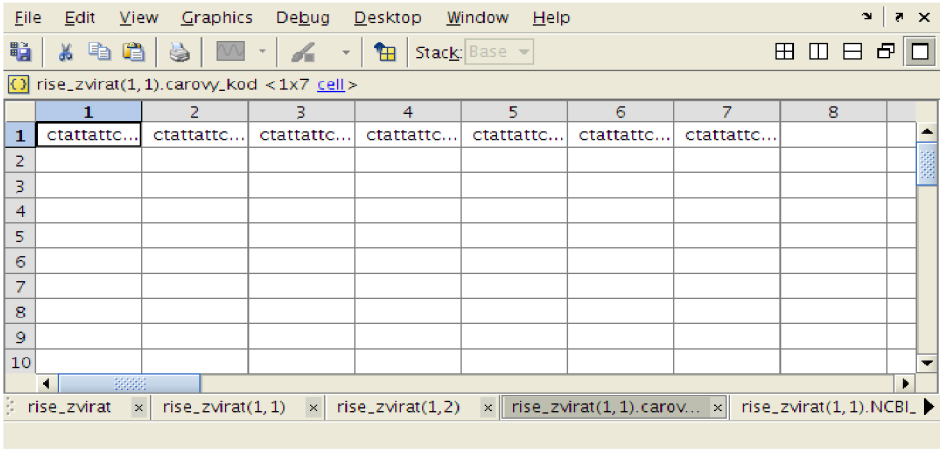
Obr. 7: Prázdná knihovna

Jako první bylo do knihovny uloženo několik sekvencí čárových kódů člověka. I když prozatím jsou všechny hodnoty u jednotlivých proměnných struktury prázdné vektory, ihned při uložení prvního jedince se v databázi objeví prvky struktury typu `cell`.

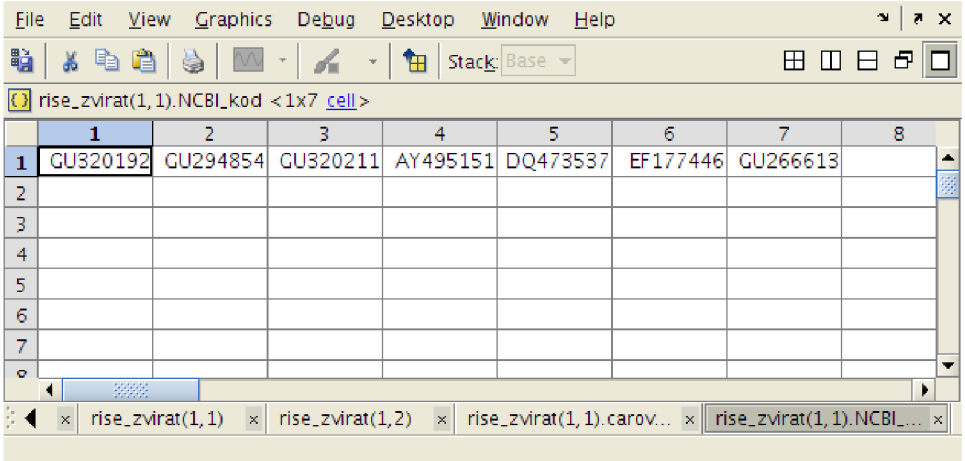


Obr. 8: Člověk rozumný se sedmi čárovými kódy

Proměnná `carovy_kod` obsahuje sedm sekvencí čárových kódů, proměnná `pocet_bazi` délky sedmi sekvencí čárových kódů a `NCBI_kod` odlišuje každou sekvenci a jedinečně ji typizuje.

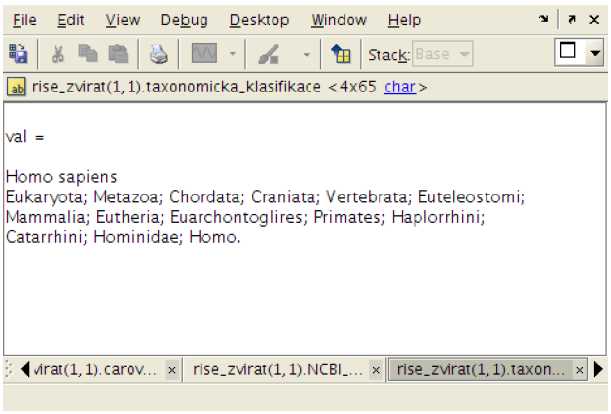


Obr. 9: Sedm sekvencí čárových kódů člověka rozumného v buňce cell



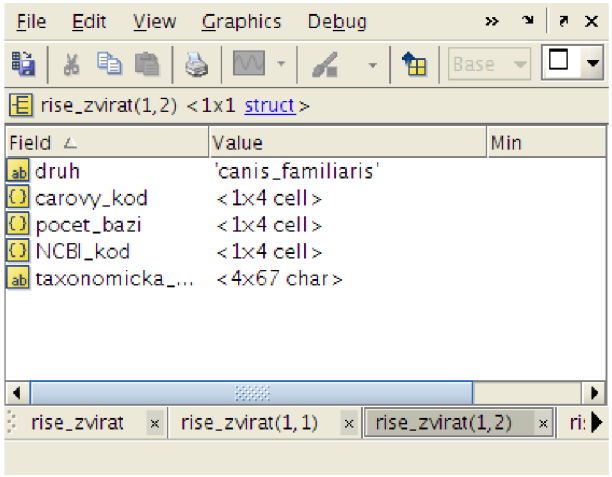
Obr. 10: Sedm jedinečných NCBI kódů pro sekvence člověka rozumného

Pro úplnou názornost ještě uvádíme obsah proměnné `taxonomicka_klasifikace`:



Obr. 11: Taxonomická klasifikace člověka rozumného čerpaná z NCBI

Jelikož byl člověk rozumný prvním druhem, kterého jsme do databáze uložili, rozměr základní struktury byl 1x1, po přidání dalšího druhu se rozměr struktury o jedna zvětší a její rozměr bude 1x2.

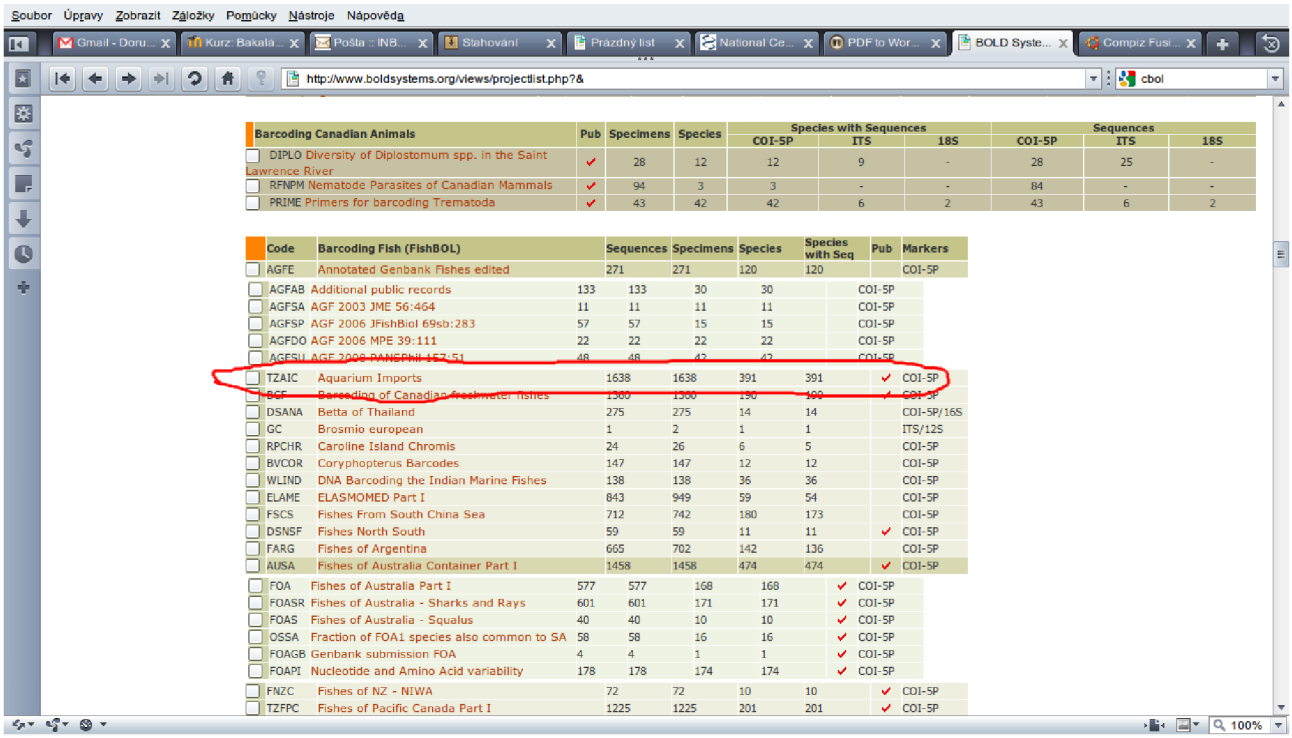


Obr. 12: Pes domácí

5.2 Získávání čárových kódů z veřejné databáze BOLD Systems

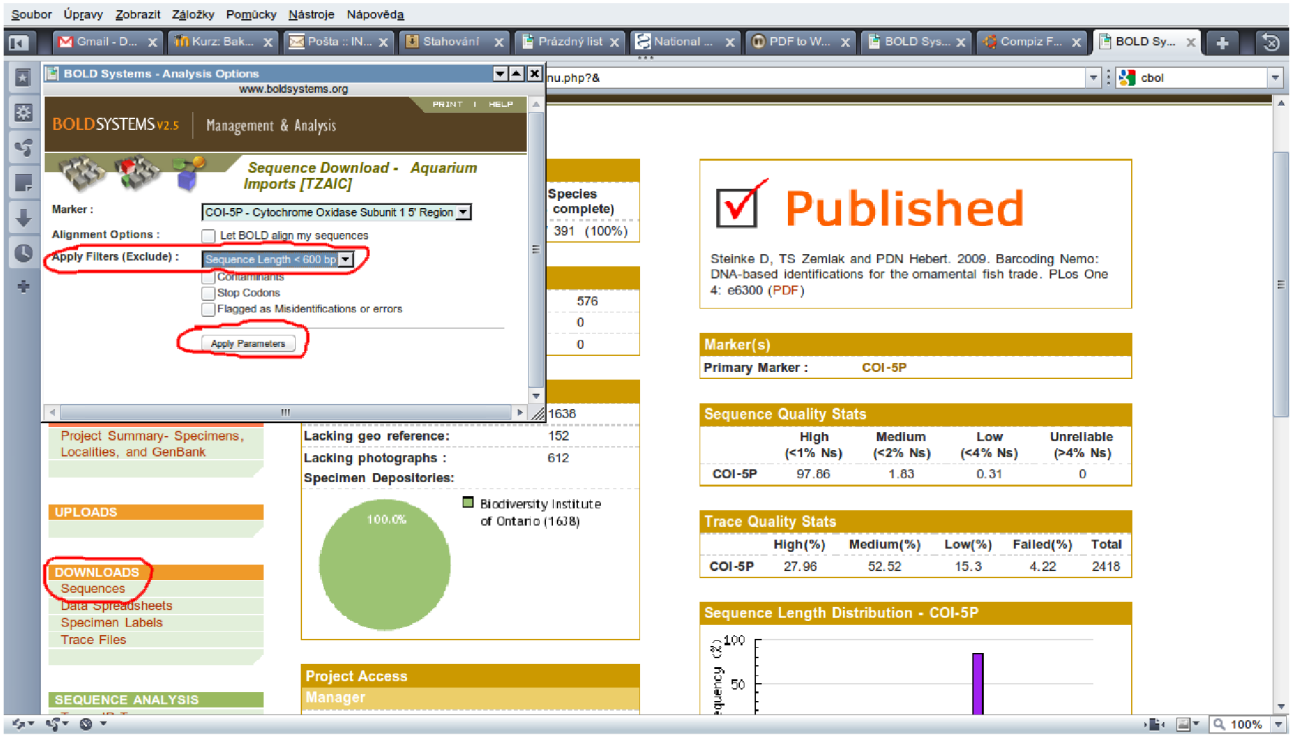
Abychom byli schopni stahovat čárové kódy z této veřejné databáze zabývající se přímo vytvářením a ukládáním čárových kódů, museli jsme si vytvořit uživatelský účet na: http://www.boldsystems.org/views/login.php

Tato internetová stránka je přímo podporována pracovní skupinou CBOL a slouží jako veřejně přístupná knihovna čárových kódů. Po zaregistrování a přihlášení se na účet se dostaneme na seznam právě probíhajících i už dokončených projektů. Pro ukázku jsme si vybrali stažení čárových kódů z projektu Barcoding fish (FishBOL), konkrétně Aquarium Imports.



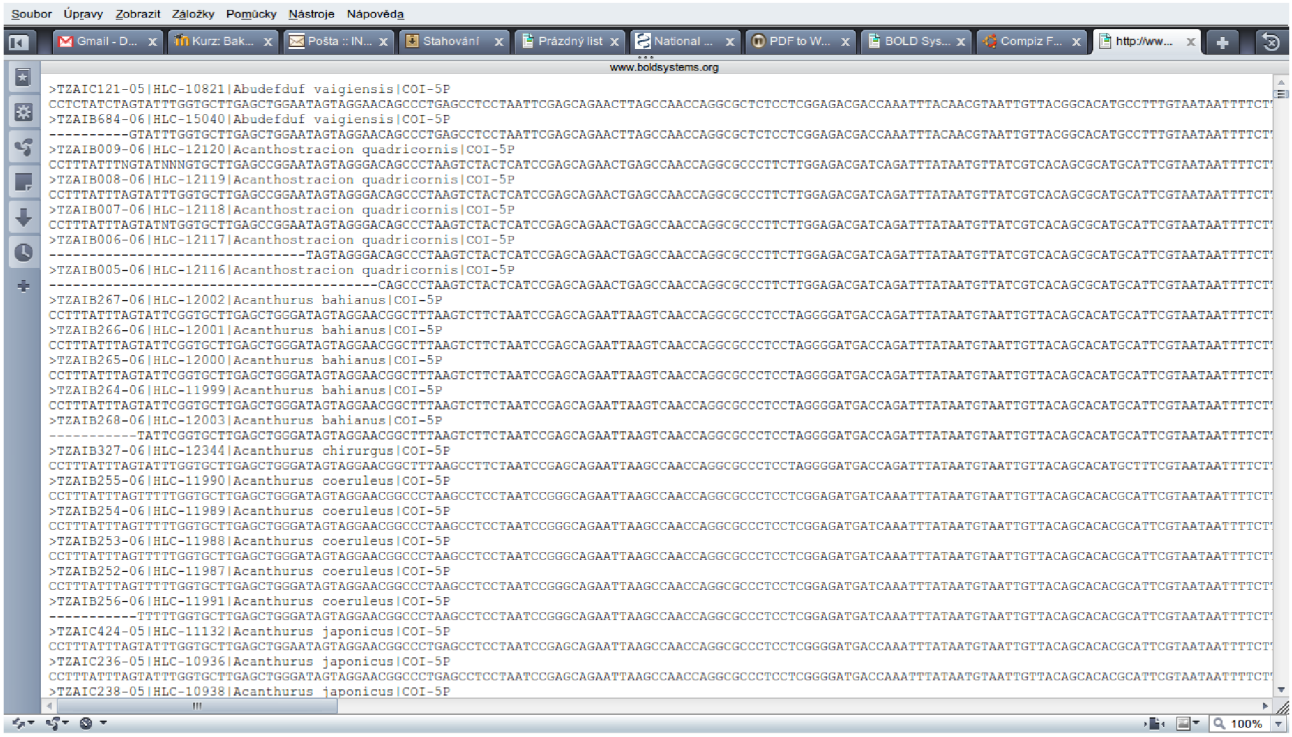
Obr. 13: Výběr skupiny druhů ke stažení z Boldsystems

Z obrázku je vidět, že odkaz Aquarium Imports obsahuje 1638 sekvencí získaných ze 1638 jedinců. Dále je také vidět, že sekvence patří 391 druhům ryb a že všechny druhy ryb mají uznaný čárový kód, který byl získán z genu CO1. Po kliknutí na tento odkaz se uživatel dostane na stránku této zpracované skupiny, zde si najde odkaz ze skupiny Downloads **Sequences** , na který klikne a zobrazí se mu okno, které mu nabízí stažení všech sekvencí ve formátu fasta. Na obrázku pod textem je vidět, jaká nabídka se nám zobrazí.



Obr. 14: Stažení sekvencí z BOLD Systems

V okně BOLD Systems – Analysis Options nic nemění pouze nastavení filtru pro minimální délku sekvencí podle uživatelovy potřeby. Poté klikne na tlačítko Apply Parameters a sekvence zadaných druhů se mu stáhnou v okně prohlížeče v následující podobě:



Obr. 15: Stažené sekvence akvarijních ryb

Takto zobrazené sekvence převede například pomocí poznámkového bloku nebo podobného textového editoru na .txt soubor. S takto upravenými daty umí následně pracovat funkce `getcbol.m`, která všechny tyto sekvence roztřídí a společně s názvem daného druhu je uloží do knihovny `carovy_kod.mat`.

5.2.1 Getcbol.m

```
function pridano = getcbol2(nazev_souboru)
```

Tato funkce má pouze jeden vstup a jeden výstup. Na vstup se přivádí název textového souboru, ve kterém jsou staženy druhy a sekvence z databáze BOLD Systems. Bude-li uživatel pokračovat v uvedeném příkladu přivede na vstup: `'aquarium_imports.txt'`

Slovní popis průběhu funkce `getcbol.m` vypadá následovně. Po spuštění funkce načte knihovnu `carovy_kod.mat`, zazálohuje starou knihovnu do souboru `carovy_kod_zaloha.mat`. Odtud může uživatel zachránit starou knihovnu v případě, kdyby došlo k chybě programu a stará knihovna by se chybně přepsala novou. Poté se otevře soubor `aquarium_imports.txt`, vytvoří si prázdnou strukturu `a` a přečte textový soubor řádek po řádku. Lichý řádek obsahuje informace o názvu druhu a proto jej uloží do proměnné `a.druhh`. Následující sudý řádek obsahuje čárový kód tohoto druhu a uloží jej do `a.carovy_kod`. Jelikož první řádek neobsahuje jen název druhu, sekvence je na BOLD Systems uloženo velkými písmeny a na NCBI je sekvence uložena malými písmeny, musí funkce upravit jak proměnnou `a.druhh`, tak i `a.carovy_kod`. Tyto úpravy jsou prováděny v cyklech od 29. do 72. řádku funkce.

V poslední části už zbývá jen sloučení staré knihovny s nově získanými daty. V cyklu porovnáváme shody mezi proměnnými `rise.druh` (druhy načtené z `carovy_kod.mat`) a `a.druh` (druhy stažené z BOLD Systems). V případě shody v některém kroku jsou čárové kódy uloženy k již vyskytujícímu se druhu za ostatní již uložené sekvence. Pokud dojde k neshodám ve všech

krocích cyklu, znamená to, že se druh uložený pod `a.druh` prozatím v knihovně nenachází. Rozměr struktury knihovny se tedy o 1 zvětší a uloží se nový název druhu i nová sekvence.

Na konci funkce je uložena knihovna `carovy_kod.mat` s nově přidanými sekvencemi a živočišnými druhy a na výstup funkce je přivedena proměnná `pridan`, což je seznam druhů, od kterých byly přidány sekvence do knihovny. Tento výstup je zobrazen v grafickém provedení této funkce.

5.3 Srovnání čárového kódu druhu se stejným genem člověka

Pro srovnávání těchto genů bylo využito dvou algoritmů dynamického programování. Prvním byl Needleman-Wunschův algoritmus, který slouží ke globálnímu zarovnání sekvencí. Druhým byl Smith-Watermanův algoritmus, který zarovnává sekvence lokálně. Oba tyto algoritmy pracují velmi podobně, proto je podrobněji rozebrán Needleman-Wunschův a následně uvedeno v čem se Smith-Watermanův liší.

5.3.1 Needleman-Wunschův algoritmus

Tento matematický model dynamického programování hledá to nejlepší možné globální zarovnání dvou sekvencí ze všech možností, které se mu nabízí. Jelikož bývá tato teorie těžce vstřebatelná a pochopitelná, bude celý problém vysvětlen rovnou na příkladu. Grafické výstupy byly kopírovány z funkce `needleman_wunsch.m`.

Předpokládejme, že chceme zarovnat 2 sekvecne x a y.

x = aagctgact , kde x_m je m-tý symbol sekvence x

y = agccta , kde y_n je n-tý symbol sekvence y

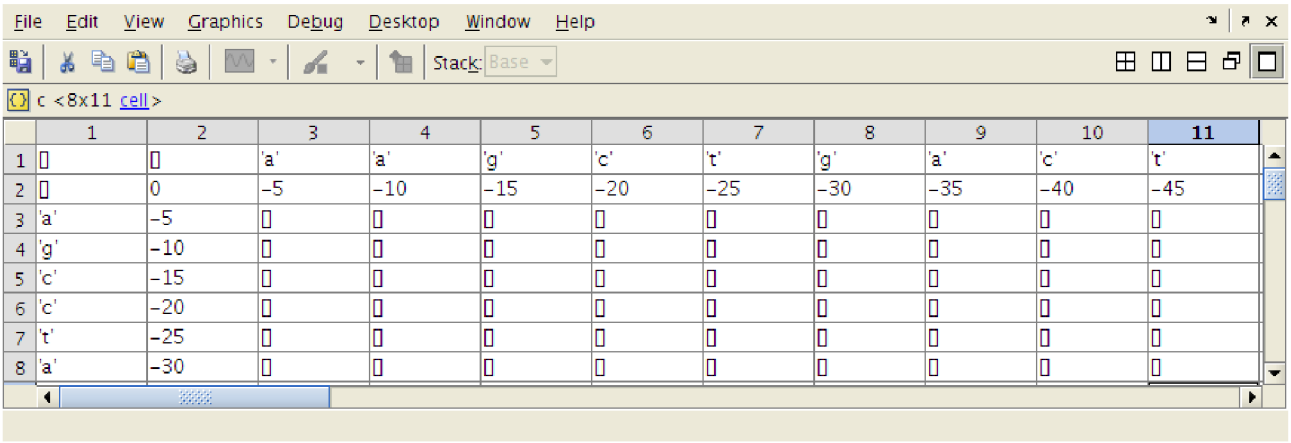
Pro numerický výpočet počítačem bychom do příkazové řádky zadali:

```
[zarovnani,sekvence] = needleman_wunsch('aagctgact','agccta')
```

Tento algoritmus vytváří matici F, kde F(m,n) je skóre nejlepšího zarovnání mezi bázemi x_m a y_n. Jednotlivé prvky matice F jsou počítány z tzv. skórovací matice a penalizace za mezeru. Skórovací matice vyjadřuje míru podobnosti jednotlivých bází mezi sebou. Je několik cest, jak tuto skórovací matici vytvořit, v tomto algoritmu, byla použita tato:

-	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

Penalizaci mezery jsme určili hodnotou `penalizace = -5`. Na počátku funkce se vytvoří prázdná matice o rozměrech (m+1)x(n+1), kde první prvek o indexech (1,1) je roven 0, hodnoty v prvním řádku jsou poté rovny 1*penalizace,2*penalizace,...,m*penalizace, obdobně je zaplněn i první sloupec hodnotami 1*penalizace,2*penalizace,...,n*penalizace.



	1	2	3	4	5	6	7	8	9	10	11
1			'a'	'a'	'g'	'c'	't'	'g'	'a'	'c'	't'
2		0	-5	-10	-15	-20	-25	-30	-35	-40	-45
3	'a'	-5									
4	'g'	-10									
5	'c'	-15									
6	'c'	-20									
7	't'	-25									
8	'a'	-30									

Obr. 16: N-W a. - Naplnění prvního řádku a sloupce

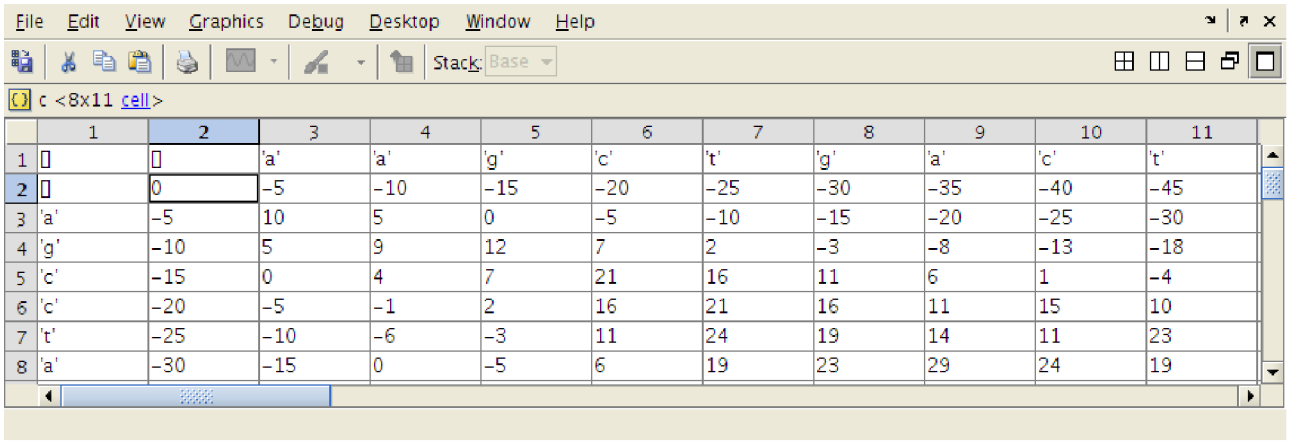
Další prvky matice jsou počítány jako maximum z následujících 3 hodnot:

$F(m,n)=F(m-1,n-1)+s(m,n)$ **(1)**, kde s(m,n) je hodnota skóre ze skórovací matice mezi dvojicí nukleotidů. První dvojice nukleotidů je aa => první hodnota s(m,n) = 10.

$F(m.n)=F(m,n-1)+penalizace$ **(2)**

$F(m.n)=F(m-1,n)+penalizace$ **(3)**

Po provedení výpočtu matlabem dostane uživatel následující matici:



	1	2	3	4	5	6	7	8	9	10	11
1			'a'	'a'	'g'	'c'	't'	'g'	'a'	'c'	't'
2		0	-5	-10	-15	-20	-25	-30	-35	-40	-45
3	'a'	-5	10	5	0	-5	-10	-15	-20	-25	-30
4	'g'	-10	5	9	12	7	2	-3	-8	-13	-18
5	'c'	-15	0	4	7	21	16	11	6	1	-4
6	'c'	-20	-5	-1	2	16	21	16	11	15	10
7	't'	-25	-10	-6	-3	11	24	19	14	11	23
8	'a'	-30	-15	0	-5	6	19	23	29	24	19

Obr. 17: N-W a. - Matice nejlepšího zarovnání F(m,n)

Nyní se zarovnání počítá pomocí tvorby tzv. zpětných šipek, kdy program vychází z pravého dolního rohu matice a vždy zpětným výpočtem určujeme, ze kterého směru sem byla daná hodnota dopočítána. Hodnota F(m,n) se tedy nyní rovná 19, musí se tedy zjistit, jak se program k této hodnotě dostal:

$F(m,n-1)+penalizace=23-5=18$

Tento výpočet nevyšel 19, z toho vyplývá, že šipka nebude směřovat nahoru. Vypočte se tedy diagonální zpětný výpočet:

$F(m-1,n-1)+s(m,n)=11-4=9$

Ani nyní výsledek nevyšel 19, musel být tedy počítán z levé hodnoty, pro kontrolu je ověřen i tento výpočet:

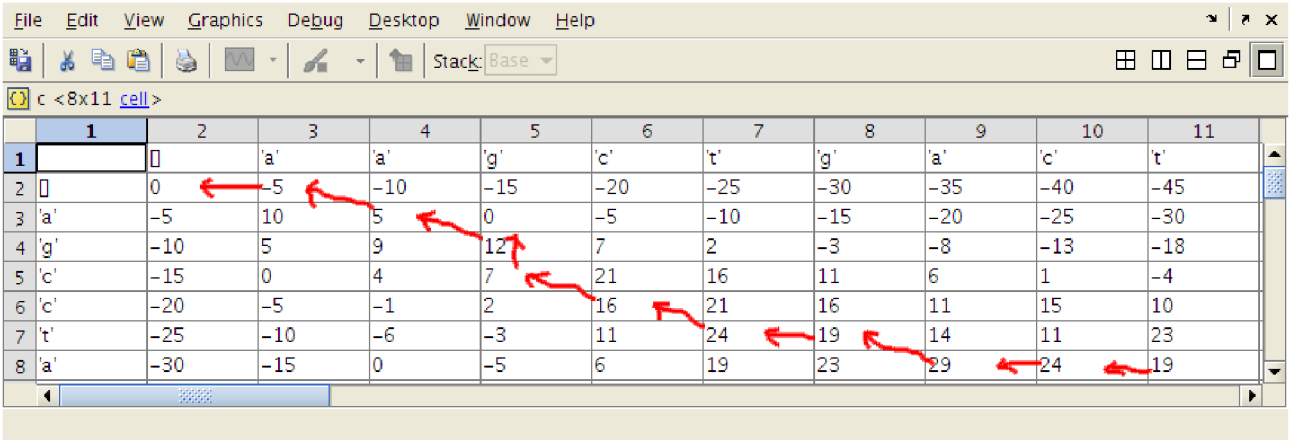
$F(m-1,n)+penalizace=24-5=19$

Tentokrát výpočet již vyšel a vyplývá z toho, že zpětná šipka bude směřovat doleva. To pro

výsledek zarovnání znamená, že m-tá báze sekvence x bude ponechána a namísto n-té báze sekvence y bude vložena mezera. Takto by se zpětnými výpočty v názorném příkladu pokračovalo až do doby, než by se uživatel dostal na hodnotu 0 v matici o souřadnicích (0,0) v matici F(m,n). Tento proces, za uživatele zvládne udělat počítač a mu se již jen zobrazí zarovnané sekvence.

Jen pro úplnost jsou zde uvedena pravidla vkládání mezer. Směřuje-li zpětná šipka doleva, mezera se vkládá do sekvence, která je v matici ve sloupci. Směřuje-li diagonálně, mezera se nevkládá a směřuje-li nahoru, mezera se vkládá do sekvence, která je v matici v řádku.

Po celém zpětném výpočtu by uživatel zjistil, že šipky v matici F(m,n) vypadají takto:



	1	2	3	4	5	6	7	8	9	10	11
1		[]	'a'	'a'	'g'	'c'	't'	'g'	'a'	'c'	't'
2	[]	0	-5	-10	-15	-20	-25	-30	-35	-40	-45
3	'a'	-5	10	5	0	-5	-10	-15	-20	-25	-30
4	'g'	-10	5	9	12	7	2	-3	-8	-13	-18
5	'c'	-15	0	4	7	21	16	11	6	1	-4
6	'c'	-20	-5	-1	2	16	21	16	11	15	10
7	't'	-25	-10	-6	-3	11	24	19	14	11	23
8	'a'	-30	-15	0	-5	6	19	23	29	24	19

Obr. 18: N-W a. - Zpětné šipky algoritmu určující zarovnání

Kdyby si uživatel nechal v příkazové řádce zobrazit zpětné zarovnání těchto dvou sekvencí pomocí funkce `needleman_wunsch.m` dostane následující výstup:

```
aag-ctgact
:||:||:|::
-agcct-a--
```

Obr. 19: N-W a. - Zarovnání

V následující grafické nástavbě programu bude algoritmus pracovat pouze s delšími sekvencemi. Jedna z nich bude stále pevná, neboť v zadání je zadáno za úkol srovnávat čárové kódy jednotlivých druhů s čárovým kódem člověka.

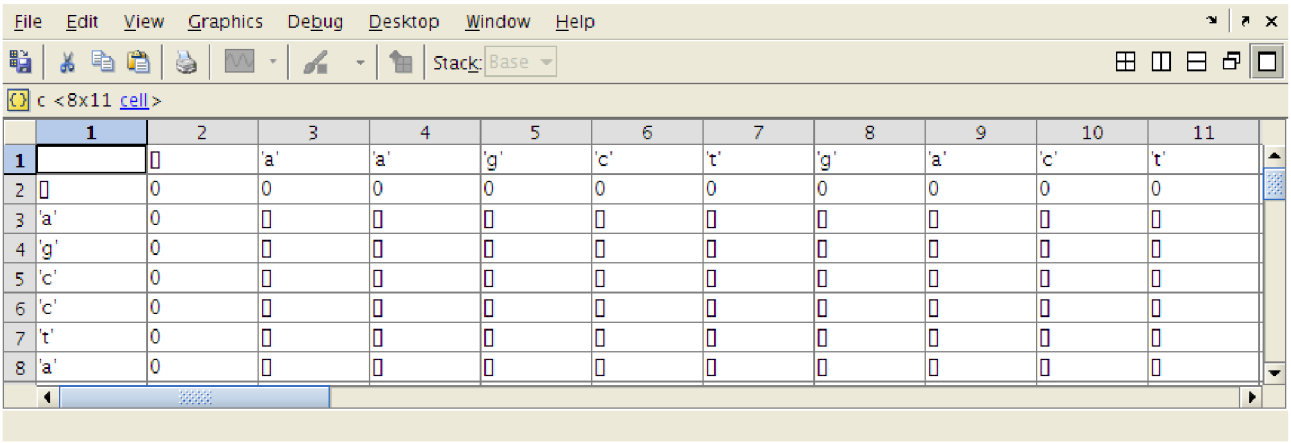
5.3.2 Smith-Watermanův algoritmus

Funkce v matlabu zarovnávající sekvence pomocí této metody má název `smith_waterman.m` a v uvedeném příkladě, je do příkazové řádky zadáno:

```
[zarovnani,sekvence] = smith_waterman('aagctgact','agccta')
```

Tento algoritmus se používá k lokálnímu zarovnání dvou sekvencí. Aby byl ukázán rozdíl mezi oběma algoritmy, bude k zarovnání použita stejná dvojice sekvencí, rovněž indexování, skórovací matice i penalizace mezery bude ponechána. Princip výpočtu je obdobný jako u Needleman-Wunschova algoritmu s několika rozdíly.

První řádek a první sloupec matice F nejsou násobky penalizací, ale nuly.

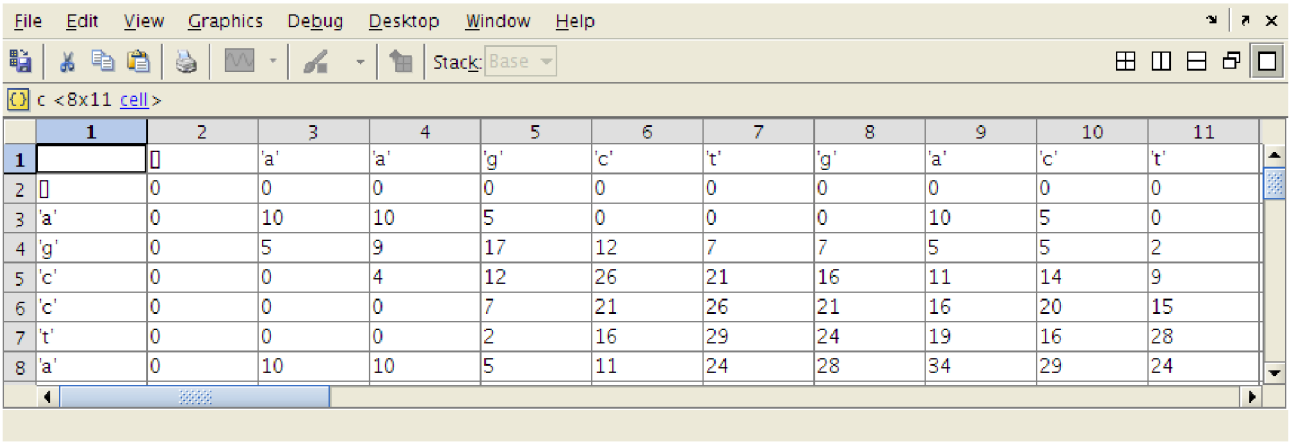


	1	2	3	4	5	6	7	8	9	10	11
1			'a'	'a'	'g'	'c'	't'	'g'	'a'	'c'	't'
2		0	0	0	0	0	0	0	0	0	0
3	'a'	0									
4	'g'	0									
5	'c'	0									
6	'c'	0									
7	't'	0									
8	'a'	0									

Obr. 20: S-W a. - Naplnění prvního řádku a sloupce

Při výpočtu maxima prvku matice F(m,n) k rovnicím **(1), (2), (3)** přibývá 4. rovnice: $F(m,n)=0$ **(4)**

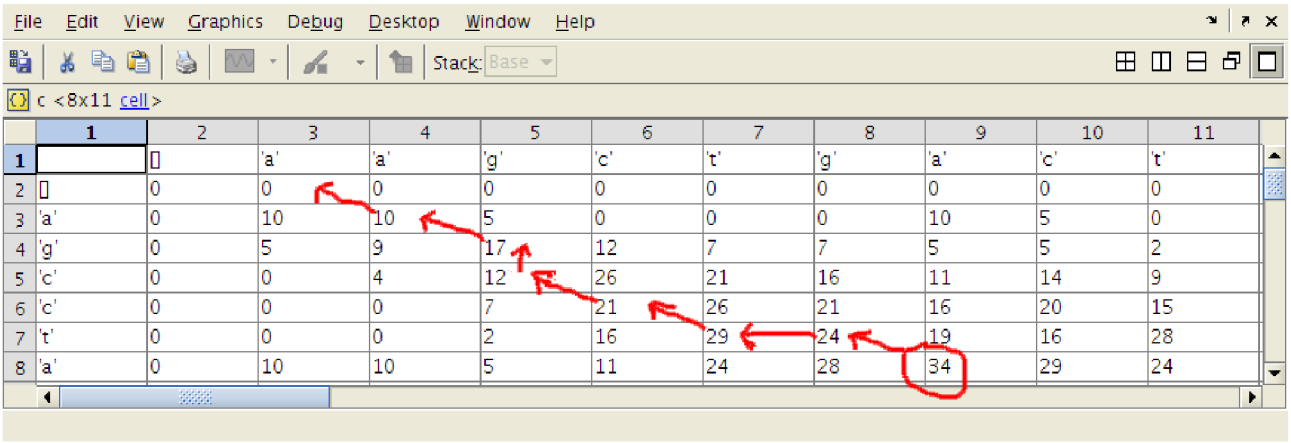
Z toho plyne že hodnoty v matici F nemohou být záporné. Po výpočtu skóre nejlepších zarovnání pro každou dvojici bází vypadá matice F následovně:



	1	2	3	4	5	6	7	8	9	10	11
1			'a'	'a'	'g'	'c'	't'	'g'	'a'	'c'	't'
2		0	0	0	0	0	0	0	0	0	0
3	'a'	0	10	10	5	0	0	0	10	5	0
4	'g'	0	5	9	17	12	7	7	5	5	2
5	'c'	0	0	4	12	26	21	16	11	14	9
6	'c'	0	0	0	7	21	26	21	16	20	15
7	't'	0	0	0	2	16	29	24	19	16	28
8	'a'	0	10	10	5	11	24	28	34	29	24

Obr. 21: S-W a. - Matice nejlepšího zarovnání F(m,n)

Největší rozdíl tohoto algoritmu nastává pravděpodobně při výpočtu zpětných šipek. Nevychází se z pravého dolního rohu matice, ale musí se najít nejvyšší hodnota skóre nejlepšího zarovnání. V tomto případě je vidět, že touto hodnotou je hodnota 34 pro dvojici páru bází AA. Z tohoto bodu jsou výpočty zpětných šipek a pravidla pro vkládání mezer stejná jako u Needleman-Wunschova algoritmu s jediným rozdílem. Výpočet končí v momentě, kdy se šipka dostane k první nule v matici, u Needleman-Wunschova algoritmu výpočet končil až ve chvíli, kdy se zpětné šipky dostaly na nulu o indexu (0,0) v matici F. Tento úsek párů bází mezi maximem v matici F a dosažením nuly v matici F je zarovnán, zbytek sekvence je umazán, z toho plyne, že zarovnané sekvence se zkrátí.



	1	2	3	4	5	6	7	8	9	10	11
1		[]	'a'	'a'	'g'	'c'	't'	'g'	'a'	'c'	't'
2	[]	0	0	0	0	0	0	0	0	0	0
3	'a'	0	10	10	5	0	0	0	10	5	0
4	'g'	0	5	9	17	12	7	7	5	5	2
5	'c'	0	0	4	12	26	21	16	11	14	9
6	'c'	0	0	0	7	21	26	21	16	20	15
7	't'	0	0	0	2	16	29	24	19	16	28
8	'a'	0	10	10	5	11	24	28	34	29	24

Obr. 22: S-W a. - Zpětné šipky algoritmu určující zarovnání

Grafické zobrazení zarovnané dvojice párů bází v příkazové řádce vypadá následovně:

```
ag-ctga
||:||:|
agcct-a
```

Obr. 23: S-W a. - Zarovnání

Ze zobrazení šipek je vidět, že v moment, kdy byla hodnota matice F 21 a hledal se další směr šipky, mohla šipka směřovat i nahoru na hodnotu 26 a poté diagonálně na hodnotu 17. Z tohoto výsledku je vidět, že jedna matice může poskytnout více možností zarovnání. V tomto případě by došlo k záměně mezery a báze C v horní sekvenci. Počítač, určil přednostně toto zarovnání, neboť ve zdrojovém kódu funkce při hledání zpětných šipek nejprve počítá diagonální výpočet, výpočet šipky směrem nahoru má v podmínkách nadefinován až jako poslední možnost.

```
while u >= 3 && v >= 3 && F{u,v} > 0
    skore = F{u,v};
    skore_diag = F{u-1,v-1};
    %skore_nahoru = F{u-1,v};
    skore_nalevo = F{u,v-1};
    if skore == (skore_diag + S(u-2,v-2))
        seq1_zarovnana = [F{1,v} seq1_zarovnana];
        seq2_zarovnana = [F{u,1} seq2_zarovnana];
        u = u - 1;
        v = v - 1;
    elseif skore == (skore_nalevo + penalizace)
        seq1_zarovnana = [F{1,v} seq1_zarovnana];
        seq2_zarovnana = ['-' seq2_zarovnana];
        v = v - 1;
    else
        seq1_zarovnana = ['-' seq1_zarovnana];
        seq2_zarovnana = [F{u,1} seq2_zarovnana];
        u = u - 1;
    end
end
```

5.4 Identifikace druhu organismu podle zadané sekvence genu CO1

Táto část práce byla vyřešena ve funkci `identifikace.m`, kde na vstup je přiváděn čárový kód u kterého chce uživatel zjistit, jakému druhu patří.

```
function [druh,znacka] = identifikace(carovy_kod)
```

Kdyby se podíval na zdrojový kód této funkce, viděl by, že celý problém řeší pouze 2 cykly `while`, které postupně prohledávají a srovnávají jeden čárový kód za druhým s uživatelem zadanou

sekvencí. V momentě shody jsou ukončeny oba cykly a funkce zobrazí výstup s názvem druhu. Pokud nedojde ke shodě v žádném kroku, funkce vypíše, že sekvence hledaného druhu se v databázi nenachází. Výstupem této funkce je ještě proměnná značka, která nese pouze binární hodnotu pro následující grafickou nástavbu.

Jelikož řešení tohoto problému bylo triviální, bylo rozhodnuto vytvořit i funkce pro hledání sekvencí podle zadaného druhu. Princip funkce je obdobný, proto ho nemá smysl dále rozebírat. Jediné, co stojí za zmínku je, že tato funkce byla napsána dvakrát.

```
function [carovy_kod,znacka] = identifikace2(druh,ncbi)

function [carovy_kod,znacka] = identifikace3(druh)
```

Funkce `identifikace2.m` najde přímo konkrétní sekvenci, která je v knihovně nahrána z veřejné databáze NCBI. Funkce `identifikace3.m` najde pouze první sekvenci v pořadí u daného druhu. Jelikož jsou všechny sekvence čárového kódu od jednoho druhu ve více jak 90% případů identické, nebylo považováno za potřebné, jak jedinečně specifikovat sekvence z databáze BOLD Systems a nalezení první sekvence z databáze je považováno za dostatečně přesné.

5.5 Převod sekvencí do numerických formátů

Jelikož numerické formáty sekvencí čárových kódů budou následně využívat algoritmy k podobnostní analýze pomocí distanční matice, rozhodlo se použít 2D a 3D numerické formáty, které nesou informaci o chemických vlastnostech bází. Základním principem těchto převodů sekvencí na numerické formáty je přiřazení specifického vektoru čísel každé bázi. U 2D formátů je to vektor o dvou prvcích, u 3D formátů se tedy logicky přiřadí vektor o třech prvcích. Výsledkem je matice, která na prvním řádku obsahuje vektor, popisující první bázi. Na druhém řádku je vektor, popisující první dvě báze. Tato informace vznikne jako součet vektorů, popisující první a druhou bázi. Takto se postupuje až do konce sekvence, až k poslední bázi, kterou nyní reprezentuje suma všech vektorů, reprezentujících všechny báze.

5.5.1 2D numerický formát

Tato metoda přiřazuje každé bázi v sekvenci 2D vektor, který leží v 1. a 4. kvadrantu kartézské soustavy souřadnic.

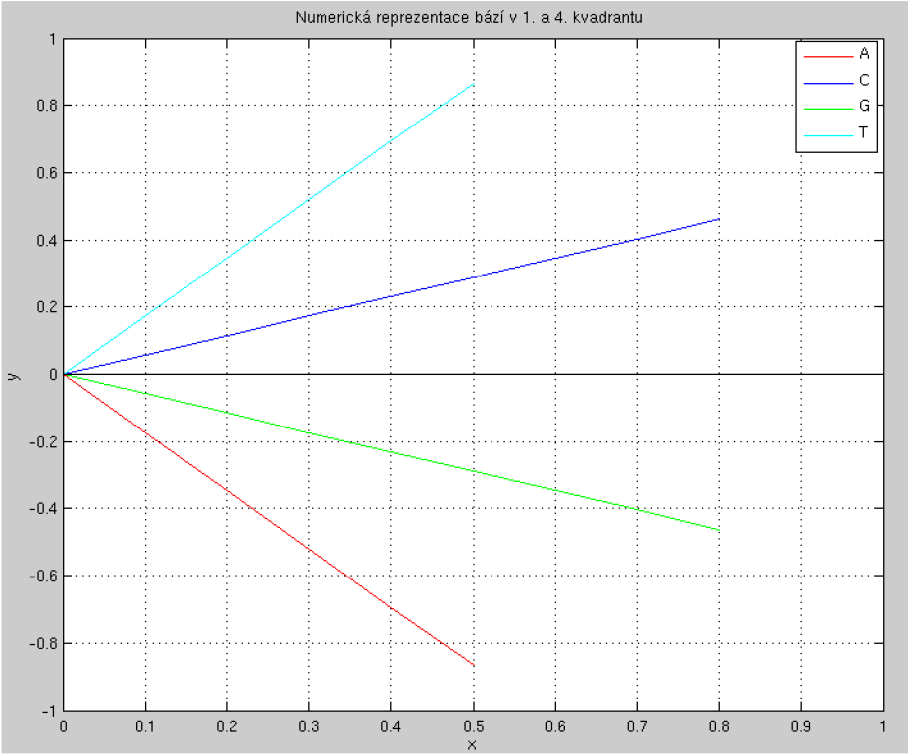
$$A = (\frac{1}{2}, -\frac{\sqrt{3}}{2})$$

$$C = (\frac{\sqrt{3}}{2}, \frac{1}{2})$$

$$G = (\frac{\sqrt{3}}{2}, -\frac{1}{2})$$

$$T = (\frac{1}{2}, \frac{\sqrt{3}}{2})$$

Kladná část osy y reprezentuje pyrimidiny, záporná puriny v bázi. Kladná část osy x představuje dojnou vazbu v dvojšroubovici DNA, záporná trojnou vazbu. Tento vektor nenese žádnou informaci o přítomnosti amino- nebo keto- skupiny v bázi. Při grafickém zobrazení vektorů v 1. a 4. kvadrantu vyjde následující graf:



Obr. 24: Grafické zobrazené bází v kartézské soustavě souřadnic

V matlabu tyto vektory 2D reprezentace sekvence čárového kódu přiřazuje funkce `reprezentace2D.m`, která je tvořena jedním cyklem `for`, který přeskakuje z jedné báze na druhou a příkazem `switch` uvnitř cyklu, který přičítá postupně jeden vektor reprezentující aktuální bázi k součtu všech vektorů předešlých bází. Na výstupu vystoupí matice, která má 2 sloupce a řádků tolik, kolik bází sekvence obsahuje.

```
>> a = reprezentace2D('atgccgtacg')

a =

    0.5000   -0.8660
    1.0000         0
    1.8660   -0.5000
    2.7321         0
    3.5981    0.5000
    4.4641         0
    4.9641    0.8660
    5.4641         0
    6.3301    0.5000
    7.1962         0
```

Obr. 25: 2D numerický formát sekvence bází molekuly DNA

5.5.2 3D numerický formát

Tato metoda přiřazuje každé bázi v sekvenci vektor o 3 prvcích. Konkrétně přiřazuje tyto vektory:

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

První souřadnice nese informaci o tom, zda je struktura báze založena na purinu (0) nebo pyrimidinu (1), druhá souřadnice rozlišuje amino- (0) a keto- skupinu (1), třetí souřadnice popisuje

dvojnou (0) a trojnou (1) vazbu vodíkových můstků mezi bázemi ve šroubovici molekuly DNA.

Princip výpočtu a algoritmus výpočtu 3D numerického formátu analyzované sekvence je naprosto stejný jako u 2D reprezentace. Na výstupu jen nevyjde matice o dvou, ale o třech sloupcích.

```
>> a = reprezentace3D('atgccgtacg')

a =

     0     0     0
     1     1     0
     1     2     1
     2     2     2
     3     2     3
     3     3     4
     4     4     4
     4     4     4
     5     4     5
     5     5     6
```

Obr. 26: 3D numerický formát sekvence bází molekuly DNA

5.6 Porovnávání distanční maticí

Porovnávání distanční maticí je způsob zpracování dvou sekvencí, pomocí kterého je možné posuzovat míru podobnosti této dvojice sekvencí. V této práci byla k této analýze vytvořena funkce `distancni_matice.m`, která počítá tzv. hodnoty pásových vektorů (band average), což jsou průměrné hodnoty prvků všech diagonál nad hlavní diagonálou distanční matice.

```
function band_average = distancni_matice(ciselna_sekvence)
```

Zvolený numerický formát, do kterého je převedena sekvence, přímo ovlivňuje výsledek analýzy porovnávání dvou sekvencí distanční maticí, neboť na vstup funkce `distancni_matice.m` je sekvence přiváděna právě v numerickém formátu. Tato funkce umožňuje výpočet distanční matice pro sekvence ve 2D a 3D numerických formátech.

Jednotlivé prvky distanční matice jsou počítány podle níže uvedeného vzorce:

$$D_{i,j} = \sqrt{\sum_{n} (x_{ni} - x_{nj})^2} \quad \textbf{(5)}$$

kde n představuje počet rozměrů numerické reprezentace bází v sekvenci, i a j jsou pořadí bází v sekvenci. Ze vzorce tedy vyplývá, že prvek distanční matice je vypočítán pro každou dvojici bází v sekvenci. Pro názornost je uveden příklad distanční matice v paměti počítače pro sekvenci z podkapitoly 5.5.2, kde byla sekvence o délce 10 bází převedena do 3D numerického formátu. Distanční matice bude tedy mít rozměr 10x10. Jelikož je počítán vzájemný rozdíl každé s každou numericky reprezentovanou bází, lze ze vzorce usuzovat, že matice bude symetrická podle hlavní diagonály.

	1	2	3	4	5	6	7	8	9	10
1	0	1.4142	2.4495	3.4641	4.6904	5.8310	6.9282	6.9282	8.1240	9.2736
2	1.4142	0	1.4142	2.4495	3.7417	4.8990	5.8310	5.8310	7.0711	8.2462
3	2.4495	1.4142	0	1.4142	2.8284	3.7417	4.6904	4.6904	6	7.0711
4	3.4641	2.4495	1.4142	0	1.4142	2.4495	3.4641	3.4641	4.6904	5.8310
5	4.6904	3.7417	2.8284	1.4142	0	1.4142	2.4495	2.4495	3.4641	4.6904
6	5.8310	4.8990	3.7417	2.4495	1.4142	0	1.4142	1.4142	2.4495	3.4641
7	6.9282	5.8310	4.6904	3.4641	2.4495	1.4142	0	0	1.4142	2.4495
8	6.9282	5.8310	4.6904	3.4641	2.4495	1.4142	0	0	1.4142	2.4495
9	8.1240	7.0711	6	4.6904	3.4641	2.4495	1.4142	1.4142	0	1.4142
10	9.2736	8.2462	7.0711	5.8310	4.6904	3.4641	2.4495	2.4495	1.4142	0

Obr. 27: Distanční matice

Z matice je vidět, že předpoklad se opravdu numericky potvrdil. V další části programu jsou počítány průměrné hodnoty diagonál nad hlavní diagonálou, již výše zmiňované hodnoty pásový vektor (band average). Pro ukázku je opět uveden výstup přímo z paměti počítače, který tyto hodnoty ukládá na výstup naprogramované funkce.

	1	2	3	4	5	6	7	8	9
1	1.2571	2.2380	3.1086	4.1120	5.1466	6.1475	7.0234	8.1851	9.2736

Obr. 28: Hodnoty pásového vektoru

Tento výstup je následně využíván grafickou nástavbou tohoto programu, který ze dvou těchto vektorů, od dvou různých druhů, vypočítá míru podobnosti obou druhů. Aby se tedy určila míra podobnosti dvou druhů, musí se dvakrát za sebou použít funkce `distancni_matice.m`.

Míra podobnosti dvou druhů je počítána jako průměr euklidovských vzdáleností jednotlivých prvků vektorů `band_average` prvního druhu a `band_average` druhého druhu. Pro zkrácení zápisu ve vzorci byly proměnné `band_average` zkráceny na BA.

$$\Delta BA = \sqrt{(\Delta BA_1^2 + \Delta BA_2^2 + \ldots + \Delta BA_n^2)} \quad \textbf{(6)}$$

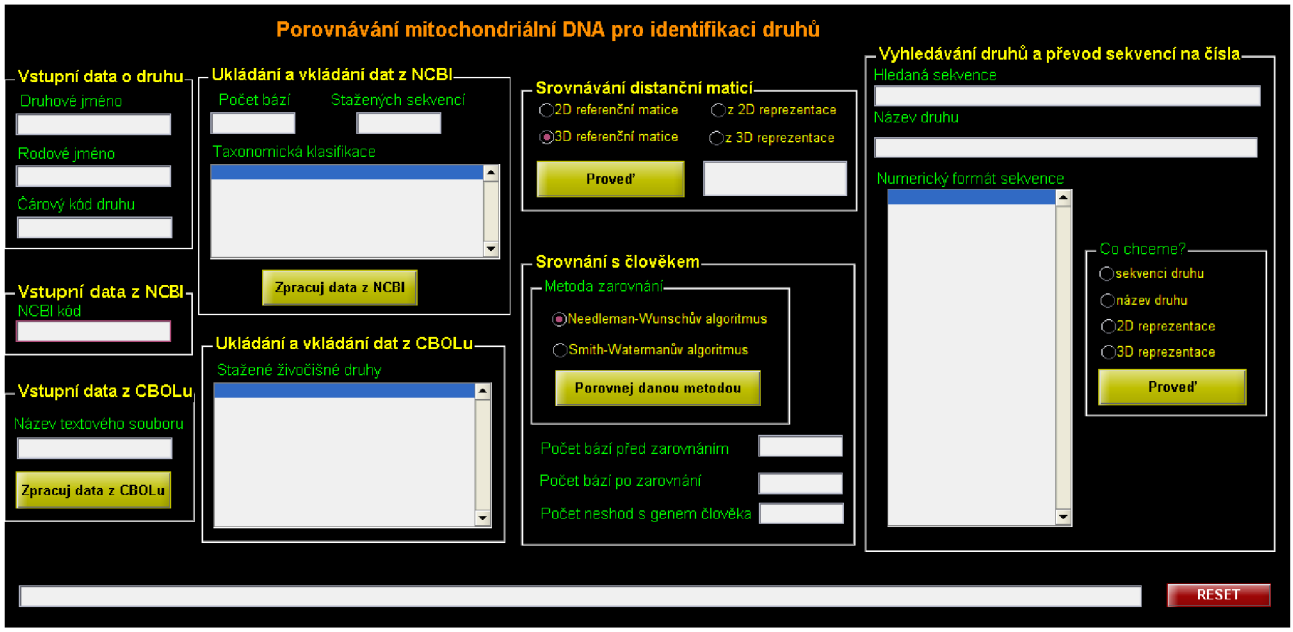
kde ΔBA_1 je rozdíl prvních prvků pásových vektorů obou sekvencí, ΔBA_2 je rozdíl druhých prvků pásových vektorů, tyto rozdíly jsou počítány až po ΔBA_n rozdíl posledních prvků pásových vektorů. ΔBA následně představuje míru podobnosti obou sekvencí. Tuto hodnotu již ale nepočítá funkce `distancni_matice.m`, ale přímo grafická nástavba programu.

6. GUI – Grafická nástavba programu

Jelikož cílovou skupinou, kteří by měli výše popsané funkce využívat, jsou převážně biologové, muselo dojít i k naprogramování graficky a hlavně uživatelsky příjemného pracovního prostředí. Důvod byl prostý, většině biologům se naježí všechny chlupy po těle, při vyslovení sousloví „příkazová řádka“. Naneštěstí programovací jazyk matlab umožňuje tvorbu grafického prostředí ke svým zdrojovým kódům. Toto doplnění bylo vytvořeno v grafické nástavbě GUI (Guide).

Každá grafická nástavba ke zdrojovým kódům se skládá ze dvou souborů. První soubor má příponu .fig, kde po otevření tohoto souboru dojde k rozvržení grafického vzhledu budoucího programu. Druhým souborem je běžný .m soubor, kde se nachází zdrojová část grafiky, doplněná o další podmínky a algoritmy, které zajišťují plynulý chod programu.

Grafická forma programu pro stahování a analýzu CO1 genu živočišných druhů se nachází na přiloženém CD pod názvy souborů `porovnavani.fig` a `porovnavani.m`. Po otevření souboru `porovnavani.fig` se spustí a zobrazí program Porovnávání mitochondriální DNA pro identifikaci druhů.



Obr. 29: Grafický vzhled programu

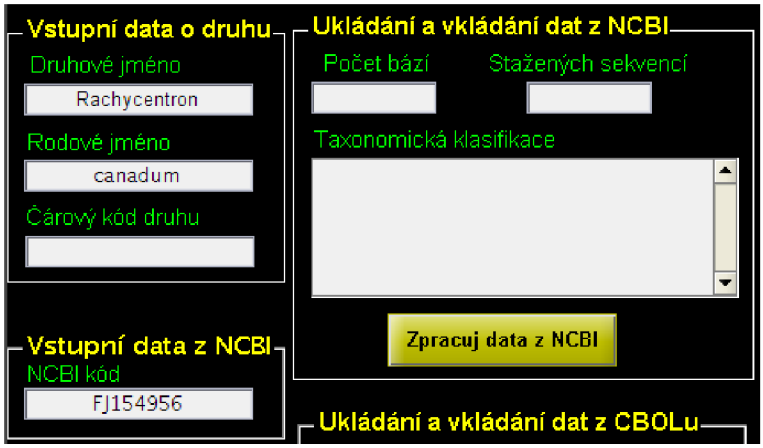
Na první pohled je vidět, že program je sestaven z několika bloků. První blok tvoří série tří oken nalevo, které slouží jako vstupy pro celý blok programu. Prvními vstupy jsou vstupní data o druhu. S tímto oknem bude uživatel nejčastěji pracovat, téměř ve všech případech, kdy bude uživatel po programu požadovat nějaký úkon, bude zadávat vstupní informace pod kolonku druhové jméno, rodové jméno nebo čárový kód druhu. Druhým vstupním blokem, jsou vstupní data z databáze NCBI. Zde se jedná pouze o NCBI kód, který jedinečně specifikuje požadovanou sekvenci druhu, kterou chce uživatel získat. Třetí a poslední vstupní blok tvoří vstupní data z CBOLu. Tento vstup je používán pouze při stahování a ukládání dat z veřejné databáze BOLD Systems.

Další část programu je tvořena dvěma bloky napravo od vstupních bloků, které slouží pro stahování a ukládání dat z veřejných databází. Pravá polovina okna slouží k výpočetním analýzám sekvencí čárových kódů, které má program umět řešit.

V dolní části okna je jeden dlouhý výpisový řádek, který slouží k výpisu aktuálních stavů a procesů o tom, co zrovna program dělá. Tento výpisový řádek má sloužit k informování uživatele o průběhu zadaného procesu a hlavně k upozornění a nápovědě při chybném používání programu. V pravém dolním rohu se nachází červené tlačítko RESET, které vynuluje všechny hodnoty ve všech

vstupních i výstupních buňkách programu.

6.1 Ukládání a vkládání dat z NCBI

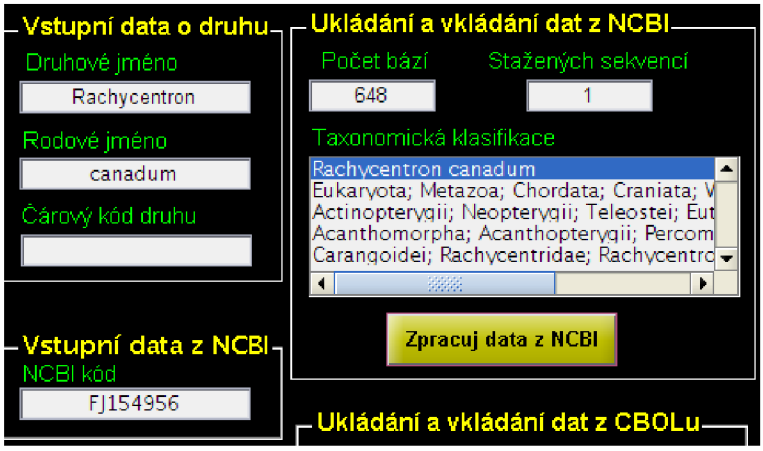


Obr. 30: Zpracování dat z NCBI

Pokud chce uživatel přidat do knihovny `carovy_kod.mat` novou sekvenci popřípadě celý nový živočišný druh z databáze NCBI, musí si nejprve dohledat NCBI kód, pod kterým jsou druh a sekvence uloženy v této databázi. Způsob hledání této informace byl popsán v kapitole 5.1.1. Pokud již uživatel zná NCBI kód, stačí mu pouze vyplnit druhové, rodové jméno a NCBI kód tak, jako na obrázku nad textem. Hledaný druh je Rachycetron canadum, v české literatuře označován jako kranasovec štíhlý.

Aby došlo k plnohodnotnému a kvalitnímu sloučení dat z databází NCBI a BOLD Systems, doporučuje se u druhového jména začít velkým písmenem. Programovací jazyk matlab rozlišuje malé a velké písmeno jednoho znaku jako dva rozdílné. V databázi BOLD Systems jsou druhová jména uloženy s velkým písmenem, pokud by uživatel zadal druhové jméno s malým písmenem, nedošlo by k chybě programu, pouze by se informace uložily pod položku, pod kterou by se následně nesloučily s případnými dalšími daty z druhé databáze.

Po vyplnění vstupních dat uživatel zmáčkne tlačítko Zpracuj data z NCBI, program začne pracovat, stahovat a ukládat data.

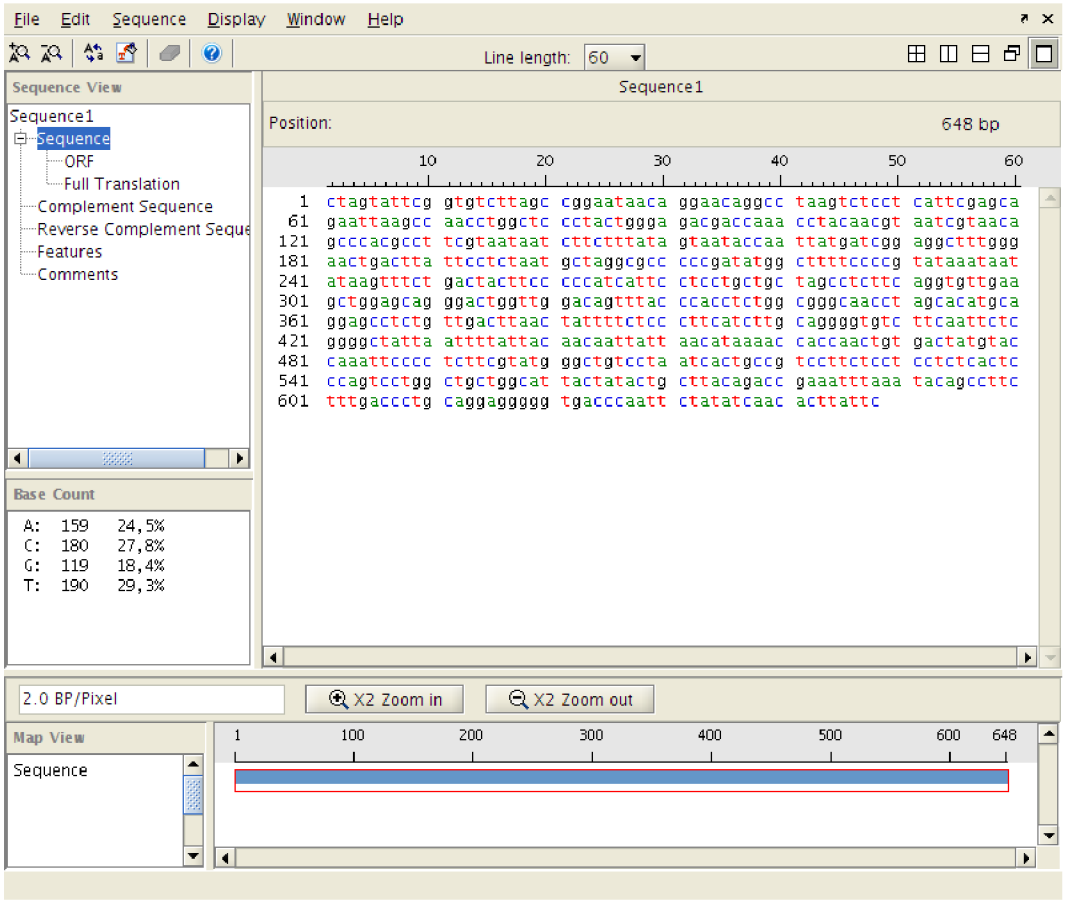


Obr. 31: Výstupní data po zpracovaní dat z NCBI

Po stisknutí tlačítka se spustí algoritmy `barcode.m` a `knihovna.m`. Po ukončení algoritmů se vyplní pole počet bází, stažených sekvencí a taxonomická klasifikace. Pole Počet bází nás informuje o délce stažené sekvence CO1 genu. Pole Stažených sekvencí zobrazuje, kolik sekvencí zadaného druhu máme již uloženo v knihovně `carovy_kod.mat`. Pokud je hodnota jedna, znamená

to, že jsme do databáze uložili první sekvenci daného druhu a že se tento druh v knihovně doposud nevyskytoval. Pole Taxonomická klasifikace zobrazuje informace o zařazení druhu v živočišné říši.

V novém okně se nám otevře Sequence viewer, což je grafická nástavba bioinformatického toolboxu pro zobrazování sekvencí molekuly DNA, s právě staženou částí genu CO1.



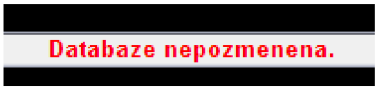
Obr. 32: Zobrazení stažené sekvence pomocí Sequence viewer

Kromě stažené sekvence, tento prohlížeč zobrazí uživateli informace o procentuálním zastoupení jednotlivých bází v sekvenci a umožní mu převést sekvenci do několika dalších formátů. Zobrazí mu ORF (open reading frame) na sekvenci, ukáže mu, jak by se úsek genu translatoval, převede sekvenci na komplementární nebo zpětnou.

Jak bylo na začátku kapitoly uvedeno, s uživatelem komunikuje výpisová řádka. V tomto případě vypsala:

Obr. 33: Výpisová řádka NCBI 1

Kdyby uživatel tlačítko Zpracuj data z NCBI zmáčkl podruhé, nedošlo by k žádné změně v knihovně a ve výpisové řádce by se objevilo:



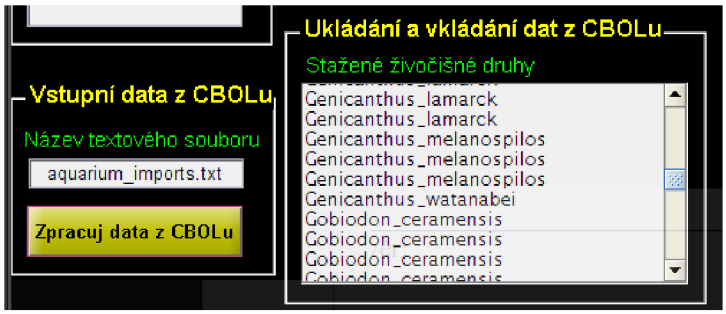
Obr. 34: Výpisová řádka NCBI 2

Pokud by uživatel nezadal jeden nebo více vstupních parametrů pro stažení dat z databáze NCBI, ve výpisu se objeví:

Obr. 35: Výpisová řádka NCBI 3

6.2 Ukládání a vkládání dat z CBOLu

Pro vložení dat z databáze BOLD Systems je třeba postupovat podle postupu popsaného v kapitole 5.2 s tím rozdílem, že uživatel již nepotřebuje znát, jak pracuje funkce `getcbol.m`. Stačí mu, když stažená data uloží do libovolně nazvaného .txt souboru a jeho název i s příponou napíše do kolonky Vstupní data z CBOLu. Následně již stačí jen zmáčknout tlačítko Zpracuj data z CBOLu a software začne pracovat za uživatele.



Obr. 36: Získávání dat z databáze BOLD Systems

Obr. 37: Výpisová řádka CBOL 1

Po ukončení algoritmu program zobrazí uživateli seznam nově uložených druhů do databáze a o průběhu operace ho opět informuje výpisový řádek.

Pokud by uživatel zapomněl, jak se do knihovny ukládají sekvence genu CO1 a názvy druhů z databáze BOLD Systems, výpisová řádka mu ihned intuitivně poradí:

Obr. 38: Výpisová řádka CBOL 2

Pokud by uživatel v tomto případě stiskl tlačítko dvakrát po sobě, veškerá data z .txt souboru se mu do knihovny nahrají dvakrát.

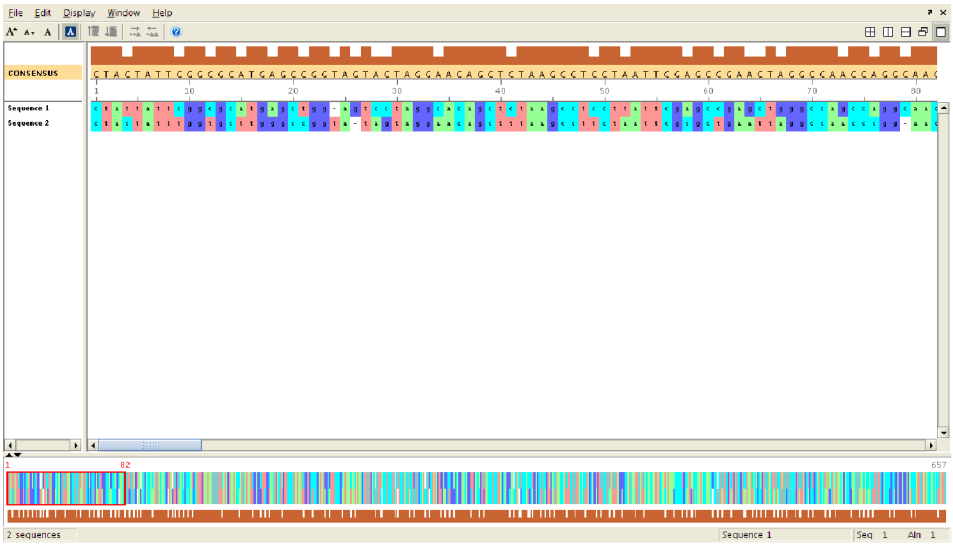
6.3 Srovnání čárového kódu se stejným genem pro člověka

Při srovnávání živočišných druhů s genem člověka má uživatel na výběr ze dvou možností zarovnání (Needleman-Wunschův a Smith-Watermanův algoritmus). Jako vstupní informace se používají všechny vstupy o živočišném druhu a vstup o NCBI kódu. Uživatel má na výběr ze tří možností nastavení vstupních dat, podle kterých bude software zarovnávat.

První možností je zapsání druhového a rodového jména druhu do vstupů. Software poté spustí vyhledávací algoritmus `identifikace3.m`, který nalezne první uloženou sekvenci v knihovně `carovy_kod.mat` zadaného druhu. Pokud se zadaný druh v databázi nenachází, výpisový řádek zobrazí:

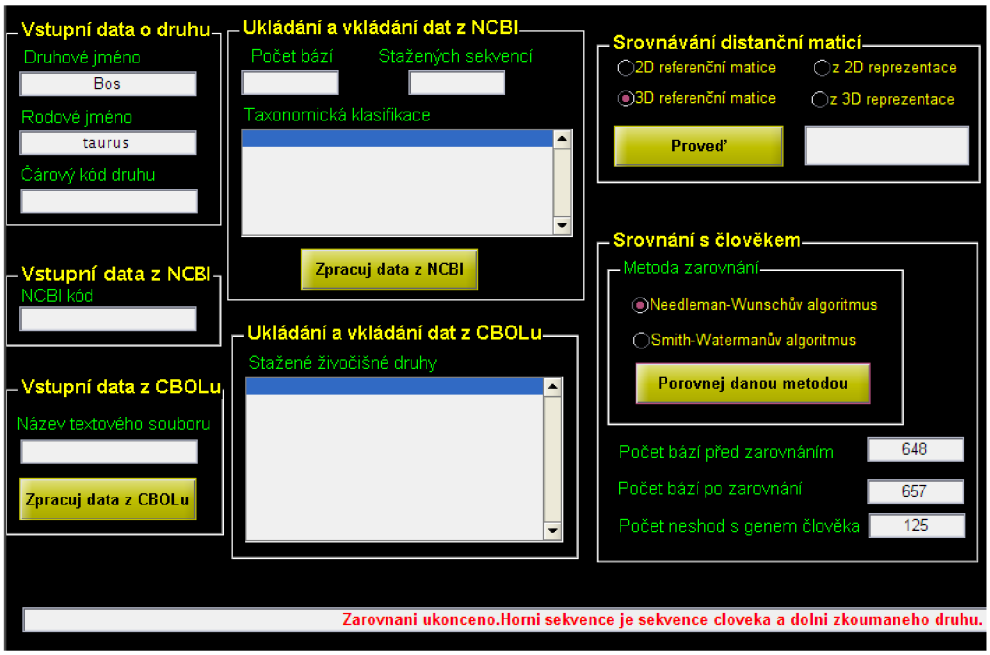
Obr. 39: Výpisový řádek srovnávání 1

Pří nalezení požadované sekvence se spustí Needleman-Wunschův nebo Smith-Watermanův algoritmus, podle toho, který z nich si uživatel vybral. Pro ukázku výstupu tohoto bloku programu je v práci uvedeno zarovnání druhu bos taurus (tur domácí), po zadání výše uvedených vstupních informací.



Obr. 40: Výstup dvojice zarovnaných čárových kódů

Zarovnané sekvence se nám zobrazí pomocí softwaru bioinformatického toolboxu Multiple Sequence Alignment Viewer, kde horní sekvence představuje čárový kód člověka a spodní sekvence sekvenci tura domácího. Přímo v okně programu Porovnávání mitochondriální DNA pro identifikaci druhů se zobrazují informace o délce sekvence tura domácího před a po zarovnání a také program počítá počet neshod čárového kódu tura domácího s člověkem. Výpisová řádka uživatele informuje, která sekvence vyjadřuje který druh ve výstupu zarovnaných sekvencí.



Obr. 41: Výstupy zarovnání v okně programu

Druhým způsobem zadání vstupních dat je k druhovému a rodovému jménu druhu přidat ještě NCBI kód, díky kterému software pomocí funkce `identifikace2.m` nalezne přímo konkrétní sekvenci uloženou v knihovně `carovy_kod.mat`.

Třetím a posledním způsobem zadání vstupních dat je zadání přímo sekvence, kterou chceme zarovnat do kolonky Čárový kód druhu.

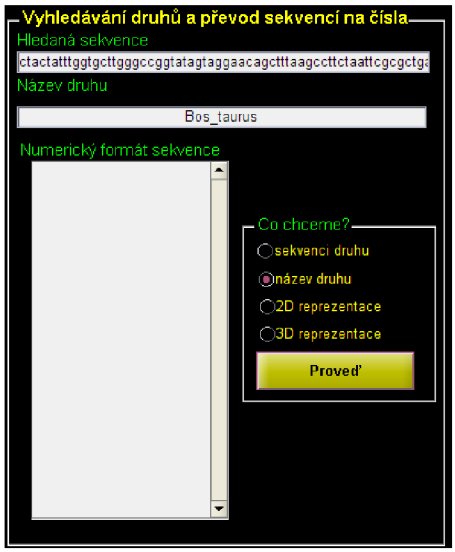
Dále je postup analýzy naprosto totožný jako u zadání vstupních dat prvním způsobem.

Obr. 42: Výpisová řádka: nápověda u zarovnání

Kdyby uživatel nezadal nebo zadal vstupní informace do jiných kolonek, výpisová řádka se mu bude snažit poradit.

6.4 Vyhledávání názvů druhů, sekvencí, převod na numerické formáty

Všechny tyto funkce zpracovává blok programu úplně v pravé části okna.



Obr. 43: Blok vyhledávání a převodu na numerické formáty

6.4.1 Vyhledávání názvů druhů a sekvencí

Pokud uživatel hledá název druhu podle zadané sekvence, zadá tuto sekvenci do vstupní kolonky Čárový kód druhu, ve výběru druhu operace vybere název druhu a stiskne tlačítko proveď. Hledaný název druhu se mu objeví v kolonce Název druhu a nebo ho výpisová řádka informuje o faktu, že druh s touto zadanou sekvencí se v knihovně nenachází. Tuto operaci řeší funkce `identifikace.m`.

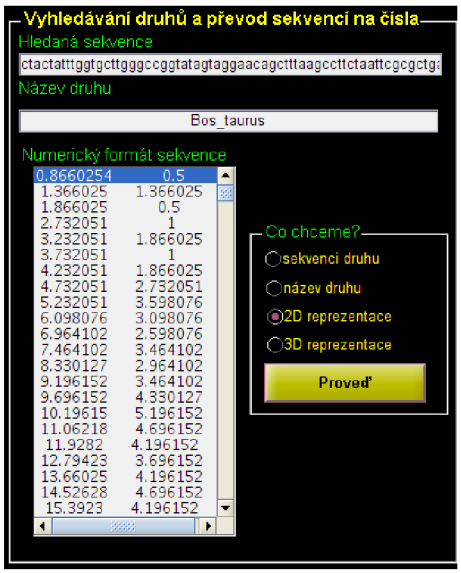
Při hledání sekvence druhu podle názvu druhu je postup obdobný jako v předešlém případě, akorát se do vstupních kolonek zadávají opačné informace. Stejně jako při zadávání vstupních informací při srovnávání sekvence DNA s genem člověka i teď program umožňuje najít první sekvenci uloženou u daného druhu při zadání druhového a rodového jména druhu a nebo konkrétní sekvenci získanou z databáze NCBI při zadání třetí vstupní informace NCBI kódu. Tyto operace jsou řešeny funkcemi `identifikace3.m` nebo `identifikace2.m`.

V obou případech při chybném zadání vstupních dat uživateli radí výpisová řádka.

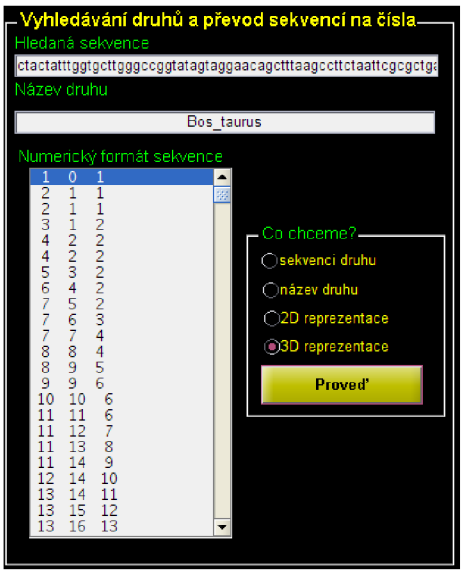
6.4.2 Převod na numerické formáty

Jak bylo nastíněno v podkapitolách 5.5.1 a 5.5.2, program umožňuje převod sekvence do jednoho 2D a jednoho 3D formátu. Tento výběr uživatel provede označením tlačítka 2D reprezentace nebo 3D reprezentace. Pro vstup na převod do numerických formátů mohou sloužit 2 kolonky v programu. První možností je ve vstupních datech Čárový kód druhu. Druhou možností je nejprve vyhledat požadovanou sekvenci pomocí názvu druhu, popřípadě ještě NCBI kódu viz předchozí podkapitola a použít jako vstup výstup předchozí funkce Hledaná sekvence. Při chybném zadání vstupních dat uživateli opět napomáhá výpisová řádka.

Převody zprostředkovávají funkce `reprezentace2D.m` a `reprezetace3D.m`. Kromě grafického výstupu, numerického formátu sekvence molekuly DNA, program tuto matici uloží i do souboru `reprezentace2D.mat` popřípadě `reprezentace3D.mat`. Díky tomuto uložení je uživateli umožněno tyto data zpracovávat i mimo tento vytvořený program.



Obr. 44: Výstup 2D numerického formátu



Obr. 45: Výstup 3D numerického formátu

Zobrazené matice nadále slouží jako vstupní data pro podobnostní analýzu distanční maticí.

6.5 Analýza podobnosti sekvencí distanční maticí

Tuto analýzu zpracovává poslední blok grafického programu, který v této práci zatím nebyl popsán.

Obr. 46: Blok srovnávání distanční maticí

Na vstup tohoto bloku jsou přiváděny sekvence v numerických formátech viz předchozí kapitola. Na pozadí bloku pracuje funkce `distancni_matice.m`, která byla popsáná v podkapitole 5.6. Jelikož výstupem funkce není míra podobnosti dvou sekvencí, ale pásový vektor distanční matice jedné sekvence, musí si uživatel nejdříve do paměti uložit referenční pásový vektor od jednoho druhu.

To provede tak, že do numerického formátu převede sekvenci druhu, u kterého chce zkoumat míru podobnosti s ostatními druhy. Poté klikne na položku 2D referenční matice nebo 3D referenční matice, podle druhu zvoleného numerického formátu sekvence části molekuly DNA. Zmáčkne tlačítko Proveď a počká než mu výpisová řádka sdělí, že pásový vektor pro referenční druh byl uložen do paměti.

Až se tento úkon provede, převede uživatel do numerického formátu stejného rozměru sekvenci čárového kódu od dalšího druhu. Označí položku z 2D reprezentace nebo z 3D reprezentace, zmáčkne znovu tlačítko proveď a na výstupu bloku se mu zobrazí míra podobnosti mezi sekvencemi obou druhů.

Pokud chce uživatel zkoumat míru podobnosti dalšího druhu se stejným referenčním druhem, stačí, když převede další sekvenci do numerického formátu a znovu označí položku z 2D reprezentace nebo z 3D reprezentace, zmáčkne tlačítko proveď a program mu vypočítá další míru podobnosti sekvencí referenčního druhu a nově zkoumaného.

Program si hodnotu referenčního pásového vektoru pamatuje až do doby než ho uživatel nechá znovu přepočítat pro novou sekvenci.

Jako příklad byl uveden výpočet míry podobnosti sekvence tura domácího s člověkem pomocí reprezentace 3D numerickým formátem sekvencí.

7. Analýza dat

V závěru této práce bude provedena analýza čárových kódů vybraných druhů pomocí naprogramovaných algoritmů. Na Ústavu biomedicínského inženýrství je člověk v centru zájmu studia, z tohoto důvodu bylo vybráno k analýze několik příbuzných druhů a pak skupina živočichů, které pocházejí ze skupin nepříliš příbuzných živočišných tříd. K analýze byli vybráni z řádu primáti čeledi hominidů šimpanz učenlivý (pan troglodytes), šimpanz bonobo (pan paniscus), orangutan bornejský (pongo pygmaeus), orangutan sumaterský (pongo abelii), gorila nížinná (gorilla gorilla). Z řádu primátů čeledi kočkodanovití byl vybrán jeden zástupce pavián anubi (papio anubis). Z řádu hlodavci čeledi myšovití byla vybrána myš domácí (mus musculus), z řádů sudokopytníci čeledi turovití byl vybrán tur domácí (bos taurus). Jakožto největšího dárce orgánů a kožních štěpů v plastické chirurgii bylo z řádu sudokopytníci na seznam přidáno ještě prase domácí (sus domestica), které patří do čeledi prasovití. Všechny výše vypsané druhy jsou zástupci kmene strunatci, třídy savci.

Evoluční teorie předpokládají, že třídy savci a ptáci mají společné předky v třídě plazi. Aby mohl být tento předpoklad ověřen, bude provedena analýza člověka se třemi zástupci ptáků a třemi zástupci plazů. Ze třídy ptáci byli vybráni racek stříbřitý (larus argentatus), z řádu dlouhokřídlí čeledi rackovití, kur domácí (gallus gallus), z řádu hrabaví z čeledi bažantovití, a sojka obecná (garrulus glandarius), z řádu pěvci čeledi krkavcovití. Ze třídy plazi budou se savcem porovnáváni ještěrka zelená (lacerta viridis), z řádu šupinatí čeledi ještěrkovití, slepýš křehký (anguis fragilis), z řádu šupinatí čeledi slepýšovití, a chameleon jemenský (chamaeleo calyptratus), z řádu šupinatí čeledi chameleonovití.

Všechny výše uvedené druhy byli zástupci kmene strunatci živočišné říše, aby bylo možné vyvodit závěry o přesnosti naprogramovaných analýz i v rámci více kmenů, budou zkoumáni ještě tři druhoví zástupci z kmene členovci třídy hmyz. Budou to včela medonosná (apis mellifera), z řádu blanokřídlí čeledi včelovití, moucha domácí (musca domestica), z řádu dvoukřídlí čeledi mouchovití, a mravenec otročí (formica fusca), z řádu blanokřídlí čeledi mravencovití.

7.1 Srovnávání s genem člověka

Všechny výše vyjmenované živočišné druhy byly pomocí programu `porovnavani.m` zarovnány s čárovým kódem člověka Needleman-Wunschovým a Smith-Watermanovým algoritmem. Sekvence čárového kódu člověka byla čerpána z veřejné databáze NCBI, byla použita pro zarovnání všech druhů, měla délku 648 bází. Pro názornost je uvedena tato sekvence přímo v textu práce.

- Čárový kód člověka používaný v N.-W. a S.-W. algoritmech:

```
ctattattcggcgcatgagctggagtcctaggcacagctctaagcctccttattcgagccgagctgggccagcc
aggcaaccttctaggtaacgaccacatctacaacgttatcgtcacagcccatgcatttgtaataatcttcttca
tagtaatacccatcataatcggaggctttggcaactgactagttcccctaataatcggtgcccccgatatggcg
tttccccgcataaacaacataagcttctgactcttacctccctctctcctactcctgctcgcatctgctatagt
ggaggccggagcaggaacaggttgaacagtctaccctcccttagcagggaactactcccaccctggagcctccg
tagacctaaccatcttctccttacacctagcaggtgtctcctctatcttaggggccatcaatttcatcacaaca
attatcaatataaaaccccctgccataacccaataccaaacgcccctcttcgtctgatccgtcctaatcacagc
agtcctacttctcctatctctcccagtcctagctgctggcatcactatactactaacagaccgcaacctcaaca
ccaccttcttcgaccccgccggaggaggagaccccattctataccaacacctattc
```

Výsledky tohoto zarovnání jsou uvedeny v tabulce, kde N_1 je počet bází sekvence zkoumaného druhu před zarovnáním, N_2 je počet bází sekvence zkoumaného druhu po zarovnání, NE je počet neshod mezi oběma zarovnanými sekvencemi a δ_r je relativní počet neshod na zarovnanou sekvenci.

První tabulka vyjadřuje výsledky zarovnání Needleman-Wunschovým algoritmem a druhá Smith-Watermanovým.

název druhu (česky)	název druhu (latinsky)	kmen	třída	řád	N_1/bp	N_2/bp	NE/bp	δ_r/%
šimpanz učenlivý	pan troglodytes	strunatci	savci	primáti	648	648	66	10,19
šimpanz bonobo	pan paniscus	strunatci	savci	primáti	600	648	112	17,28
orangutan bornejský	pongo pygmaeus	strunatci	savci	primáti	883	885	315	35,59
orangutan sumaterský	pongo abelii	strunatci	savci	primáti	883	884	305	34,5
gorila nížinná	gorilla gorilla	strunatci	savci	primáti	642	649	75	11,56
pavián anubi	papio anubis	strunatci	savci	primáti	621	653	148	22,66
myš domácí	mus musculus	strunatci	savci	hlodavci	648	658	141	21,43
tur domácí	bos taurus	strunatci	savci	sudokopytníci	648	657	125	19,03
prase domácí	sus domestica	strunatci	savci	sudokopytníci	648	652	138	21,17
racek stříbřitý	larus argentatus	strunatci	ptáci	dlouhokřídlí	686	702	169	24,07
kur domácí	gallus gallus	strunatci	ptáci	hrabaví	648	664	135	20,33
sojka obecná	garrulus glandarius	strunatci	ptáci	pěvci	686	699	180	25,75
ještěrka zelená	lacerta viridis	strunatci	plazi	šupinatí	883	894	380	42,51
slepýš křehký	anguis fragilis	strunatci	plazi	šupinatí	648	662	147	22,21
chameleon jemenský	chamaeleo calyptratus	strunatci	plazi	šupinatí	648	672	199	29,61
včela medonosná	apis mellifera	členovci	hmyz	blanokřídlí	651	687	247	35,95
moucha domácí	musca domestica	členovci	hmyz	dvoukřídlí	721	793	389	49,05
mravenec otročí	formica fusca	členovci	hmyz	blanokřídlí	654	692	240	34,68

Tab. 1: Tabulka výsledků pořízených N.-W. a.

název druhu (česky)	název druhu (latinsky)	kmen	třída	řád	N_1/bp	N_2/bp	NE/bp	δ_r/%
šimpanz učenlivý	pan troglodytes	strunatci	savci	primáti	648	648	66	10,19
šimpanz bonobo	pan paniscus	strunatci	savci	primáti	600	600	65	10,83
orangutan bornejský	pongo pygmaeus	strunatci	savci	primáti	883	650	95	14,62
orangutan sumaterský	pongo abelii	strunatci	savci	primáti	883	649	80	12,33
gorila nížinná	gorilla gorilla	strunatci	savci	primáti	642	643	69	10,73
pavián anubi	papio anubis	strunatci	savci	primáti	621	625	122	19,52
myš domácí	mus musculus	strunatci	savci	hlodavci	648	658	141	21,43
tur domácí	bos taurus	strunatci	savci	sudokopytníci	648	657	125	19,03
prase domácí	sus domestica	strunatci	savci	sudokopytníci	648	652	138	21,17
racek stříbřitý	larus argentatus	strunatci	ptáci	dlouhokřídlí	686	658	131	19,91
kur domácí	gallus gallus	strunatci	ptáci	hrabaví	648	658	129	19,6
sojka obecná	garrulus glandarius	strunatci	ptáci	pěvci	686	656	138	21,04
ještěrka zelená	lacerta viridis	strunatci	plazi	šupinatí	883	661	159	24,05
slepýš křehký	anguis fragilis	strunatci	plazi	šupinatí	648	662	147	22,21
chameleon jemenský	chamaeleo calyptratus	strunatci	plazi	šupinatí	648	671	198	29,51
včela medonosná	apis mellifera	členovci	hmyz	blanokřídlí	651	679	240	35,35
moucha domácí	musca domestica	členovci	hmyz	dvoukřídlí	721	546	163	29,85
mravenec otročí	formica fusca	členovci	hmyz	blanokřídlí	654	681	231	33,92

Tab. 2: Tabulka výsledků pořízených S.-W. a.

Z tabulek je vidět, že výsledky procentuální podobnosti sekvencí čárového kódu vyšly ve třinácti z osmnácti případů téměř totožně a volba algoritmu zarovnání neměla na výsledek vliv. Pak ale nastalo pět případů, kdy Needleman-Wunschův algoritmus selhal a začal podávat nesměrodatné výsledky. Podrobnější zdůvodnění a vysvětlení tohoto jevu bude uvedeno v závěru práce.

7.2 Počítání podobnosti čárových kódů distanční maticí

I pro tuto analýzu byly použity stejné živočišné druhy jako v předešlé podkapitole. Postup provedení pomocí grafické nástavby programu byl popsán v podkapitolách 6.4 a 6.5, kdy nejprve byl vyhledán čárový kód člověka, který byl převeden do 2D numerického formátu, z něj byl vypočítán referenční pásový vektor pomocí distanční matice. Poté byl stejný postup opakován pro referenční pásový vektor z 3D numerického formátu. Pro rychlý průběh získávání dat, bylo tedy využito faktu, že referenční pásové vektory zůstávají uloženy v paměti.

V dalším kroku byl pokaždé vyhledán čárový kód určitého druhu, převeden do 2D numerického

formátu, byla vypočítána míra podobnosti čárových kódů. Následně hned poté byl tentýž čárový kód převeden do 3D numerického formátu, byla vypočítána druhá míra podobnosti, výsledky byly zapsány do tabulky a přešlo se na další zkoumaný druh.

název druhu (česky)	název druhu (latinsky)	kmen	třída	řád	P_{2D}	P_{3D}	N/bp
šimpanz učenlivý	pan troglodytes	strunatci	savci	primáti	14,29	34,18	648
šimpanz bonobo	pan paniscus	strunatci	savci	primáti	46,53	23,62	600
orangutan bornejský	pongo pygmaeus	strunatci	savci	primáti	100,24	101,88	883
orangutan sumaterský	pongo abelii	strunatci	savci	primáti	64,44	79,6	883
gorila nížinná	gorilla gorilla	strunatci	savci	primáti	90,48	83,1	642
pavián anubi	papio anubis	strunatci	savci	primáti	86,4	35,31	621
myš domácí	mus musculus	strunatci	savci	hlodavci	378,28	644,74	648
tur domácí	bos taurus	strunatci	savci	sudokopytníci	218,8	327,31	648
prase domácí	sus domestica	strunatci	savci	sudokopytníci	258,01	267,42	648
racek stříbřitý	larus argentatus	strunatci	ptáci	dlouhokřídlí	76,43	78,61	686
kur domácí	gallus gallus	strunatci	ptáci	hrabaví	13,16	35,46	648
sojka obecná	garrulus glandarius	strunatci	ptáci	pěvci	108,06	172,62	686
ještěrka zelená	lacerta viridis	strunatci	plazi	šupinatí	222,4	71,53	883
slepýš křehký	anguis fragilis	strunatci	plazi	šupinatí	85,03	279,83	648
chameleon jemenský	chamaeleo calyptratus	strunatci	plazi	šupinatí	282,45	482,88	648
včela medonosná	apis mellifera	členovci	hmyz	blanokřídlí	829,54	727,41	651
moucha domácí	musca domestica	členovci	hmyz	dvoukřídlí	637,28	354,87	721
mravenec otročí	formica fusca	členovci	hmyz	blanokřídlí	754,61	483,13	654

Tab. 3: Tabulka vypočtených měr podobností distanční maticí

Míra podobnosti je takové bezrozměrné číslo, které čím nabývá menší hodnoty, tím označuje danou dvojici čárových kódů živočišných druhů za podobnější. Kromě měr podobností byla do tabulky přidána vždy ještě délka sekvence daného druhu, která byla porovnávána s čárovým kódem člověka o délce 648 bp.

Při pohledu na vypočítané výsledky v tabulce je vidět značný rozptyl ve výsledcích mezi zástupci jednotlivých zkoumaných skupin. Co tato čísla ve skutečnosti znamenají je vysvětleno v závěru bakalářské práce.

8. Závěr

V průběhu vypracovávání prvních čtyř kapitol této práce bylo zjištěno, že od roku 2004 existuje metoda, která každému živočišnému druhu přiřadí specifický úsek genu CO1, uloženého na mitochondriálním genomu. Tato metoda se začala v praxi využívat a po schválení základních pravidel a regulí začaly první živočišné druhy získávat sekvence párů bází s označením barcode flag. Využití mitochondriální DNA má své výhody i nevýhody. Největší nevýhodou této metody je bezesporu fakt, že se dá aplikovat pouze na živočichy. Pro rostliny ani houby prozatím nebyl objeven gen natolik specifický, aby plnil funkci čárového kódu. Oproti tomu nesmírnou výhodou je obrovské množství kopií mitochondriální DNA v každé buňce živočicha, toto bude nesmírně cenný fakt v momentě, kdy se začne sestrojovat čtečka čárových kódů. Izolací a vytěsněním sekvence čárového kódu v praxi přímo z buňky se ale tato práce nezabývala. Úkolem této práce bylo nasimulovat tento proces pouze virtuálně na procesoru počítače.

V páté kapitole se práce posunula od teorie k praktické aplikaci a tvorbě výpočetních algoritmů. Prvním velkým úkolem bylo vytvořit funkce, které budou získávat data z veřejně přístupných genových databází. Konkrétně bylo za úkol čerpat informace z databází NCBI a BOLD Systems. Pro dolování dat z databáze NCBI byly napsány funkce `barcode.m`, se kterou pracuje funkce `knihovna.m`. I přes snahu ošetření podmínkami všech možných situací, které mohou nastat při hledání způsobu a místa uložení sekvence čárového kódu v databázi, se ukázalo, že ne vždy se při stahování dat uživatel shledá s pozitivním výsledkem. Nevýhoda této databáze se ukázala v její neřízenosti a možnosti ukládat do ní v podstatě téměř cokoliv a kdykoliv. U spousty druhů najdete pomocí vyhledávače tamější stránky, že se hledaný mitochondriální genom na stránce nachází, poté ho uživatel i nalezne. V momentě, kdy ale uživatel začne stahovat data pomocí funkcí vytvořených pro matlab, zjistí, že u sekvence genomu nejsou uloženy všechny potřebné informace (např. které geny sekvence obsahuje, že nemá uloženy počáteční a koncové pořadí báze genu) a kvůli nedostatku těchto informací následně program není schopen požadované informace z databáze získat. Další velkou nevýhodou této databáze jako zdroje informací byl poměrně omezený počet druhů, které mají v databázi uloženy mitochondriální genom, z převážné části více jak 50% uložených mitochondriálních genomů se jedná o mitochondriální genom člověka. Jako výborný zdroj automaticky stahovaných informací se ovšem tato databáze ukázala pro ukládání taxonomické klasifikace druhu. Jelikož je to jediná a ne zcela podstatná výhoda databáze nebyla NCBI shledána za dostatečný zdroj informací.

Tento fakt byl jeden z hlavních důvodů hledání dalšího zdroje požadovaných informací pro vytvoření zkušební knihovny čárových kódů. Jako výborný další zdroj se zdála být databáze BOLD Systems podporována pracovní skupinou CBOL, neboť se přímo zabývá získáváním čárových kódů přímo ze vzorků a ukládání jich v elektronické podobě na internetovou síť. Po seznámení se s touto databází byla vytvořena funkce `getcbol.m`, která získává informace o názvu druhu a jeho čárovém kódu ze staženého .txt souboru uloženého na pevném disku uživatelova počítače. I když se tento postup zdá na první pohled poměrně složitý, jeho výhodou zůstává fakt, že BOLD Systems umožňují uložení do .txt souboru i větší množství sekvencí a názvů druhů najednou. Toto číslo nemusí být v žádném případě malé. Databáze a funkce `getcbol.m` umožňují uživateli několika málo kliknutími rozšířit svoji knihovnu čárových kódů o jeden až tisíce nových druhů nebo sekvencí. BOLD Systems je řízená a spravovaná databáze, takže se data v ní vyskytují pouze v jednom uceleném formátu. Díky tomu při programování nenastal téměř žádný výskyt chyby ze strany databáze, pouze jednou se v sekvenci čárového kódu objevil znak W. Tento problém byl vyřešen pár řádky zdrojového kódu navíc. Nevýhodou se ukázalo, že při stahování sekvencí, se uživateli do .txt souboru uloží pouze informace o druhu a o sekvenci. Txt soubor neobsahuje žádné další bližší informace o primerech, které byly potřebné k vytěsnění sekvence z mitochondriálního genomu a dalších informacích, které jsou potřebné k získání statusu barcode flag u sekvence. Všechny tyto informace jako jsou primery, trace files, obrazová dokumentace druhu, lokace

výskytu, taxonomická klasifikace databáze obsahuje. Jsou jen uloženy pod jinými odkazy a bohužel v rámci této bakalářské práce nebyly vytvořeny algoritmy k přepracování do formátu, který by byly pro matlab kompatibilní. Ze všech těchto uvedených faktů vyplývá, že BOLD Systems je mnohem prospěšnější databází pro tvorbu knihovny čárových kódu pro programovací prostředí matlab než databáze NCBI. Jen pokud bych v závěru směl uvést jako autor práce jednu osobní výtku k databázi. Jako celek mi databáze přišla a stále se mi zdá jako značně nepřehledná a neintuitivní, na neznalého uživatele se vyvalí spousta tabulek, grafů, odkazů a téměř žádný vysvětlující text. Jen najít obyčejný vyhledávač mi trvalo pěkných pár minut. To je ale jen nepodstatná osobní výtka, oproti faktu, že tato databáze je bezesporu spolehlivý zdroj získaných čárových kódů, které uživatel hledá. I přes mírné nedostatky, výhody této databáze převažují.

Data získaná z obou databází jsou uložena a sloučena v souboru `carovy_kod.mat`. Tento soubor představuje knihovnu informací o čárových kódech a živočišných druzích, se kterou matlab umí pracovat. Pokud již byl druh do knihovny uložen, obsahuje knihovna informace alespoň o jedné sekvenci čárového, kódu, jeho délce a pokud byl čárový kód získán z databáze NCBI, obsahuje ještě NCBI kód uložené sekvence a taxonomickou klasifikaci druhu. Jak bylo uvedeno v předchozích odstavcích. Tyto informace ještě nestačí k udělení statusu barcode flag v podobě v jaké jsou uloženy v matlabovém formátu.

Další algoritmy, které byly v rámci bakalářské práce vytvořeny, se již nezabývaly získáváním dat, ale jejich matematickou analýzou. Princip jejich funkce a chod byl podrobně popsán a vysvětlen přímo v kapitole 5. Ke zhodnocení jejich funkčnosti a použitelnosti, byla vytvořena kapitola 7, ze které bude hlavně vycházet následující zbytek závěru.

Prvním úkolem analýzy bylo srovnávání čárových kódů živočišných druhů s čárovým kódem člověka. Na tento úkol byly vytvořeny dva algoritmy Needleman-Wunschův a Smith-Watermanův. Pohlédnete-li na naměřená data získaná výpočtem těchto algoritmů na vzorku vybraných živočišných druhů, zjistíte, že algoritmy podávají ve většině případů téměř identické výsledky. Tento fakt je způsoben použitím totožné skórovací matice u obou algoritmů. Pak ale nastaly případy, ve kterých se ukázal Needleman-Wunschův algoritmus jako naprosto nevhodný nástroj analýzy. Připomeňme si, že sekvence čárového kódu člověka, se kterou byly čárové kódy živočišných druhů zarovnávány, měla délku 648 bází. Jak je z výsledků vidět, při zarovnávání sekvencí mnohem delších než je použitý čárový kód člověka mnohonásobně roste počet neshod v sekvenci a tento fakt poté následně neprávem říká, že jsou si oba druhy mnohem méně podobné než si ve skutečnosti opravdu jsou. Důvod je prostý, Needleman-Wunschův algoritmus vytváří globální zarovnání obou sekvencí a z velmi rozdílných délek sekvencí vyplývá i obrovské množství vložených mezer, které poté program bere jako neshody a započítá je do procentuálního výskytu neshod. Neznamená to, ale že algoritmus v těchto případech zarovnává špatně, grafický výstup tohoto zarovnání je naprosto v pořádku, pouze je nepraktický k výpočtu relativní podobnosti obou sekvencí. Tato chyba se začne projevovat i pokud globálně zarovnáme výrazně kratší sekvenci než je délka 648 bází u čárového kódu člověka. Srovnáte-li výsledky relativních podobností těchto růžově označených sekvencí s relativní podobností stejných dvojic sekvencí zarovnaných lokálně Smith-Watermanovým algoritmem, uvidíte, že relativní podobnost se na tomto rozdílu délek liší průměrně o 20%, což už je velmi významná chyba. Z těchto výsledků vyplývá, že při stažení výrazně delší nebo kratší sekvence než průměrně 650 bází je v tomto programu výhodnější použít globálního Smith-Watermanova zarovnání.

Pohlédnete-li na naměřená data podruhé a namísto délky zarovnávaných sekvencí, si začnete všímat přímo živočišných druhů, kterým čárové kódy patří, uvidíte očekávaný fakt, že nejpodobnější čárový kód máme se šimpanzem. Odborná literatura o čárových kódech od P. Herberta a D. N. Stoecklea uvádí, že se naše čárové kódy liší řádově v 60 bázích. Z tohoto pohledu se dá považovat označený počet neshod 66 u obou algoritmů se šimpanzem učenlivým za poměrně přesný. Dále tato literatura uváděla, že s gorilou nížinnou se náš čárový kód liší řádově v 70 bázích.

I v tomto případě se programové algoritmy chovaly podle očekávání, v N.-W. případě vypočítaly 75 neshod a v případě S.-W. algoritmu 69 neshod.

Z výsledku S.-W. algoritmu pro řád hominidů se dá usuzovat, že čárový kód těchto zástupců se liší řádově v 10-15% od čárového kódu člověka. Je zajímavé, že zástupce řádu kočkodanů, kteří patří také mezi primáty, se od člověka v sekvenci čárového kódu liší procentuelně přibližně totožně jako ostatní zkoumaní savci. Dále je vidět, že se nevyplnil předpoklad, že by se čárový kód člověka měl více shodovat s čárovými kódy plazů než ptáků. Výsledky vyšly přesně opačně, i když rozdíly v procentech nejsou nijak markantní, je-li pominut výsledek u chameleona. Všichni tito uvedení zástupci, mimo řád hominidů a chameleona, se lišili průměrně ve 20% bází. Při zarovnání čárových kódů vybraného hmyzu, se rozdíly zvýšily na více než 30%.

V další části bakalářské práce měl být vytvořen vyhledávač živočišných druhů podle uživatelem zadané sekvence, tato funkce byla vytvořena a zprovozněna. Podle autorova názoru je mnohem přínosnější funkcí vyhledávání čárových kódů podle zadaného názvu druhu. Z tohoto důvodu byla vytvořena i možnost tohoto vyhledávání.

Posledním blokem bakalářské práce bylo vytvoření algoritmů pro převod čárových kódů do numerických formátů, pomocí kterých měla být počítána míra podobností dvou sekvencí využitím distanční matice a pásových vektorů. Úspěšnost tohoto snažení a získané výsledky jsou popsány v následujících odstavcích.

Při zpětném pohledu na naměřená data v tabulce musíme v první řadě věnovat pozornost primátům. Zde metoda správně určila šimpanze učenlivého jako geneticky nejbližšího čárovému kódu člověka pomocí 2D reprezentace čárových kódů. Pomocí 3D reprezentace hodnota vyšla poněkud větší, ale ve srovnání s výsledkem u šimpanze bonobo, který je nám také velmi příbuzný, je rozdíl jen nepatrný. Srovnáme-li míry podobností šimpanzů u obou zástupců s orangutany a gorilou, pak vychází, že byli opravdu určeni jako geneticky nejbližší předkové ve všech případech. Zarážející jsou ale výsledky u orangutana bornejského, který byl výpočtem označen jako druh poměrně vzdálený od ostatních primátů. Nejprve se zdálo, že tato zdánlivá chyba byla způsobena rozdílnou délkou porovnávaných sekvencí. V dalších výpočtech se ukázalo, že délka sekvence asi nemá na výsledek markantní vliv. Za zmínku stojí ještě výsledek podobnosti člověka s paviánem anubim, který je podle 3D reprezentace téměř nejbližší genetický příbuzný i přes fakt, že patří do odlišné čeledi primátů.

Při srovnání výsledku míry podobnosti ostatních savců s člověkem v tabulce zjistíme, že je nám myš geneticky více vzdálena než sudokopytníci podle 2D numerického formátu sekvencí, čemuž by se dalo i věřit, při pohledu na 3D numerickou reprezentaci ale ihned vidíme, že nám myš má být více geneticky vzdálena než někteří vybraní zástupci hmyzu.

U výsledků podobností ptáků a plazů je vidět, že i zde se nachází spousta podezřelých čísel u obou způsobů numerických reprezentací. Nejzvláštnější výsledek bezesporu je podobnost čárového kódu kura domácího s čárovým kódem člověka, který je mu podle míry podobnosti více podobný než čárový kód řady primátů u obou numerických reprezentací.

Při shrnutí všech doposud získaných výsledků analýzy podobnosti distanční maticí a počítání průměrů jejich pásových vektorů, vyjde jeden jediný závěr. Tato metoda se ukázala jako velmi nepřesná pro zkoumání podobností sekvencí molekuly DNA na genu CO1 mitochondriálního genomu. Jedinou skupinu druhů, kterou poměrně spolehlivě odlišila na měřeném vzorku byl hmyz od obratlovců a to pouze v případě, že zanedbáme výsledky vypočítané u myši domácí.

Nemusí to nutně znamenat, že metoda porovnání podobností distanční maticí je špatná metoda. Její výsledky značně ovlivňuje typ převodu sekvence na numerický formát. Je možné, že při volbě jiného numerického převodu by analýza nabyla naprosto jiných výsledků.

Při srovnání všech použitých metod určování podobnosti čárového kódu s čárových kódem

člověka se jako nespolehlivější ukázal Smith-Watermanův algoritmus, který podával nejpřesnější výsledky ve všech směrech při jakýkoliv sekvencích mitochondriální DNA na vstupech.

Úkolem práce bylo seznámit se a zpracovat srozumitelnou rešerši o metodě udělování čárového kódu živočišným druhům, protože o tomto problému prozatím neexistuje téměř žádná česká literatura, zabývá se touto problematikou poměrně značná část práce. Po dostatečném teoretickém zpracování nastaly další body práce týkající se konkrétní praktické aplikace. Po dohodě s vedoucím práce se rozhodlo vytvořit několik porovnávacích algoritmů s cílem zjistit, zda budou vhodné pro aplikaci porovnávání u této metody katalogizace živočišné říše. Ukázalo se, že ne všechny již v praxi používané druhy algoritmů a zpracování se dají úspěšně použít i na tuto metodu zpracování genetické informace. Nezbývá než doufat, že se podařilo všechny problémy vysvětlit a předvést výstižně a i pro čtenáře ne zcela znalého veškeré teorie srozumitelně.

V posledním bodě závěru by měla být zhodnoceno sloučení linuxové verze Matlabu, operačního systému Linux s operačním systémem Windows a verzí Matlabu naistalovaném na tomto operačním systému. Z našeho sledování a testování se nepotvrdilo, že by tato změna uživatelského prostředí měla nějaký vliv na funkčnost jednotlivých naprogramovaných funkcí. Při spuštění grafické podoby našeho programu se ale již objevily mírné nesrovnalosti s rozložením jednotlivých grafických komponent na ploše. Dále se ukázalo, že grafické prostředí GUI zřejmě ještě nemá všechny vnitřní knihovny plně sloučené s konzolí Matlab, neboť při spuštění grafického prostředí programu přímo z příkazové řádky nefungují základní příkazy, které by v rámci skriptu nebo funkce normálně běžně fungovaly. Tento problém se vyskytl jak při spuštění na operačním systému Linux, tak i při spuštění na operačním systému Windows. Není to nedostatek jen programu v grafickém prostředí GUI v rámci této bakalářské práce, tento nedostatek sloučení grafického editoru GUI s konzolí Matlab provázel většinu studentů v průběhu několika cvičení, ve kterých vytvářeli jiné grafické programy. Jediné prozatímní řešení spočívá v otevření zdrojového kódu grafického programu (.m soubor) a spuštění programu z textového editoru přes zelený trojúhelník nebo přes otevření .fig souboru přes grafický editor a následné spuštění programu přes zelený trojúhelník. Při tomto postupu se knihovny prostředí GUI a konzole Matlab plně sloučí a program funguje.

9. Seznam použité literatury

[1] GLENN, Martina. ArtMuseum.cz : Pravěk [online]. Konečná verze. 1999-2009 , 21. 5. 2009 [cit. 2009-10-25]. Windows-1250. Text v češtině. Dostupný z WWW: <http://www.artmuseum.cz/smery_list.php?smer_id=124>.

[2] DOLEŽAL, Mgr. Tomáš PhD.. GENETICKÁ DIVERZITA. NEODARWINISMUS : Rozluštění chemické struktury DNA [online]. Upravené vydání. České Budějovice [ČR] : University of South Bohemia, Faculty of Biological Sciences, , 2008 , Leden 2009 [cit. 2009-10-27]. Text v češtině. Dostupný z WWW: <http://apendix.bf.jcu.cz/Dolezal/vyuka/dna/DNA.htm>.

[3] ALBERTS, Bruce, et al. Základy buněčné biologie : Úvod do molekulární bilogie. 2. vyd. Translation Prof. RNDr. Arnošt Kotyk, DrSc. Ústí nad Labem : Espero publishing, 1998. ISBN 80-902906-2-0. Získávání energie v mitochondriích a chloroplastech, s. 407-430.

[4] MOYNA, Guillermo. Lecture 36 : Nucleotides and nucleic acids [online]. 1999 [cit. 2009-12-12]. Dostupný z WWW: <http://tonga.usip.edu/gmoyna/biochem341/lecture36.html>.

[5] LANG, B. Franz, et al. A Comparative Genomics Approach to the Evolution of Eukaryotes and their Mitochondria. In J. Eukaryot. Microbiol.. [s.l.] : [s.n.], 1999. s. 320-326.

[6] SNUSTAD, D. Peter, SIMMONS, Michael J. Principles of Genetics. 2nd edition. [s.l.] : [s.n.], 1999. The molecular genetics of mitochondria, s. 503-507.

[7] EFENBERK, Aleš. MIMOJADERNÁ DĚDIČNOST U ČLOVĚKA. [s.l.], 2008. 32 s. MASARYKOVA UNIVERZITA; Přírodovědecká fakulta; Ústav experimentální biologie; Oddělení genetiky a molekulární biologie. Vedoucí bakalářské práce prof. RNDr. Jiřina Relichová, CSc.

[8] HERBERT, Paul. D. N., STOECKLE, Mark Y. Čárový kód života. Scientific American. 2008, č. listopad 2008, s. 24-29. české vydání.

[9] HEBERT, Paul D. N., STOECKLE, Mark Y. BARCODE OF LIFE. Business Source Complete : Scientific American [online]. 2008 [cit. 2009-10-30], s. 82-88. Anglické vydání. Dostupný z WWW: <http://web.ebscohost.com/ehost/detail?vid=1&hid=111&sid=bb39e15e-8e49-4502-b703-0fba2890b0f0%40sessionmgr113&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=bth&AN=34236720>. ISSN 0036-8733.

[10] Barcode of life : National Center for Biotechnology Information, US National Library of Medicine [online]. [cit. 2009-10-30]. Text v angličtině. Dostupný z WWW: <http://www.ncbi.nlm.nih.gov/Genbank/barcode.html>.

[11] DNA barcoding of animal and plant species as an approach for their molecular identification and describing of diversity [online]. 2009 Jul-Aug [cit. 2009-10-30]. Text v angličtině. Dostupný z WWW: <http://www.ncbi.nlm.nih.gov/pubmed/19799325?itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum&ordinalpos=6>.

[12] Consortium for the barcode of Life [online]. Washington [USA] : CBOL, May 2004 [cit. 2009-10-30]. Text v angličtině. Zdroj první části práce hlavně pdf soubory přiložené na úvodní straně. Dostupný z WWW: <http://www.barcoding.si.edu/>.

[13] DESALLE, R., EGAN, MG., SIDDALL, M.. The unholy trinity: : taxonomy, species delimitation and DNA barcoding [online]. Londýn [GB] : PubMed, 2005 Oct [cit. 2009-10-30]. Text v angličtině.Dostupný z WWW: <http://www.ncbi.nlm.nih.gov/pubmed/16214748?ordinalpos=1&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_SingleItemSupl.Pubmed_Discovery_RA&linkpos=5&log$=relatedreviews&logdbfrom=pubmed>.

[14] FRÉZAL, L., LEBLOIS, R.. Four years of DNA barcoding: : current advances and prospects [online]. Paříž [Francie] : PubMed, 2008 Sep , 2008 Jun 3 [cit. 2009-10-30]. Text v angličtině. Dostupný z WWW: <http://www.ncbi.nlm.nih.gov/pubmed/18573351?itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum&ordinalpos=16>.

[15] HANNER, Robert, DWG, CBOL. Proposed Standards for BARCODE Records in INSDC (BRIs) [online]. November 6, 2005 , 19 December 2005 [cit. 2009-12-01]. Text v angličtině. Link na originál konečného standardu pro barcode. Dostupný z WWW: <http://www.barcoding.si.edu/PDF/DWG_data_standards-Final.pdf>.

[16] Genbankread [online]. 1984-2009 [cit. 2009-12-23]. Dostupný z WWW: <http://www.mathworks.fr/access/helpdesk/help/toolbox/bioinfo/ref/genbankread.html>.

[17] Needleman–Wunsch algorithm. In Wikipedia : the free encyclopedia [online]. St. Petersburg (Florida) : Wikipedia Foundation, 21.09.2004, last modified on 13.05.2010 [cit. 2010-05-19]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Needleman–Wunsch_algorithm>.

[18] Merlin.fit.vutbr.cz [online]. datum vydani neuvedeno [cit. 2010-05-19]. DNA projekt. Dostupné z WWW: <http://merlin.fit.vutbr.cz/DNA/>.

[19] MELOUN, Radek. Identifikace podobných proteinových struktur s odlišnou konektivitou [online]. Brno : Masarykova univerzita, 2006. 35 s. Bakalářská práce. Masarykova univerzita, Fakulta informatiky. Dostupné z WWW: <http://is.muni.cz/th/98895/fi_b/bp.pdf>.

[20] MAĎERÁNKOVÁ, Ing. Denisa. Deterministické metody ve zpracování genomických dat. Brno, 2010. 27 s. Pojednání. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství.

10. Seznam zkratek a příloh

10.1 Seznam zkratek

- DNA – deoxyribonukleová kyselina
- mtDNA – mitochondriální deoxyribonukleová kyselina
- mRNA – mediátorová ribonukleová kyselina
- tRNA – transferová ribonukleová kyselina
- rRNA – ribozomální ribonukleová kyselina
- ATP – adenosintrifosfát
- ADP – adenosindifosfát
- CBOL – The Consortium for the Barcode of Life
- DWG – Database Working Group
- INSDC – International Nucleotide Sequence Database Collaboration
- NCBI – National Center for Biotechnology Information
- BRIs – nahrávky čárových kódů v INDSC
- BoLD – University of Guelph's Barcode of Life Database
- ABBI – All Birds Barcoding Initiative
- FISH-BOL – Fish Barcode of Life
- MBI – Mosquito Barcoding Initiative
- N.-W. a. – Needleman-Wunschův algoritmus
- S.-W. a. – Smith-Watermanův algoritmus

10.2 Seznam příloh

- CD s elektronickou verzí bakalářské práce, kompletním vytvořeným programem, zdrojovou textovou formou informací čerpaných z databází NCBI a BOLD Systems, obrázky použitými v textu, elektronickou formou ostatních příloh
- Papírová forma zdrojových kódů
- Manuál k softwaru