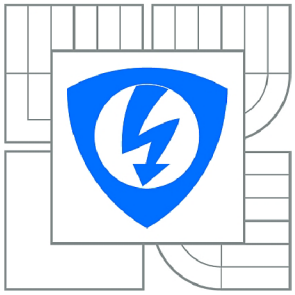
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

MOLEKULÁRNÍ TAXONOMIE POMOCÍ MITOCHONDRIÁLNÍ DNA

MITOCHONDRIAL DNA FOR MOLECULAR TAXONOMY

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

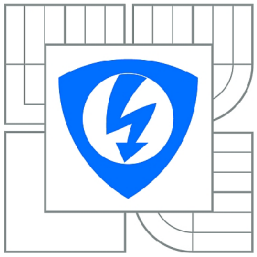
AUTOR PRÁCE — KATEŘINA LESÁKOVÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2013



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Studentka: Kateřina Lesáková ***ID:*** 138944
Ročník: 3 ***Akademický rok:*** 2012/2013

NÁZEV TÉMATU:

Molekulární taxonomie pomocí mitochondriální DNA

POKYNY PRO VYPRACOVÁNÍ:

1) Zpracujte literární rešerši na téma metody pro molekulární taxonomii, DNA barcoding a numerické metody reprezentace genomických sekvencí. 2) Vyberte numerickou reprezentaci, která by mohla být vhodná pro distanční porovnávání mitochondriálních sekvencí, a tuto metodu programově implementujte. 3) Implementovanou metodu vyzkoušejte na vybraném souboru sekvencí z databáze BOLD. 4) Pro vybranou třídu organismů vytvořte referenční numerické sekvence alespoň dvaceti druhů. 5) Proveďte klasifikaci souboru sekvencí pomocí distančního porovnávání s referenčními numerickými sekvencemi. Výsledky diskutujte.

DOPORUČENÁ LITERATURA:

[1] SCHWARTZ, Jeffrey. Molecular Systematics and Evolution. Encyclopedia of Molecular Cell Biology and Molecular Medicine, 2nd ed. Wiley: Weinheim, 2005, 696 s., ISBN: 3-527-30550-2.
[2] BLAXTER, Mark. The promise of a DNA taxonomy. Phil. Trans. R. Soc. Lond. B., 2004, roč. 359, s. 669-679.

Termín zadání: 11.2.2013 ***Termín odevzdání:*** 31.5.2013

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti bakalářské práce:

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

UPOZORNĚNÍ:

ABSTRAKT

Tato bakalářská práce se zabývá molekulární taxonomií pomocí mitochondriální DNA. V úvodu je popsána taxonomie obecná a molekulární. V teoretické části je dále popsána struktura a složení živočišné buňky, deoxyribonukleové kyseliny a mitochondriální deoxyribonukleové kyseliny. Další část obsahuje informace o DNA barcodingu a numerických metodách reprezentace genomických sekvencí. V praktické části je popsán program zvolené numerické metody pro zpracování genomických sekvencí a programy pro tvorbu a přiřazování sekvencí referenčním druhům.

KLÍČOVÁ SLOVA

Molekulární taxonomie, mitochondriální DNA, DNA barcoding, numerické reprezentace

ABSTRACT

This work deals with mitochondrial DNA and molecular taxonomy. Structure and composition of animal cell, deoxyribonucleic acids and mitochondrial ribonucleic acids are described in the introduction. Another part contains information of DNA barcoding and numerical representation of genomic sequences. Programs are described in the practical part.

KEYWORDS

Molecular taxonomy, mitochondrial DNA, DNA barcoding, numerical representation

LESÁKOVÁ, K. *Molekulární taxonomie pomocí mitochondriální DNA.* Brno: FEKT VUT v Brně 2013. 42 s. Vedoucí bakalářské práce Ing. Denisa Maděránková.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Molekulární taxonomie pomocí mitochondriální DNA“ jsem vypracovala 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 bakalářské práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědoma následků porušení ustanovení §11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestně právních důsledků vyplývajících z ustanovení §152 trestního zákona č. 140/1961 Sb.

Brno . . . .

(podpis autora)

PODĚKOVÁNÍ

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

V Brně dne . . . .

(podpis autora)

Obsah

SEZNAM OBRÁZKŮ

ZDROJE OBRÁZKŮ

Obrázek 1: (http://schoolworkhelper.net/scientific-taxonomy)
Obrázek 2: (http://www.animalport.com/img/Animal-Cell.jpg)
Obrázek 3: (http://patf-biokyb.lf1.cuni.cz/wiki/_media/wiki/user/komponenty_obrazky/mitochondrie.png)
Obrázek 4 : (http: //applejacksgirl.blogspot.cz/2012/04/nucleotides.html, http://www.bionet-skola.com/w/Nuk lein ske_kiseline)
Obrázek 5: (http://ehrig-privat.de/ueg/dna_-_structure.htm)
Obrázek 6 : (http://cs.wikipedia.org/wiki/Mitochondri%C3%A1ln%C3%AD_DNA)
Obrázek 7: http://robotika.cz/articles/gerda/pcr-expansion.png)
Obrázek 8: (http://sitfu.com/2010/10/every-species-to-get-own-dna-barcode-international-barcode-of-life/)
Obrázek 9: (Cristea, Conversion of nucleotides sequences into genomic signals,Bio-Medical Engineering Center)
Obrázek 10, 11: (Fenglan Bai, Tianming Wang. A 2-D graphical representation of protein sequences based. Chemical Physics Letters.)
Obrázek 12: (Randic, Milan, Graphical representations of DNA as 2-D map. Chemical Physics Letters)
Obrázek 13: (Randic, Milan.On characterization of DNA primary sequences by a condensed.Chemical Physics Letters)
Obrázek 14: (Tomaž Pisanski, Jure Zupan, Milan Randic. On representation of DNA by line distance matrix. Journal of Mathematical Chemistry)
Obrázek 15: (Nella Lers...,Four-color map representation of DNA or RNA sequences. Chemical Physics Letters.)
Obrázek 16: (Xiaofeng Guo, On the Characterization of DNA Primary Sequences by Triplet of Nucleic Acid Bases. J. Chem. Inf. Comput. Sci.)
Obrázek 17: (Jun Wang, Yi Zhang. Characterization and similarity analysis of DNA sequences grounded. Chemical Physics Letters.)

SEZNAM TABULEK

Úvod

Taxonomie představuje hierarchický systém, do kterého můžeme zařadit dle určitých znaků živočichy a rostliny. Tento systém se vytvářel pouze pomocí určitých strukturálních znaků, jako jsou velikost, barva … atd. V případě, kdy z organismu máme k dispozici jen kousek tkáně, musíme hledat jiné způsoby identifikace organismů.

Právě molekulární taxonomie při třídění organismů do hierarchických skupin využívá molekulárními znaky, pod kterými rozumíme znaky uložené v sekvencích informačních makromolekul, jako jsou DNA, RNA a proteiny.

Systém, který se o toto třídění snaží, se nazývá DNA barcoding, který v sobě zahrnuje samotné vytváření databází živočichů, ale také síť organizací zabývající se touto problematikou. Standardní metody analýzy DNA sekvencí pracují s jejich symbolickou reprezentací. Alternativně se používají numerické metody reprezentace genomických dat. [1], [2]

Cílem této práce je seznámit se s molekulární taxonomií, buněčnou biologií, DNA barcodingem a numerickými metodami zpracování genomických sekvencí. Z těchto metod najít nejvhodnější pro zpracování použitých dat a programově tuto metodu realizovat. Pomocí zvolené metody vytvořit v programovém prostředí referenční numerické sekvence dvaceti druhů a poté provést klasifikaci souboru s databáze BOLD pomocí distančního porovnání s referenčními sekvencemi.

1 TEORETICKÁ ČÁST

1.1 Taxonomie

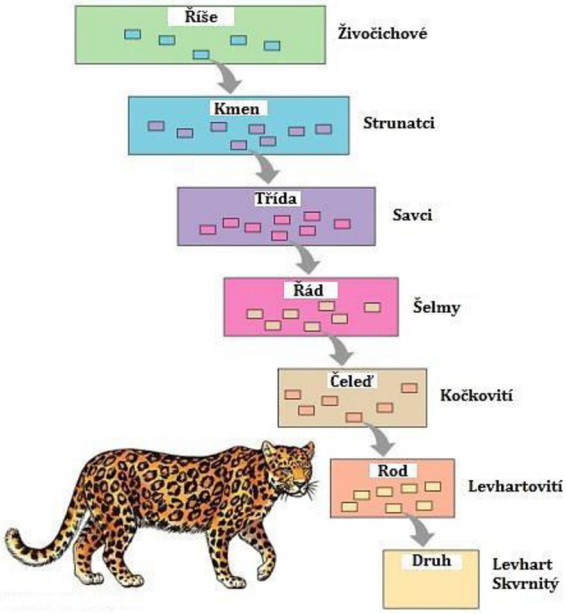
Taxonomie je věda studující taxony, jejich rozeznávání, ohraničování, klasifikaci, vnitřní strukturu, vznik a historický vývoj.

Taxon je jakýkoli přirozený a rozlišitelný soubor žijících i vymřelých organismů, který je do té míry vyhraněný, že jej přijímáme i jako jednotku klasifikace. Je to systematická jednotka tvořená skupinou organismů, pro které je společný určitý stupeň příbuznosti, a které se svými znaky odlišují od ostatních systematických jednotek stejné úrovně. Čím je taxon nižší, tím podobnější organismy v sobě zahrnuje. Jednotlivé taxony jsou hierarchicky seřazeny do kategorií. Kategorie jsou seřazeny následovně: říše, kmen, oddělení, třída, řád, čeleď, rod a druh. Jména druhů jsou dvouslovná, jsou složena z rodového jména a druhového přívlastku. Této formě říkáme binomická nomenklatura. Vymezení taxonů vyšších než druh a vytvoření jejich klasifikace je v současné době spojeno s fylogenetikou, která se snaží o rekonstrukci historického vývoje organismů a o zjištění jejich příbuznosti. Taxonomie je významná z hlediska manipulace, ochrany, bezpečného ověřování a opakovaného pozorování určitého druhu. Protože každý druh má svou specifickou ekologii, etologii a genetické vlastnosti. [1]

Znak je jakákoli vlastnost organismu. Daná situace znaku se označuje jako stav. Stavy mohu být odlišné kvalitativně (např. přítomnost/nepřítomnost), meristicky (počtem prvků) nebo spojitě kvantitativně (např. délka těla). Neklasifikujeme znaky, ale organismy s plným souborem jejich znaků včetně těch, které se projevují jen v určitých etapách ontogeneze, tzn. vývoji jedince od zárodečného vývoje do jeho zániku. Holomorfologie je stav, kdy organismus nemusí vykazovat určitý znak v dané etapě vývoje. Zdrojem znaků mohou být údaje strukturální (morfologické), ontogenetické a morfometrické (rozměry a jejich poměry), cytotaxonomické (stavba chromozomů) a další. V našem případě nás budou zajímat zdroje molekulárně-biologické, jako jsou údaje o molekulární struktuře genomu a jeho složení. [1]

Strukturální znaky jsou nejčastěji používány díky snadnější dostupnosti informací. Zvláštní skupinou je skupina znaků makromolekulárních, poskytovaná studiem nukleových kyselin (DNA, RNA) i jejich bezprostředních produktů – proteinů. Tyto znaky mohou být získány metodami přímými (např. mapování genů, sekvenování nukleotidů v DNA a RNA, aminokyselin v proteinech) a nepřímými (např. zjišťování imunologické vzdálenosti různých organismů). Všechny tyto znaky jsou v diagnostické i evolučně chápané taxonomii, pro fylogenetiku mají však mnohem větší význam znaky získané metodami přímými. Jen tyto znaky jsou plně srovnatelné se znaky získanými z oborů nemolekulárních. [1]

Při použití znaku makromolekulárních přetrvávají dva základní problémy. Jak dalece můžeme se sekvenčními údaji manipulovat, protože určitá míra kalibrace molekulárních dat je nezbytná. A jak spojit molekulární data s údaji z tradičních zdrojů, protože početnost molekulárních dat by mohla mnohonásobně převýšit počet znaků pocházejících z jiných zdrojů. [1], [2]



Obrázek 1: Příklad zařazení Levharta skvrnitého do taxonomického sytému

1.2 Molekulární taxonomie

Molekulární taxonomie při třídění organismů do hierarchických skupin využívá molekulárními znaky, pod kterými rozumíme znaky uložené v sekvencích informačních makromolekul, jako jsou DNA, RNA a proteiny. Nejdůležitější je primární struktura DNA, protože právě zde dochází k novinkám ve formě mutací a informace zde obsažená se předává do dalších generací podle jasných pravidel. Primární sekvence DNA tedy v sobě nese informaci o historii daného lokusu DNA nebo proteinu. Daný lokus (pozice, kterou na chromozomu zaujímá jeden nebo více genů) nám může prozradit identitu jedince a jeho příbuznost s ostatními jedinci v populaci, druhovou příslušnost nebo příbuzenský vztah tohoto druhu.

Molekulární fylogenetika, je fylogenetika založená na molekulárních znacích, zabývá se rekonstrukcí fylogeneze (příbuzenských vztahů), neboť všechny organismy prošli v minulosti určitou fylogenezí, snaží se tedy vystopovat pořadí větvení taxonů (kladogeneze) a všímá si taky vývoje vlastností v rámci linií. [1]

Existují dva základní přístupy používání molekulárních znaků pro fylogenetickou rekonstrukci. První fenetický přístup používá při porovnání organismů všechny znaky. Na tomto přístupu je založena numerická taxonomie, kde se analyzuje podobnost mezi

organismy. Druhý kladistický přístup pro rekonstrukci fylogeneze používá převážně synapomorfie, což jsou sdílené znaky vyskytující se alespoň u dvou skupin organismů.

Základní taxon by měl být monofyletický, což znamená, že jeho členové si jsou vzájemně příbuzní více, než je kdokoli z nich příbuzný druhu mimo tento taxon neboli, že zahrnuje všechny potomky jednoho předka. Uznávaný je také parafyletický taxon, který zahrnuje společného předka a některé jeho potomky (např. plazi). Nepřípustný je polyfyletický taxon, který nezahrnuje všechny potomky společného předka a nezahrnuje ani tohoto předka, protože ten neměl znaky, které by ho opravňovaly do tohoto taxonu patřit. Snaha je, aby klasifikace organizmů byla v souladu s jejich fylogenezí. Znalost fylogeneze nějaké skupiny nám ukazuje, které taxony nesmíme vytvářet a které naopak můžeme. [3]

Teorie neutrální evoluce vysvětluje velké množství polymorfismů. Podle ní drtivá většina znaků neovlivňuje fenotyp, a je proto "neviditelná" pro přírodní výběr. Tyto znaky jsou selekčně neutrální. Frekvence těchto znaků neboli alel (konkrétní formy genu) je v populaci ovlivňována výhradně náhodou tzv. genetickým driftem (posun ve frekvenci jednotlivých alel v populaci). Ten je způsoben tím, že do následující generace se dostane konečný počet náhodně vybraných alel z rodičovské populace. Frekvence alel během generací tedy náhodně fluktuuje, dokud se jednou nesníží na 0, kdy mluvíme o vymizení alely, nebo nezvýší na 100, kdy se jedná o její fixaci. Genetický drift funguje rychleji v malých populacích. V takových populacích proto vliv driftu převáží i nad působením přírodního výběru v případě mírně selekčně výhodných nebo mírně selekčně nevýhodných mutací. [3]

1.2.1 Molekulární znaky

Molekulární znaky jsou, jak již bylo zmíněno, především molekuly DNA, RNA a proteinů. Tyto znaky mají oproti ostatním znakům sloužících pro tvoření taxonomie svoje výhody i nevýhody.

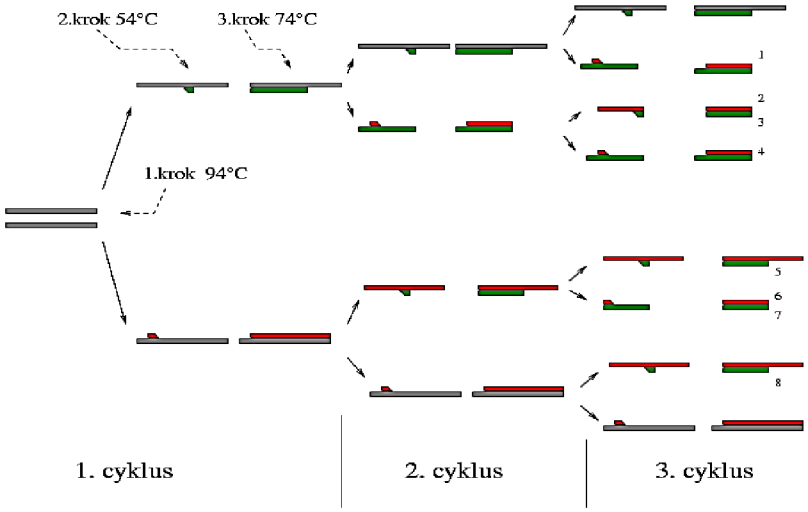
Mezi výhody můžeme zahrnout to, že tyto znaky pocházejí z úrovně, kde vznikají evoluční novinky neboli mutace, většinou známe jejich dědičnost, jsou nezávislé na prostředí i samy na sobě a často jsou selekčně neutrální. Znaků je obrovské množství a jsou použitelné na všech úrovních taxonomie. Další výhodou je možnost je jednoznačně popsat, protože nabývají několika diskrétních stavů, jsou v podstatě digitální. Jsou lépe vážitelné a lépe se zde kvantifikuje stupeň nejistoty. Vybrané znaky můžeme s opatrností použít jako molekulární hodiny pro odhad stáří.

Nevýhodou molekulárních znaků je, že neposkytují informaci o anagenezi (změna vlastností taxonu nebo vývojové linie během fylogeneze), tedy že většina molekulárních znaků se vůbec neprojeví na fenotypu. Financování získávání molekulárních znaků je stále nákladnější než získávání morfologických znaků. Poslední nevýhodou je občasné nenávratné zničení organismu nebo jeho části při získávání dat. [3]

Sekvence DNA je základní úrovní, kterou molekulární taxonomie studuje, protože tyto znaky vznikají mutacemi DNA. Klasickou metodou sekvenace DNA je Sangerova dideoxy metoda. Je to speciální forma PCR s jedním primerem a fluorescenčně značenými dideoxynukleotidy (ddNTP, ddATP, ddGTP, ddCTP). Je nepostradatelná, pokud potřebujeme osekvenovat jeden určitý úsek DNA. Pro tuto metodu je nejprve potřeba připravit úsek DNA v mnoha identických kopiích. K tomu se používá metoda PCR (polymerázová řetězová reakce). [3]

Amplifikace znamená zesílení, v případě DNA hovoříme o zmnožení. Polymerázová řetězová reakce je biochemická reakce, která využívá enzym DNA-polymerázu ke kopírování DNA takovým způsobem, že se produkt hromadí geometrickou řadou.

Princip spočívá v tom, že enzym DNA-polymeráza syntetizuje komplementární (doplňující) vlákno podle templátu jednovláknové DNA tak, že přidává k existujícímu úseku druhého vlákna nové nukleotidy ve směru 5' > 3'. K tomu ještě potřebuje krátký úsek druhého vlákna, tzv. primer, který bude za vhodných teplotních podmínek tvořit vodíkové můstky s komplementární sekvencí v templátovém vlákně, hovoříme o nasedání primeru. Takto určíme DNA-polymeráze, od kterého místa a kterým směrem má začít syntetizovat komplementární vlákno. Extenze primeru znamená jeho prodlužování přidáváním dalších nukleotidů na 3'-konci. Takto vznikne pouze jedna kopie. V PCR je zajištěno množení kopií použitím dvou primerů, které nasedají na komplementární sekvence ve dvou templátových vláknech. Templátová vlákna získáme denaturací původně dvouvláknové DNA. Primery na vlákna nasedají v protisměrné orientaci, takže po opakovaných cyklech denaturace, nasedání primeru a extenze primeru DNA-polymerázou vznikají další produkty sloužící jako templáty pro nový reakční cyklus. Máme-li tedy na začátku k dispozici dvě templátová vlákna, v prvním cyklu vzniknou dvě kopie, pro další cyklus máme k dispozici čtyři vlákna, podle kterých vzniknou další čtyři kopie. Celkem osm templátových vláken slouží v dalším cyklu k syntéze dalších osmi vláken, takže se produkt hromadí geometrickou řadou. [4]

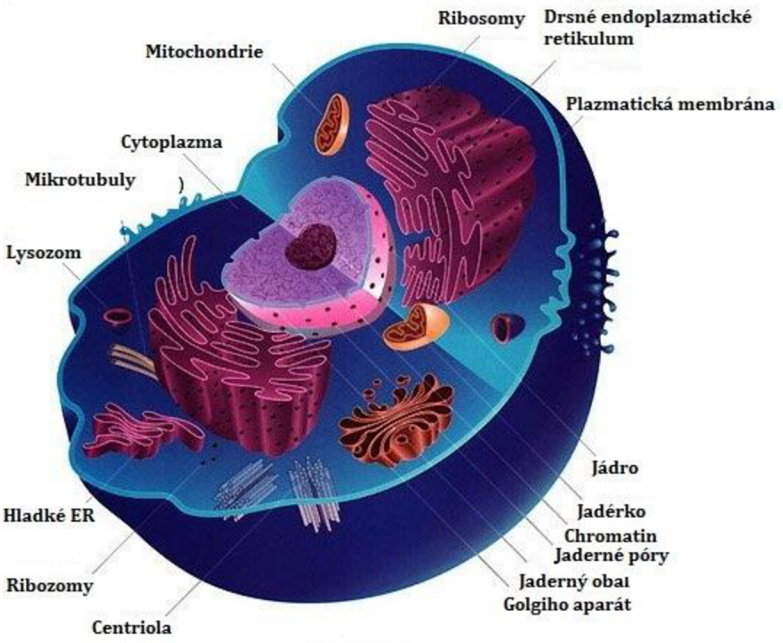


Obrázek 2: Schéma polymerázové řetězové reakce

1.3 Biologický základ

1.3.1 Živočišná buňka

Buňka je nejjednodušší a nejmenší známý útvar schopný samostatné existence a rozmnožování, je to základní stavební a funkční jednotka všech živých organismů. Každá buňka má svůj vlastní genetický aparát, proteosyntetický aparát a metabolický systém umožňující vytvářet a využívat energii nutnou pro její existenci. Velikost eukaryotické buňky (buňky s diferencovaným jádrem a s biomembránovými strukturami) se pohybuje přibližně v rozmezí 10-100 μm. [1], [2]



Obrázek 3: Živočišná buňka

Buňka je ohraničena cytoplazmatickou membránou, která izoluje vnitřní prostředí buňky od prostředí vnějšího. Je tvořena dvojvrstvou fosfolipidů, které jsou uspořádány tak, že hydrofobní konce mastných kyselin směřují k sobě a fosfátové hydrofilní části směřují od sebe směrem ven z buňky. Ve fosfolipidové dvojvrstvě jsou přítomny molekuly bílkovin, které jsou buď z části, nebo úplně zanořeny do dvojvrstvy. Do hlavních funkcí buňky zařazujeme výměnu látek, především živin a metabolitů mezi buňkou a prostředím, pomocí přenašečů, integrálních proteinů a proteinových komplexů. [2]

Cytoplazma vyplňuje celý vnitřní obsah buňky. Je to viskózní koncentrovaný roztok, který obsahuje molekuly organických i anorganických látek. Často může obsahovat kapénky nebo krystalky odpadních látek (buněčné inkluze). Buněčné jádro je od okolí cytoplazmy ohraničeno dvojitou jadernou membránou s jadernými póry, které umožňují komunikaci mezi jádrem a cytoplazmou. Vnitřek jádra je vyplněn polotekutou hmotou (karyoplazmou), v nichž se nacházejí vláknité útvary tzv. chromozomy obsahující deoxyribonukleovou kyselinu (DNA). V jádře se často nachází jedno nebo více jadérek. [2]

Endoplazmatické retikulum je membránový systém plochých váčků a kanálků, který navazuje na obal jádra pomocí membrány. Může být drsné, kdy jsou na membránách přítomny ribozomy, nebo hladké bez ribozomů, kde se především syntetizují glykolipidy. Ribozomy jsou bílkovinná tělíska obsahující ribozomovou DNA (r-RNA), můžou být buď volné, nebo mohou být navázané na endoplazmatické retikulum. Obsahují dvě nestejné podjednotky a jejich hlavní funkcí je proteosyntéza. Golgiho aparát je soustava měchýřků propojených kanálky, ve kterých probíhají biochemické reakce upravující látky, které se vytvořily v endoplazmatickém retikulu. [1]

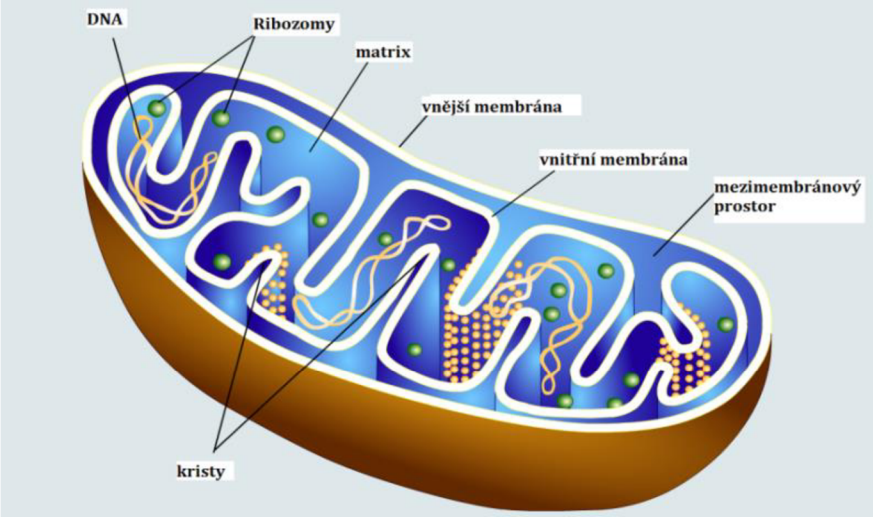
Cytoskelet je systém tvořící kostru buňky. Skládá se z vlákének - mikrofilament a trubiček – mikrotubulů, které v buňce tvoří svazky, které se mohou zkracovat a prodlužovat a tak umožňují pohyb struktur uvnitř buňky. Součástí cytoskeletu je i dělící vřeténko, důležité při buněčném dělení. [2]

1.3.2 Mitochondrie

Mitochondrie jsou malé organely, řádově mají jeden μm, a může jich být v buňce několik set. Mají hladkou vnější membránu, která je mimořádně dobře propustná pro polární látky. Vnitřní membrána je tvořeny záhyby neboli kristy. Je bohatě osazena transmembránovými komplexy dýchacího řetězce, syntázy ATP (složitý transmembránový bílkovinný komplex, schopný syntézy ATP) a membránovými přenašeči. Uvnitř je mitochondriální matrix, obsahující enzymy aerobních metabolických drah. Při buněčném dýchání se uvolňuje v mitochondriích energie zabezpečující životní děje v buňce. [1]

Mitochondrie patří mezi semiautonomní organely symbiotického původu. Jejich autonomie spočívá v tom, že se mohou v buňce rozmnožovat dělením, ale nově vznikat nemohou. Mají vlastní kružnicovou DNA, která je umístěna v mitochondriální matrix a vlastní ribozomy vyznačující se vlastní proteosyntézou. Jejich DNA kóduje jen menší část proteinů organely. Většina proteinů je kódována jadernou DNA a syntetizována v cytosolu.

Předpokládá se, že mitochondrie vznikla vsunutím K-fialové bakterie do vnitřku buňky při jejím vývoji. Během tohoto vývoje až do podoby dnešních eukaryotických buněk tento endosymbiont (organismus, který žije uvnitř těla nebo buněk dalšího organismu) převedl mnohé ze svých základních genů do jaderných chromozomů. Nicméně mitochondrie stále nese charakteristické znaky svého bakteriálního předchůdce. Bylo zjištěno, že u savců je mtDNA přenášena pouze přes samičí zárodečnou linii. V savčích spermiích je počet kopií mtDNA nízký, naopak u savčích oocytů (samičí pohlavní buňka) je počet kopií extrémně vysoký. Detailní morfologická studie ukázala, že mitochondrie spermií jsou převedeny na vajíčka během fertilizace (spojení vajíčka aspermie) a následně jsou ztraceny brzy během embryogeneze (zárodečný vývoj). Mechanismus tohoto odstranění je zatím neznámý. [1], [5]



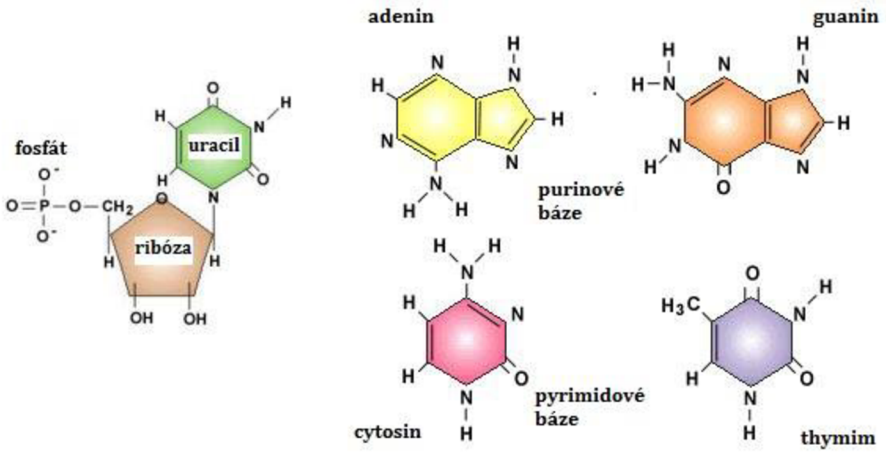
Obrázek 4: Mitochondrie

1.3.3 Deoxyribonukleová kyselina – DNA

DNA je obsažena v jádrech všech eukaryotických buněk. Je základním genetickým materiálem většiny organismů a slouží k uchování genetické informace. DNA je polynukleotid, tzn. je tvořena řetězci vzájemně spojených desítek až milionů nukleotidů. Nukleotid se u DNA skládá z pěti-uhlíkového monosacharidu - pentózy, konkrétně 2deoxy-D-ribózy, dále dusíkatých bází adeninu, guaninu, které podle molekulární struktury patří do skupiny purinů a bází cytosinu a tyminu, které patří do pyrimidů.

Dále tyto báze můžeme dělit podle biochemických vlastností, a to nejdříve podle síly vazby mezi nukleotidy, kdy adenin a tymin spojují dva vodíkové můstky, cytozin a guanin spojují vodíkové můstky tři. Jako poslední faktor pro jejich klasifikaci slouží obsažený radikál. Adenin a cytosin obsahují aminoskupinu NH_3, tymin a guanin obsahují keto skupinu C=O. Na obrázku 5 můžeme vidět jejich chemické schéma.

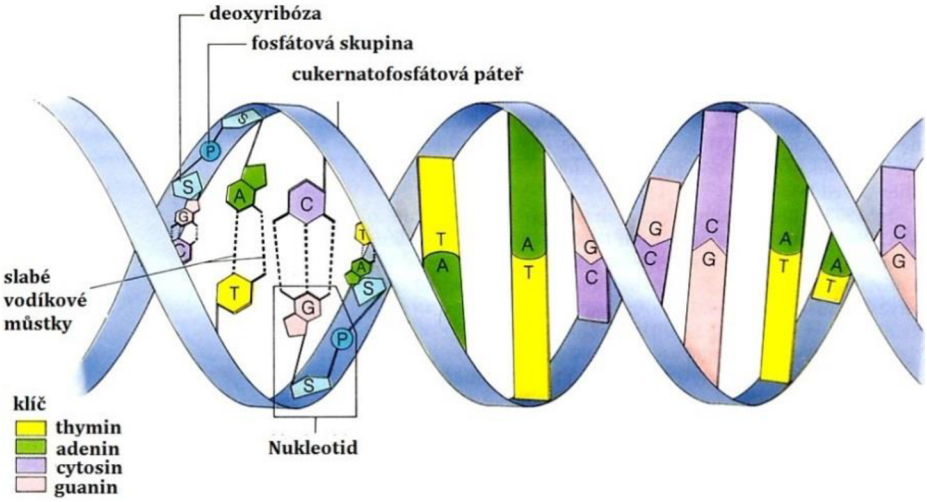
Poslední složkou nukleotidu je zbytek kyseliny trihydrogenfosforečné. Nukleotid je sloučenina nukleozidu s kyselinou fosforečnou. Nukleozidem se označuje sloučenina, která vznikla spojením purinové nebo pyramidové báze s deoxyribózou pomocí N-glykozidové vazby. Jednotlivé nukleotidy jsou spojeny fosfodiesterovými vazbami mezi sacharidy a fosfáty. Spojení probíhá tak, že na 3´-OH skupinu nukleotidu se váže 5´- fosfátová skupina nukleotidu následujícího. V důsledku toho mají pospojované řetězce polaritu, konec řetězce končící OH skupinou sacharidu se označuje jako 3‘ (tři s čárkou konec) a konec končící fosfátovou skupinou jako 5‘. [1] [6] [7]



Obrázek 5: Schéma nukleotidů

Pořadí jednotlivých nukleotidů spojených v polynukleotidový řetězec je tzv. primární struktura nukleových kyselin (NK) určující genetickou informaci organismu. Sekundární struktura NK je dána prostorovým uspořádáním polynukleotidového řetězce. Dochází ke spojení vláken dvoušroubovice pomocí vodíkových vazeb mezi bázemi obou řetězců. Báze, mezi kterými vzniká vazba, označujeme jako komplementární. K vazbě mezi určitou dvojicí bází tedy dochází na základě komplementarity. U DNA je cytosin komplementární s guaninem, ty jsou vzájemně propojeny třemi vodíkovými vazbami, naopak adenin je komplementární s tyminem a spojují je dvě vodíkové vazby, jak již bylo zmíněno dříve. Tyto páry se vytvářejí uvnitř struktury šroubovice. Molekula DNA je tvořena dvěma polynukleotidovými řetězci stočenými do pravotočivé šroubovice. Oba řetězce jsou vzájemně komplementární a mají opačnou polaritu. V jednom řetězci jsou uspořádány ve směru 5´-3´a ve druhém řetězci 3´-5´. Je to tak zvané antiparalelní uspořádání. [6] [7]

Na jeden závit dvoušroubovice připadá 10,5 párů bází, což odpovídá úseku dlouhému 3,4 nm. Vzdálenost mezi dvěma sousedními páry je 0,34 nm. Vnější část šroubovice tvoří opornou strukturu a představuje páteř DNA, jejíž nevětší vzdálenost od osy šroubovice je 1 nm. Terciální strukturu DNA označujeme jako nadšroubovici neboli superhelix poté, co se zavede do dvoušroubovité DNA další vinutí. [1]



Obrázek 6: Schéma deoxyribonukleové kyseliny

1.3.4 Přenos genetické informace

Na základě dogmatu molekulární biologie je přenos genetické informace možný pouze z nukleové kyseliny do nukleové kyseliny nebo do proteinu, ale přenos z proteinu na nukleové kyseliny není možný. Způsoby přenosu genetické informace jsou následující.

Replikace je proces, kdy se tvoří kopie DNA nebo RNA tedy, že se informace přenáší z molekuly do molekuly stejného typu. Při DNA replikaci se rozplete dvoušroubovice, která slouží jako templát a podél každého řetězce se nasyntetizuje nové komplementární vlákno. Vzniknou tak dvě dceřiné molekuly DNA, které obsahují jedno mateřské vlákno a jedno nově nasyntetizované. Jsou zde zapotřebí různé enzymy, z nich nejdůležitější je DNA polymeráza. U RNA se rovnou vytváří nové vlákno za pomocí RNA replikázy. [2]

Transkripce je proces, při němž se genetická informace přepisuje z molekuly DNA do RNA za pomocí enzymu RNA polymerázy. DNA se rozplete podle jednoho řetězce je nasyntetizována RNA, která se odpojí a řetězce DNA se opět spojí. Vzniká tak transkript, který je komplemetární s DNA vláknem.

Translace je překlad genetické informace z pořadí nukleotidů, konkrétně RNA, do pořadí aminokyselin. Každá aminokyselina je na mRNA kódována trojicí nukleotidu neboli tripletem. Kodon, tj. pořadí nukleotidů v tripletu, je základní jednotkou genetického kódu. Genetický kód je systém pravidel, podle kterých jednotlivé kodony určují přiřazení aminokyselin do peptidového řetězce popřípadě začátek a konec jeho syntézy. Genetický kód je degenerovaný, což znamená, že daná aminokyselina je kódována více než jedním z celkových 64 kodonů. Bílkoviny poté vznikají z aminokyselin přinášenými molekulami tRNA, které nasedají podle komplementarity bázi na svých antikodonech na kodony mRNA a spojují se pomocí peptidových vazeb. [1], [6]

Gen je jednotka genetické informace. Obsahuje informaci o primární struktuře funkční molekuly translačního proteinu, nebo funkční molekuly produktu transkripce, která nepodléhá translaci. Strukturní gen obsahuje informaci o primární struktuře polypeptidu jako translačního produktu. Strukturní geny mohou být složené z exonů, kódujících úseků, a intronů, které podléhají posttranskripčnímu sestřihu vlákna RNA. Po sestřihu se exonové úseky spojí v jeden řetězec RNA, který vstupuje do procesu translace za tvorby primární struktury polypeptidu. [1], [2], [8], [6]

1.3.5 Mitochondriální DNA

Mitochondriální DNA (mtDNA) je relativně malá, bohatá a snadno izolovatelná molekula DNA a pro tyto vlastnosti se stala oblíbeným cílem prvních projektů sekvenování genomu a nukleotidové sekvence DNA. Sled nukleotidů z lidské mtDNA byla první doložená kompletní sekvence mitochondriálního genomu. Struktura a genové uspořádání mtDNA je mezi savci zachovávána.

Savčí mitochondriální genom je kruhová uzavřená dvoušroubovice DNA molekuly o 16,6 kb. Vlákna dvoušroubovice DNA mohou být odlišná na základě skladby G-T bází, které způsobují různou hustotu každého řetězce, který pak můžeme označit jako těžký nebo lehký. Většina informací je zakódována v těžkém řetězci (H) s geny pro dvě rRNA, pro čtrnáct tRNA a pro dvanáct polypeptidů. Lehký řetězec (L) obsahuje kódy pro osm tRNA a jediný polypeptid. Všech třináct bílkovin je součástí enzymového komplexu oxidativní fosforylace. Savčí mtDNA jen výjimečně vykazuje úspornost organizace. Genům chybí introny s výjimkou jedné regulační oblasti. Molekula rRNA a tRNA jsou neobvykle malé. Některé z genů kódujících proteiny se překrývají a část terminačních kodonů není v mnoha případech kódována, ale je generována post-transkripčně polyadenylací mtRNA. [5]

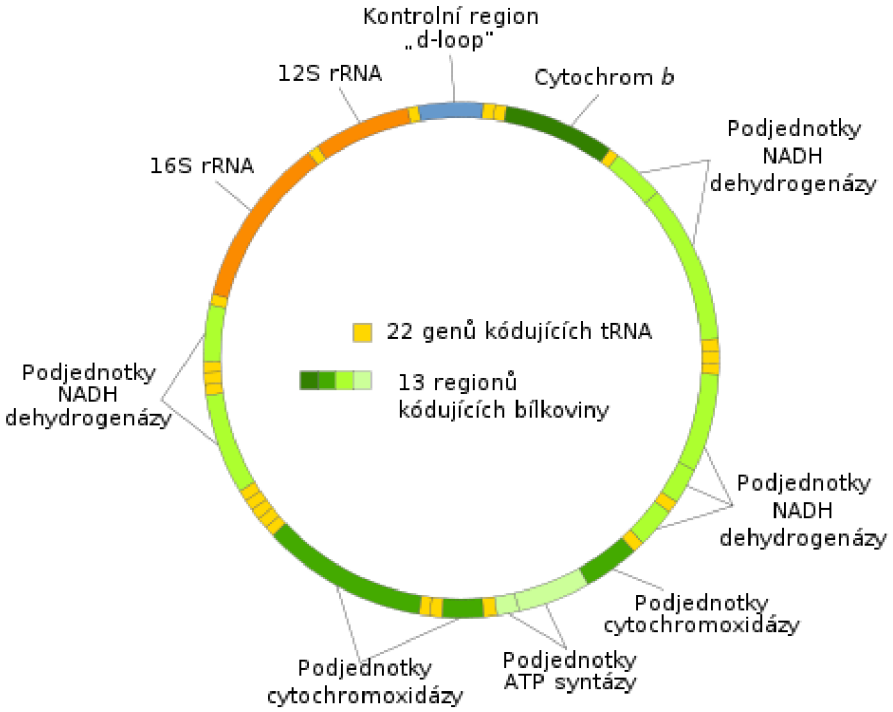
Brzy po získání mtDNA sekvencí bylo k dispozici srovnání s mitochondriálními proteinovými sekvencemi, kde byly odhaleny odchylky od standardního genetického kódu, později byly nalezeny rozdíly v používání kodonů jednotlivých druhů. Například kodon TGA v mtDNA je u většiny fylogenetických skupin používán jako kodon tryptofanu spíše než jako terminační kodon.

Dalším překvapivým rysem mitochondriálního genetického systému je jeho použití jako zjednodušený dekódující mechanismus, který umožňuje překlad všech kodonů s méně než dvaatřiceti druhy tRNA. Toto snížení množství rRNA je dosaženo použitím jediné tRNA s U v první antikodonové pozici, která rozpozná všechny kodony z čtyř-kodonové rodiny. Dvacet dva druhů tRNA je dostatečný počet pro kódování všech třinácti mitochondriálních genů pro syntézu proteinů.

U buněk obratlovců, které jsou metabolicky aktivní, velká část mtDNA obsahuje tří vláknovou strukturu, nazývanou jako posunutá smyčka nebo D-smyčka. V této struktuře se nachází nukleová kyselina, která je komplementární s L-vláknem a vytlačuje H-vlákno. Oblast D-smyčky je ohraničena geny pro tRNAPhe a tRNAPro, která se vyvinul jako hlavní kontrolní místo pro vyjádření mtDNA, obsahující místo počátku replikace a stimulátory pro transkripci. [5]

Mitochondrie nejsou samostatné subjekty v buňce. Replikace a transkripce závisí na jaderně kódovaných faktorech. U obratlovců jsou všechny mitochondriální ribozomální proteiny kódovány a syntetizovány mimo organelu, pomocí dodané aminokyselinové tRNA syntázy. Enzymy různých katabolických drah nacházejících se v mitochondriích

jsou také kódovány jadernou DNA. Dokonce i enzym pro oxidativní fosforylace je hybridního genetického původu. Všechny jaderně kódované polypeptidy určeny pro mitochondrie jsou syntetizovány na cytoplazmatických ribozomech, obvykle s oddělitelnou N-koncovou sekvencí pro mitochondriální cílení a jsou následně importovány do mitochondrie. [5]



Obrázek 7: Mitochondriální DNA

1.4 DNA Barcoding

Až dosud byly biologické vzorky identifikovány za pomoci morfologických znaků, jako je tvar, velikost či barva částí těla. V některých případech vyškolený technik mohl provádět běžné označení podle morfologických "klíčů", kde se postupovalo krok za krokem a existoval jakýsi univerzální návod na to, co hledat. Ale ve většině případů je k identifikaci potřeba zkušený profesionální taxonom. Je-li vzorek poškozen, nebo je-li v nezralém stádiu vývoje, může být problém i pro odborníka provést identifikaci. Metoda DNA barcodingu umí tyto problémy vyřešit.

Barcode v překladu znamená čárový kód, což nám napovídá, jak by tato metoda mohla vypadat. O DNA barcodingu se poprvé mohla vědecká komunita dozvědět v roce 2003, kdy výzkumná skupiny Paula Heberta z univerzity Guelph v Ontariu publikovala článek s názvem "Biologické identifikace prostřednictvím DNA barcodingu". Ten spočívá převážně v tom, že navrhuje nový systém určování živočišných a rostlinných druhů včetně objevování druhů nových organismů pomocí krátkého úseku DNA ze standardizované oblasti genomu. Tyto sekvence DNA mohou být použity k identifikaci různých druhů, podobně jako skener v supermarketech používá známé černé pruhy čárového kódu k identifikaci vašeho nákupu. To ale neznamená, že tradiční taxonomie pomocí morfologických znaků se stala méně důležitou. DNA barcoding spíše slouží k dvojímu účelu, a to jako nový nástroj v taxonomickém toolboxu rozšiřující její znalosti a jako inovativní zařízení pro laiky, kteří potřebují rychlou identifikaci druhů. [9]

Úsek genu, který je používán pro téměř všechny živočišné skupiny, je úsek dlouhý 648 párů bází a nachází se v mitochondriálním cytochromu c oxidasy genu 1 ("CO1"). Tento úsek se ukázal jako velmi účinný při určování ptáků, motýlů, ryb, much a mnoha dalších zvířecích skupin. Výhodou použití COI je, že je to sekvence dost krátká na to, aby mohla být rychle sekvenována a dost dlouhá na to, aby mohla sloužit k identifikaci variací mezi druhy. Barcode neboli čárový kód COI není efektivní pro identifikaci rostlin, protože se vyvíjí příliš pomalu, ale dvě genové oblasti v chloroplastu, matK a rbcL, byly schváleny jako barcodové oblasti pro suchozemské rostliny.

1.4.1 Komponenty DNA barcodingu

Projekty zabývající se barcodingem mají čtyři komponenty. Druhová identifikace pomocí DNA čárových kódů začíná vzorkem. Vzorky se získávají z různých zdrojů. Některé z nich jsou shromažďovány v určité oblasti, jiné pocházejí z rozsáhlých sbírek umístěných v přírodních historických muzeích, zoologických zahradách, botanických zahradách, semenných bankách, zmrazených tkáňových sbírkách, herbářích, akváriích a jiných úložištích biologických materiálů. [10]

Druhou komponentou jsou laboratoře, kde technici používají malý kousek tkáně ze vzorku pro získání jeho DNA. Barcode určité oblasti je izolován, tyto oblasti jsou replikovány pomocí procesu zvaného PCR amplifikace a následně jsou tyto oblasti sekvenovány. Sekvence je reprezentována sérií písmen CATG reprezentujících nukleové kyseliny - cytosin, adenin, tymin a guanin. Nejlépe vybavené laboratoře molekulární biologie mohou produkovat DNA barcodové sekvence během několika hodin. Tyto údaje jsou pak umístěny v databázi pro pozdější analýzu.

Databáze jsou jedním z nejdůležitějších prvků barcodingové iniciativy. Důležitá je výstavba veřejné referenční knihovny identifikátorů živočišných i rostlinných druhů, které by mohly být použity pro přiřazení neznámých vzorků tkání do známých druhů. V současnosti existují dvě hlavní barcodové databáze, které naplňují tuto roli. The International Nucleotide Sequence Database Collaborative je partnerskou databází genových databází GenBank v USA a the Nucleotide Sequence Database of theEuropean Molecular Biology Lab v Evropě a DNA Data Bank of Japan. Tyto databáze se dohodli na datových standardech CBOL pro barcodové záznamy.

Druhá databáze Barcode of Life Database (BOLD) byla vytvořena a je udržována University of Guelph v Ontariu. Tato databáze nabízí vědcům způsob, jak shromažďovat, spravovat a analyzovat data DNA barcodingu. BOLD slouží také jako úložiště pro barcodové záznamy, kde je možno čárové kódy vyhledávat, ukládat vzorky dat a obrázky, stejně jako sekvence a stopové soubory. Poskytuje identifikační motor založený na aktuálním čárovém kódu knihovny a monitoruje počet záznamů čárových kódů sekvencí a druhů zprostředkování. [10]

Poslední komponentou je analýza dat. Vzorky jsou označeny podle nejbližší nalezené shody s referenčním záznamem v databázi. CBOL's Data Analysis Working Group vytvořila the Barcode of Life Data Portal, který nabízí výzkumným pracovníkům nové a pružnější způsoby ukládání, správy, analýzy a zobrazení jejich barcodových dat. [11], [9]

Jako výzkumná iniciativa má DNA barcoding některé vlastnosti velkých, koordinovaných projektů jako je Human Genome a některé charakteristiky taxonomického výzkumu, který se tradičně skládá z individualistických projektů. Stejně jako u Human Genome je cílem DNA barcodingu výstavba obrovské, on-line a volně dostupné databáze sekvencí. Stejně jako taxonomický výzkum barcoding často provádějí výzkumní pracovníci, kteří se zaměřují na jednu taxonomickou skupinu v různých geografických regionech nebo rozmanitost druhů v jednom místě. Je to BARCODE datový standard, který umožňuje produktům individualistických projektů celého světa začlenit se do globální iniciativy. Účastníci v rámci této iniciativy DNA barcodingu přicházejí v mnoha konfiguracích, včetně konsorcií, databází, sítí, laboratoří a projektů, které se pohybují ve velikosti od místních po globální. [10]

Obrázek 8: Zobrazení mtDNA sekvence v podobě čárového kódu, kde jsou jednotlivé nukleotidy zobrazeny pomocí barevných pruhů

1.4.2 Konsorcia barcodingu

1.4.2.1 iBOL

The International Barcode of Life projekt (iBOL) je největší biodiverzitní genomická iniciativa, která kdy byla podniknuta. Stovky vědců biodiverzity, genomiky, techniků a etiků z 25 zemí společně pracují na výstavbě bohatě parametrizované referenční knihovny DNA čárových kódů, která bude základem pro identifikace systému pomocí DNA analýzy pro všechny formy buněčného života. Při stavbě knihovny, budou iBOL účastníci také budovat infrastrukturu potřebnou k použití v reálném světě - situacích, jako je zachování ekosystému, monitorování, forenzní analýzu a kontrolu zemědělských škůdců a invazních druhů. iBOL si klade za cíl vytvořit 5 000 000 záznamů čárových kódů z 500.000 druhů organismů v pěti letech. Deset pracovních skupin vyčleněných na různé taxonomické skupiny nebo typy přírodních stanovišť tvoří jádro této činnosti. [12], [10]

1.4.2.2 CBOL

Konsorcium pro Barcode of Life (CBOL) je mezinárodní iniciativa věnovaná rozvoji DNA barcodingu jako globálního standardu pro identifikaci biologických druhů. Založena byla v roce 2004 prostřednictvím podpory z Alfred P. Sloan Foundation. CBOL podporuje barcoding prostřednictvím pracovních skupin, sítí seminářů, konferencí, osvěty a školení, ale negeneruje žádná barcodová data. CBOL má 200 členských organizací z padesáti zemí a provozuje z úřadu sekretariát, který se nachází v Národním muzeu Smithsonian Institution of Natural History ve Washingtonu. [10], [13]

1.4.2.3 ECBOL

ECBOL je Evropské konsorcium pro Barcode of Life, bylo založeno jako součást úsilí výzkumné infrastruktury Evropského distribučního institutu taxonomie (EDIT). [10]

1.4.2.4 CCDB

CCDB neboli kanadské centrum pro DNA Barcoding na univerzitě v Guelph je síť laboratoří, je to největší "barcodingová továrna", generuje stovky tisíc datových záznamů za rok a vzdělává mnoho výzkumných pracovníků z celého světa. CBOL organizuje "Leading Labs Network ", vedoucí sít 15 laboratoří v 11 zemích, které poskytují technickou pomoc a výcvik sobě navzájem a novým pracovníkům. [10]

1.5 Numerické metody reprezentace genomických sekvencí

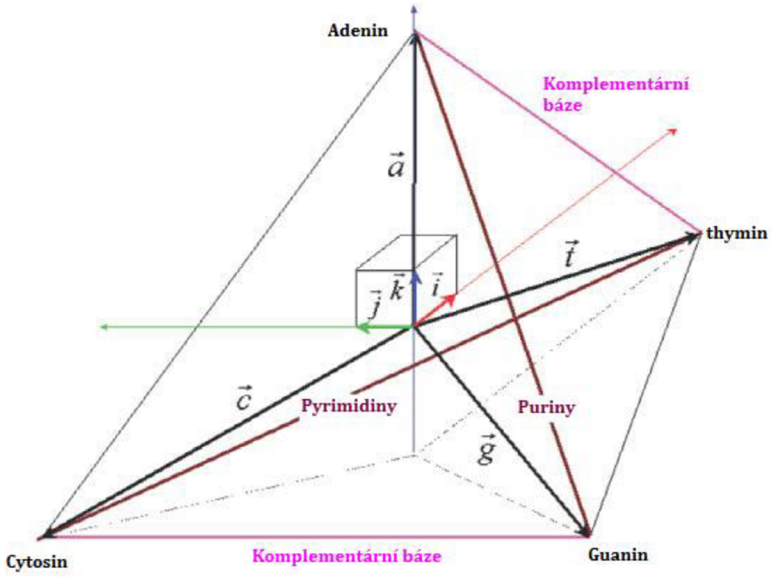
Numerické reprezentace (numerické mapování) slouží k předzpracování genomických dat, jako jsou DNA a RNA či data proteomická. Předzpracování je nutné pro následnou analýzu. Můžeme si představit, že genomická sekvence je jednorozměrný signál reprezentovaný symboly A, C, G, T nebo U. Práce se symboly je však obtížnější než s čísly. Některé numerické mapy však mohou způsobit ztrátu informačního obsahu, nebo naopak zdůrazní vlastnosti, které nejsou v symbolických sekvencích patrné. [7]

1.5.1 Převod nukleotidových sekvencí do genomického signálu

Zde je definována originální čtyřstěnná reprezentace (tetrahedron) genetického kódu (GC), která vystihuje jeho strukturu, degeneraci a vývoj trendů.

V DNA dvoušroubovici jsou naproti sobě v paralelních řetězcích vždy pyrimidové báze proti purinovým a to tak že páry tvoří A-T a C-G. Každá báze definuje svůj směr v prostoru pomocí čtyř odpovídajících základních vektorů, které jsou symetricky umístěny vzhledem k sobě navzájem a vždy orientované směrem k jednomu z rohů čtyřstěnu. Dále jsou báze propojeny na základě vodíkových můstků, které je pojí a poté podle toho, zda obsahují amino-skupinu nebo keto-skupinu. Pomocí tohoto nukleotidového čtyřstěnu je možné odvodit numerické mapování bez ztráty informace o biochemických vlastnostech a dále ho rozšířit na aminokyseliny a proteiny.

Možnost snížit rozměr reprezentace promítnutím GC čtyřstěnu na přiměřeně orientované rovině je také analyzována a vede k některým rovnocenným komplexním reprezentacím GC. Na těchto základech je optimální symbolicko-digitální mapování lineárních řetězců nukleových kyselin do reálných nebo komplexních genomických signálů odvozených od nukleotidů, kodonů a aminokyselin. Převedení sekvencí nukleotidů a polypeptidů do digitálních genomických signálů nabízí možnost využít širokou škálu signálu pro jejich zpracování a analýzu. Je prokázáno, že některé základní rysy nukleotidových sekvencí mohou být lépe získány touto reprezentací. [7] [14]

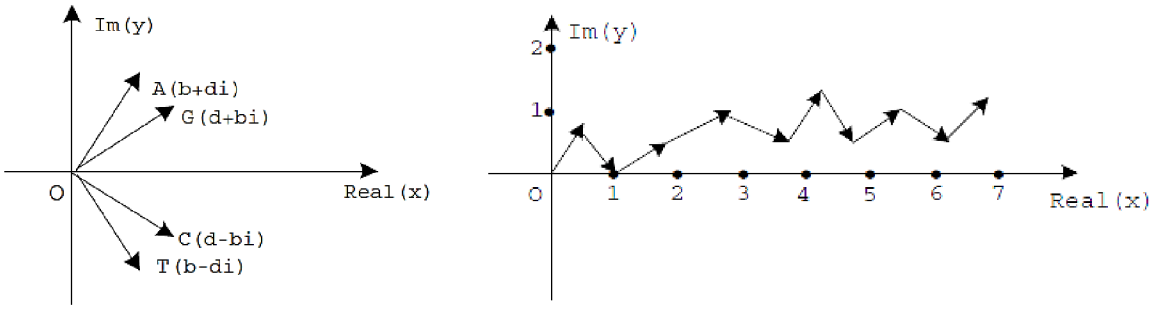


Obrázek 9: Struktura tetrahedronu

1.5.2 2D grafické znázornění genomických sekvencí

Grafické znázornění DNA poskytuje jednoduchý způsob prohlížení, třídění a porovnávání různých genových struktur. Grafická reprezentace spočívá v tom, že bázím adeninu (A), guaninu (G), tyminu (T) a cytozinu (C) se přiřadí jeden ze čtyř směrů (x, -x, y, -y). Tyto směrové šipky při grafické reprezentaci za sebe řadíme podle pořadí bází a do takové orientace, jaká vyplývá z grafu pro rozložení purinových a pyramidových bází.

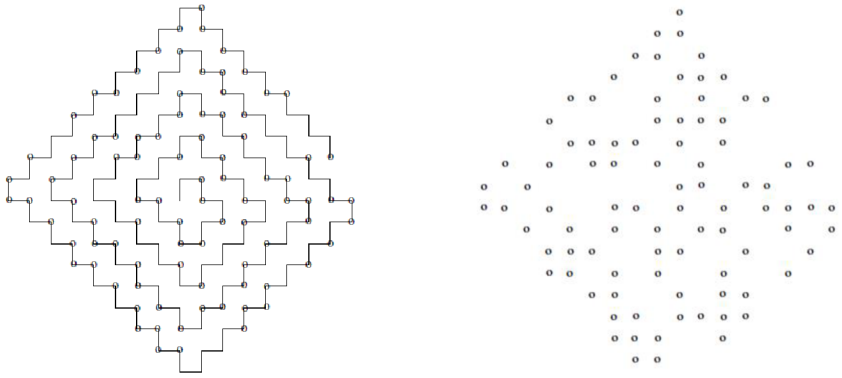
První šipka začíná v počátku souřadného sytému a následující šipka začíná v bodě, kde předchozí šipka končí. Reprezentace DNA je doprovázena nějakou ztrátou informace spojené s křížením a překrýváním výsledné křivky. V důsledku toho z těchto reprezentací není možné rekonstruovat základní DNA sekvence. Je možné 2D grafické znázornění proteinových sekvencí na základě nukleotidových tripletů kodonů v polovině komplexní roviny, která nemá žádnou pravidelnou trasu nebo rozklad, takže shoda mezi proteinovými sekvencemi a proteinovými grafy je jedna ku jedné. [15]



Obrázek 10: Graf rozložení DNA a směrová křivka sekvence ATGGCATGCA

1.5.3 Grafické znázornění DNA jako 2D mapy

Popisuje úpravu kompaktní reprezentace DNA sekvencí, které transformuje do 2D diagramu, ve kterém mají body celočíselné souřadnice. V důsledku doprovodné numerické charakterizace DNA je diagram poměrně jednoduchý a přímočarý. To je důležitá výhoda, zvláště u sekvencí DNA, které mají tisíce nukleových bází. Přístup začíná s kompaktní reprezentací DNA založené na klikaté spirálovité šabloně použité pro umístění bodů spojených s binárními kódy nukleových kyselin a následné potlačení podkladové klikaté křivky. Pouze pomocí vzdáleností mezi body mající stejnou souřadnici x nebo stejnou souřadnici y lze postavit mapu profilu pomocí celočíselné aritmetiky. [16]



Obrázek 11: Grafická 2D mapa před a po odstranění podkladové křivky

1.5.4 Charakterizace primární sekvence DNA dle zhuštěné matice

Zkrácená charakterizace primárních sekvencí DNA se děje na základě matice se čtyřmi řádky a sloupci, které jsou spojeny se čtyřmi nukleovými bázemi A, G, C a T. Zkrácené matice pro primární sekvenci DNA můžou sloužit jako zdroj invariantních sekvencí, které by umožnily kvantitativní porovnávání DNA z různých zdrojů. Možná strategie porovnávání DNA sekvencí, které mají obvykle různé délkya tudíž by se špatně porovnávali, je zastoupení této sekvence vhodně konstruovanými maticemi. Místo srovnávání sekvencí se srovnávají matice. Pokud jsou matice numerické, můžeme založit srovnání mezi nimi na základě porovnávání s maticemi invariantními. Je snaha matice konstruovat co nejmenší, protože tehdy nevykazují takové výpočetní a koncepční problémy. [17]

	A	G	C	T
A	AA	AG	AC	AT
G	GA	GG	GC	GT
C	CA	CG	CC	CT
T	TA	TG	TC	TT

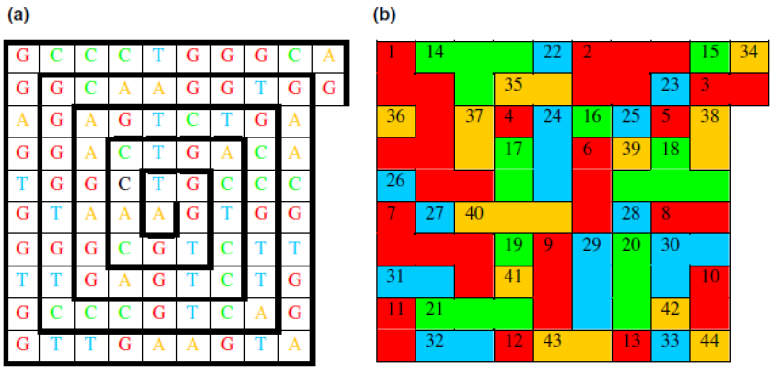
Obrázek 12: Matice bází

1.5.5 Reprezentaci DNA maticí vzdáleností

Přístup je založen na výstavbě čtyř samostatných submatic pro čtyři nukleotidy. První řádek každé z nich zachycuje vzdálenost mezi vybranými nukleotidy a zbývajícími nukleotidy stejného typu. Základní myšlenkou těchto grafických reprezentací DNA je usnadnění vizuální kontroly dlouhých sekvencí DNA, které by snad mohly pomoci kvalitativně rozpoznat podobnosti a odlišnosti mezi různými DNA sekvencemi nebo jejich částmi. Pozdější grafické znázornění DNA umožňuje výstavbu charakterizace DNA, které tak nabízejí kvantitativní přístup k měření stupně podobnosti a odlišnosti mezi podobnými sekvencemi DNA numericky. Základní myšlenkou je uspořádat DNA sekvenci tvořenou bázemi A, C, G a T do sady čísel. Tímto je možné vyjádřit číselný stupeň podobnosti mezi dvěma sekvencemi nebo jejich částmi. Matice nastíní délky úseček, které potom tvoří řádek. [18]

1.5.6 Čtyřbarevné mapy reprezentující DNA nebo RNA sekvence

Předkládáme novou 2D grafickou reprezentaci DNA nebo RNA sekvence založenou na čtyřbarevné mapě. Toto zastoupení je poměrně kompaktní a umožňuje tak člověku nejen provést vizuální kontrolu podobnosti nebo odlišnosti mezi sekvencemi DNA, ale také vede k jejich numerické charakterizaci. Prvním krokem v konstrukci čtyřbarevné mapy představující kódování sekvence například exonu lidského beta-globinu je reprezentovat kódující sekvence jako spirály čtvercových buněk. Spirála začíná středovou buňkou obsahujícího nukleotid A, a končí poslední buňkou na periferii, která obsahujíce nukleotid G. V důsledku toho je každý segment čtvercové sítě označený jedním ze čtyř písmen A, C, G a T. Pokud vymažeme sdílené strany mezi sousedními buňkami, které mají stejný název (barvu), pak dostane čtyřbarevnou mapu představující kódování sekvence exonu beta-globinu lidského genu. [19]



Obrázek 14: a) spirála buněk s nukleotidy, b) výsledná čtyřbarevná mapa

1.5.7 Charakterizace primární sekvence DNA pomocí trojice bází nukleových kyselin

Výstavba souboru menších matic o rozměru 4x4 reprezentující DNA primární sekvence, jsou založeny na výčtu všech 64 tripletů aminokyselinových bází. Vedoucí hodnota z konstruovaných matic byla vybrána jako neměnná pro výstavbu vektoru charakterizující DNA. Další varianty se považují za odvozené matice DNA a zahrnují 64-dílný vektor, jehož složky se skládají z pořadí tripletů XYZ, kde x, y, z se rovná A, G, C, T. Výstavba tabulky odlišností či podobností je založená na různých invariantech pro soubor sekvencí DNA. Základní úkol je srovnání DNA sekvence se sekvencí invariantní přítomnou v tabulce. Invariantní sekvence je zde považována za číslo nezávislé na označení písmeny A, G, C, A představující báze. Existují dva typy matic, v první matici individuální záznam odpovídá individuální páru bází a druhá, kde záznamy shrnují informace o různých XY párů bází. Matice, které shrnují informace párů XY byly označeny jako kondenzované matice. Mají velikost 4x4, protože je možných 16 párů, které čtyři písmena mohou generovat: AA, AC, AG, AT, CA, CC, atd. To vede k 10 různým párům bází (AA, AC, AG, AT, CC, CG, CT, GG, GT, a TT), které jsou považovány za výstavbu invariantů, které jsou nezávislé na uspořádání bází. [20]

1.5.8 Charakterizace primární sekvence DNA na základě průměrných vzdálenosti mezi bázemi

Zde je číselná charakterizace DNA primární sekvence založená na výpočtu průměrné vzdálenosti mezi páry bází nukleových kyselin. To vede k zastoupení DNA kondenzované symetrické matice o rozměru 4x4, jejíž prvky dávají průměrnou vzdálenost mezi dvojicemi bází X, Y v DNA (X, Y= A, C, G, T). Jako invariantní považujeme vedoucí hodnotu odvozené matice. Další strukturně příbuzné invarianty byly získány vytvořením další matice "vyššího řádu" odvozené z původní 4x4 matice, zvýšením svých prvků na vyšší hodnoty. Vhodně normalizované vedoucí hodnoty těchto matic nabízejí novou charakterizaci primární sekvence DNA. Důležitým úkolem při analýze dostupných údajů o DNA je odhadnout stupeň podobnosti mezi konečnými množinami řetězců nukleových bází. Standardní postupy uvažují rozdíly mezi řetězci v důsledku smazání nebo vložení, komprese nebo expanze a substituce prvků řetězce. [21]

Úsek DNA sekvence									
1	2	3	4	5	6	7	8	9	10
A	**T**	**G**	**G**	**T**	**G**	**C**	**A**	**C**	**C**
11	12	13	14	15	16	17	18	19	20
T	**G**	**A**	**C**	**T**	**C**	**C**	**T**	**G**	**A**

Tabulka 1: DNA sekvence

Upravené číslování DNA sekvence									
1	1	1	2	2	3	1	2	2	3
A	**T**	**G**	**G**	**T**	**G**	**C**	**A**	**C**	**C**
3	4	3	4	4	5	6	5	5	4
T	**G**	**A**	**C**	**T**	**C**	**C**	**T**	**G**	**A**

Tabulka 2: Upravené indexování DNA sekvence

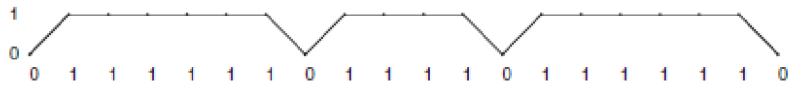
1.5.9 Charakteristická a podobnostní analýza DNA sekvencí znázorněna 2D grafickou reprezentací

Na základě klasifikace čtyř bází nukleových kyselin, se sníží DNA primární sekvence na tři (0, 1) sekvence a to non-A, non-G a non-C. Spojujeme každou sekvenci s maticí tím, že dáváme 2D grafické znázornění a získáme tak tří-složkový vektor s položkami majícími součet maximálních a minimálních vlastních čísel matic. Zavedený vektor se použije pro charakterizaci a porovnání kódující sekvence. Matematicky neměnný je součet maximální a minimální hodnoty takových matic, který se vypočítá pro popis DNA sekvence.

Čtyři báze DNA sekvence mají být rozděleny do tří skupin podle jejich chemické struktury, tj. non-A= G, C, T, non-G = A, C, T a non-C =A, G, T, dále se označí prvky non-A, non-G a non-C jako 1 a prvky A, G, a C jako 0. Podobně můžeme definovat "non-T", ale to není nutné, protože je to závislé na ostatních a vyplývá ze zápisu, že je přítomno když se všechny tři prvky rovnají 1. [22]

Tři (0,1) sekvence																			
Non-A	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1
Non-G	1	1	0	0	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0
Non-C	1	1	1	1	1	1	0	1	0	0	1	1	1	0	1	0	0	1	1

Tabulka 3: Tři sekvence utvořené podle nepřítomnosti daného nukleotidu



Obrázek 13: Grafická reprezentace tři složkového vektoru sekvencí

1.6 Denzita nukleotidů

Tato metoda se jeví jako nejoptimálnější ke zpracování mtDNA.

Tuto numerickou reprezentaci můžeme nazývat jako nukleotidovou denzitu nebo denzitní vektory. Denzita nukleotidů je jednoduchá a efektivní metoda numerické reprezentace symbolické sekvence DNA. Vyjadřuje průměrné zastoupení jednotlivých nukleotidů v definované části sekvence. Žádné dvě rozdílné sekvence nemají stejné nukleotidové denzity. Zpětná rekonstrukce symbolické sekvence z denzity nukleotidů není jednoznačná. Především na začátku a konci sekvence není možné přesně určit polohu některých nukleotidů, tudíž můžeme říci, že tento typ numerické reprezentace ztrácí informační obsah o přesné pozici některých nukleotidů. Při výpočtu denzity nukleotidů se nejprve ze symbolické sekvence vytvoří indikační vektory uA[n], uC[n], uG[n], a uT[n], které obsahují na pozici n hodnotu 1 pokud se daný nukleotid na pozici v sekvenci vyskytuje nebo hodnotu 0 v případě, že se nukleotid nevyskytuje. Následně jsou vypočítány jednotlivé denzitní vektory podle rovnice (1).

$$d_X[n] = \frac{\sum_{i=n-W/2}^{n+W/2} u_X[i]}{W},\ n = 1 \ldots N \qquad (1)$$

kde N je délka sekvence, W je velikost posuvného okna, X je typ nukleotidu. Velikost posuvného okna W musí být liché číslo, protože pozice n musí být centrálním prvkem v okně. Posuvné okno se pohybuje po celé délce indikačních vektorů a vrací průměr z hodnot v okně. Takto získáme set čtyř denzitních vektorů. Pro eliminaci vlivu začátku a konce sekvence se na začátky a konce indikačních vektorů přidává $W/2$ nul. Velikost posuvného okna je volitelná. Nejmenší možná smysluplná velikost je 5. Volba velikosti okna je závislá na délce sekvence a na požadované rozlišovací schopnosti výsledného signálu. [23]

Při využití nukleotidové denzity jako vstupního signálu do různých metod analýzy sekvencí bude potřeba velikost posuvného okna pro každou metodu optimalizovat. Grafická reprezentace nukleotidových denzit je možná dvojím způsobem. Nejjednodušší je samostatné vykreslení jednotlivých denzitních vektorů. Tento způsob vykreslení však neposkytuje dodatečnou vizuální informaci pro lidského operatéra. Druhý a vhodnější způsob vizualizace tkví v sumaci denzitních vektorů pro nukleotidy podobných biochemických vlastností a současném vykreslení komplementárních sumovaných vektorů. Tím se získají tři obrazy, které lze dobře vizuálně porovnávat pro různé sekvence. Sumace vektorů se řídí biochemickými vlastnosti nukleotidů dle dělení na purinové/pyrimidinové nukleotidy, nukleotidy tvořící silnou/slabou vazbu a nukleotidy obsahující amino/keto skupinu. [23]

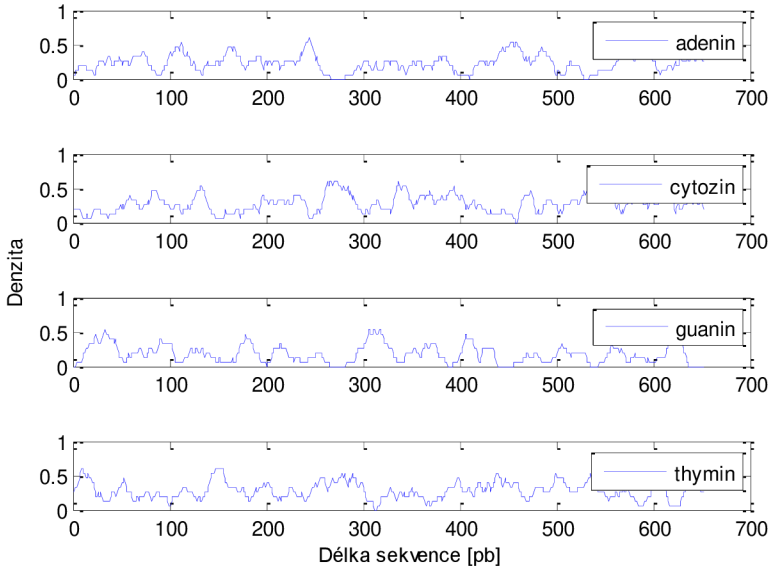
2 PRAKTICKÁ ČÁST

2.1 Programy

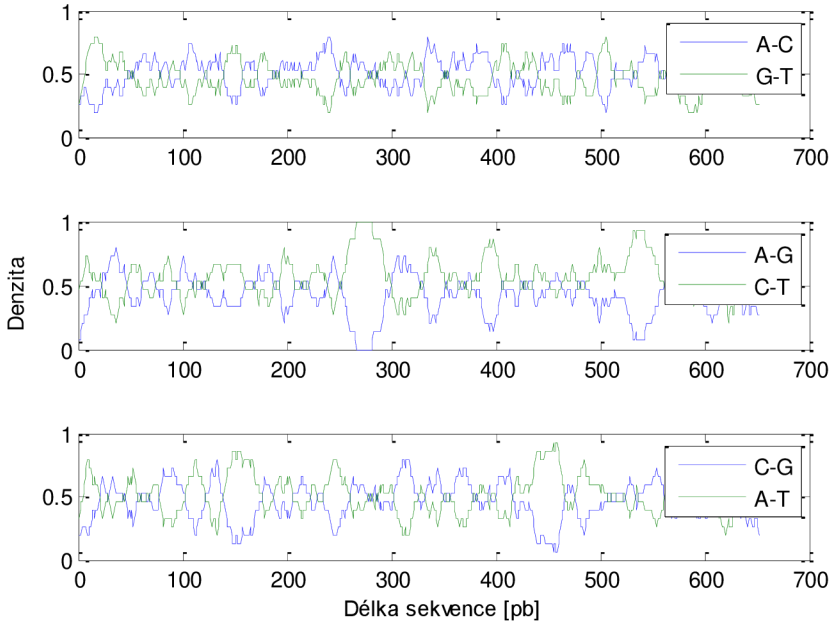
V následujících odstavcích detailně popisuji jednotlivé funkce a skripty vytvořené v programovém prostředí Matlab s Bioinformatics toolboxem, ve kterých realizuji výpočty denzitních vektorů jednotlivých sekvencí. Poté vytvářím referenční denzitní vektory, jež jsou následně přiřazovány k zástupcům ze souboru z databáze BOLD na základě minimální euklidovské vzdálenosti.

2.1.1 Funkce *Denzita*

Funkce s názvem ***Denzita*** slouží k výpočtu čtyř denzitních vektorů, kde vstupem je jedna sekvence DNA. Funkce načítá sekvenci ve formátu FASTA pomocí interní funkce ***fastaread***. Tento formát je definován tak, že začátek každé sekvenční položky začíná znakem '>', za nímž následuje jeden řádek informací o daném proteinu a od dalšího řádku následuje vlastní sekvence organismu. Po načtení sekvence následuje inicializace denzitních vektorů vytvořením vektoru nul o délce načtené sekvence. Tyto vektory se v cyklu přepisují na vektory s jedničkou v místě, kde se vyskytuje daný nukleotid. Takto nám vzniknou 4 indikační vektory uA[n], uC[n], uG[n] a uT[n]. Vektory se poté prodlouží na obou koncích o W/2 nul zaokrouhleno dolů, kde W je velikost výpočetního okna. Připraví se další inicializační vektory pro výpočet denzity, které mají stejnou délku jako samotná sekvence. Do nich se ukládají hodnoty výpočtu nukleotidové denzity pomocí průměrování hodnot indikačních vektorů v posuvném okně. Délka okna v této funkci je nastavena fixně na hodnotu 15. Následuje výpočet sum denzitních vektorů dle biochemických vlastností pomocí součtu vektorů a následné vykreslení výsledků. Pro testování programu jsem použila sekvenci rybky *Alepes djedaba* z databáze BOLD o délce 655 bp.



Obrázek 14: Denzitní vektory sekvence Alepes djedaba s délkou okna W=15



Obrázek 15: Sumy denzitních vektorů, Alepes djedaba, W=15

2.1.2 Funkce *Struktura*

Do funkce s názvem ***Struktura*** vstupuje celý soubor obsahující mnoho sekvencí ve fasta formátu. Výstup funkce je ve formátu struktury obsahující v jednom poli názvy organismů vstupujících sekvencí a v poli druhém samotné sekvence. Soubor je zde načítán pomocí funkce ***fastaread***, čímž získáme dvě cell buňky, z nichž jedna obsahuje sekvence a druhá obsahuje zmíněný popisný řádek, z něhož je potřeba získat samotný název organismu. Popisný řádek má následující formát:

>WLIND033-07|EF609500|WL-M33|Alepes djedaba|COI-5P.

Získání názvu je realizováno prohledáváním řetězce pomocí příkazu ***strfind***, kde nalezneme polohu znaků '|', kde za třetím nebo čtvrtým z nich se vyskytuje název. Ten získáme pomocí příkazu ***numel***, který zjistí počet lomítek, a poté složitějším indexováním cell buňky obsahující popisné řádky jednotlivých sekvencí. Jednotlivé názvy jsou postupně ukládány do dříve inicializované cell buňky.

Výsledná struktura poté vzniká za použití příkazu ***struct***, kde do prvního pole pojmenovaného název ukládám do sloupce názvy z cell buňky vystupující z cyklu. Do druhého pole s názvem sekvence je uložena transponovaná cell buňka obsahující sekvence. Podoba výsledné struktury je znázorněna na obrázku 16. Názvy organismů jsem nakonec příkazem ***xlswrite*** exportovala do *Microsoft Excel* s názvem *Taxonomy*.

Název:	Sekvence:
'Abudefduf vaigiensis'	CCCCTATCTAGT ...
'Abudefduf vaigiensis'	CCCCTATCTAGT ...
'Bodianus oxycephalus'	CCCCTATCTAGT ...
'Bodianus oxycephalus'	CCCCTATCTAGT ...
'Bodianus oxycephalus'	CCCCTATCTAGT ...
...	...

Obrázek 16: Schéma struktury obsahující názvy a sekvence

2.1.3 Skript pro vytváření referenčních denzitních vektorů

Tento skript s názvem ***ReferSek*** slouží k tvorbě referenčních sekvencí. Před jeho samotným spuštěním bylo předem potřeba udělat ručně několik operací. První z nich byla vybrat trojici zástupců pro každou z dvaceti referenčních sekvencí. Tato trojice byla vždy načtena z vytvořené Struktury do prostoru *Workspace*, poté byla vícenásobně zarovnána pomocí funkce ***multialign*** a zpětně zkopírována do nového *Microsoft Excel* z názvem *Zarovnane_sekvence.* Od následujícího kroku může být skript spouštěn pomocí ikony Run.

Pomocí příkazu ***xlsread*** se do *Workspace* načte dvacet zarovnaných trojic sekvencí z *Microsoft Excel*. Vektorem U jsou nadefinované pozice sekvencí v buňce označené txt vzniklé z příkazu ***xlsread***. V následujících několika cyklech dochází k výpočtu denzitních vektorů pro každou trojici sekvencí za pomocí volání vytvořené funkce ***Denzita***. Z takto vytvořených dvanácti denzitních vektorů jsou vyčleňovány denzitní vektory tak, že se utvoří čtyři matice, z nichž každá obsahuje tři denzitní vektory jedné báze. Obsažené vektory jsou zprůměrovány a jako vzniklé referenční denzitní vektory od jedné trojice sekvencí jsou vloženy opět do matice, kterou jsem postupně pro každou trojici exportovala do *Microsoft Excelu*. Do buňky s názvem *Tab_Ref_sek* jsou postupně ukládány referenční denzitní vektory do jednotlivých polí odpovídající jedné trojici sekvencí.

2.1.4 Funkce pro výpočet euklidovské vzdálenosti

Funkce s názvem ***Euklid_vzd.m*** je funkce sloužící k výpočtu euklidovské vzdálenosti. Vstupují do ní vždy dvě sekvence, které mohou mít rozdílnou délku a výstupem je hodnota minimální euklidovské vzdálenosti, která je vybrána z hodnot euklidovských vzdáleností odpovídající posunům kratší sekvence podél delší. Kratší sekvence se bude posouvat z leva doprava podél delší sekvence, celkově o tolik pozic o kolik se vzájemně liší, a v každém okamžiku posunu se bude počítat euklidovská vzdálenost. Nejprve zjistíme, která sekvence je delší a tudíž, která se jako kratší podle ní

bude posouvat a bude docházet k výpočtu vektoru euklidovských vzdáleností pomocí následujícího vzorce:

$$D_E = \sqrt{(A_r - A_i)^2 + (C_r - C_i)^2 + (G_r - G_i)^2 + (T_r - T_i)^2} \qquad (2)$$

kde D_E je bezrozměrná euklidovská vzdálenost, písmena *A,C,G,T* značí jednotlivé denzitní vektory a indexy *r,i* je označují jako referenční a indikační. Na závěr se vybere nejmenší hodnota z vektoru euklidovské vzdálenosti, které se normalizuje přepočtem této vzdálenosti připadající na jednu bázi, aby výsledky nebyli ovlivněné různými délkami sekvence.

2.1.5 Skript výpočtu euklidovské vzdálenosti a přiřazování zástupců

Skript s názvem ***Euklid.m*** slouží k samotnému výpočtu a porovnávání euklidovských vzdáleností mezi referenčním denzitními vektory a denzitními vektory testovaných zástupců. K testovaným zástupcům jsou přiřazování zástupci referenčních denzit, kde byla tato vzdálenost nejmenší, a tudíž k sobě mají z genetického hlediska nejblíže.

Skript začíná vytvořením struktury s názvem Tabulka, která obsahuji podstruktury, kde v prvním poli obsahuje názvy zástupců referenčních denzit, které byly exportovány z *Microsoft Excelu* s názvem *Taxonomy*. Druhé pole obsahuje referenční denzitní vektory, které byly načítány ze cell buňky s názvem *Tab_Ref_sek* vytvořené ve skriptu ***ReferSek*** a ukládány opět jako vnitřní struktura.

Následuje výpočet denzitních vektorů všech zástupců z dříve načtené Struktury obsahující jejich název a sekvenci pomocí vyvolání funkce ***Denzita.*** Do proměnné *Sek1* se načítá vždy čtveřice denzitních vektorů, odpovídající jedné testovací sekvenci a tyto denzitní vektory se porovnávají s dvaceti referenčními denzitními vektory, které se postupně ukládají do proměnné *Sek2*. Tyto proměnné poté vstupují do vytvořené funkce ***Euklid_vzd.m***, kde dochází k výpočtu dvaceti euklidovských vzdálenosti. Z tohoto vektoru se nalezne poloha s nejmenší hodnotou euklidovské vzdálenosti, která je poté použita jako index pro nalezení jména odpovídajícího referenčního zástupce ve struktuře *Tabulka* v poli s názvy. Tímto se do proměnné *Prirazeno_k* vygeneruje sloupec názvů referenčních sekvencí, které se poté kopírují *z Workspace* do *Microsoft Excelu* k předem nakopírovaným názvům testovaných zástupců. Takto lze porovnat, jací zástupci se k testovaným přiřadili.

2.1.6 Práce s programy

Při práci s programy se musí postupovat následovně. Za prvé se v *Command Window* vyvolá funkce ***Struktura.m*** a jako vstupní parametr se zadá název souboru obsahující sekvence ve formátu fasta. Ten musí být zadaný do závorky v apostrofech a s příponou .fas. Například: ***Struktura('Coral_fishes.fas').*** Tímto se vytvoří struktura obsahující názvy a sekvence testovaného souboru. Za druhé spustíme ikonou *Run* funkci ***ReferSek.m***, která vytvoří referenční denzity, musí mít k dispozici soubor *Microsoft Excel* s názvem *Zarovanane_sekvence*, který obsahuje vybrané sekvence, které jsou již zarovnány a ze kterých se denzity tvoří. V tomto programu tedy probíhá výpočet denzit pomocí volání funkce ***Denzita,*** kde je možno změnit délku okna W.

Poslední krok probíhá v programu ***Euklid.m***, který se též spustí ikonou *Run*, z něhož vystupuje proměnná *Prirazeno_k* obsahující v sloupci názvy přiřazených zástupců v počtu shodném s počtem vstupujících sekvencí ve struktuře. Pracuje s odvoláním se na funkci ***Euklid_vzd.m***. Je k němu opět potřeba mít k dispozici *Microsoft Excel* a názvem *Taxonomy*, který obsahuje dvacet názvů referenčních sekvencí. Tento skript pracuje s proměnnou *Strukura a Tab_Ref_sek* z předchozích souborů, proto je nepřípustné mezi jednotlivými programy mazat proměnné z prostoru *Workspace*. Pro porovnání výsledků je nutné, jak už bylo zmíněno, překopírování názvů do *Microsoft Excelu* a to názvů ze *Struktury* a paralelně k nim vložit názvy z proměnné *Prirazeno_k.*

2.2 Zhodnocení výsledků

Byla vyhodnocena úspěšnost identifikace sekvencí do referenčních druhů. V tabulce 4 je soupis referenčních druhů ze souboru *Coral fishes from South China Sea*, jejich taxonomické zatřídění a počet sekvencí jedinců daného druhu. Jedná se o sekvence z databáze BOLD, kde byl získán soubor o 341 sekvencích ve formátu fasta. Jedná se tedy o korálové ryby z Jihočínského moře. Průměrná délka těchto sekvencí se pohybuje okolo 680 pb (párů bazí). Celý soubor obsahuje celkem 51 živočišných druhů a jsou zde zastoupeny dvě třídy – *Actinopterygii* a *Elasmobranchii*.

Tyto sekvence byly zpracovány výše popsanými funkcemi, v případě referenčních sekvencí bylo dosaženo stoprocentní úspěšnosti. Ke všem zástupcům shodného druhu se přiřadily odpovídající referenční sekvence vytvořené pomocí sekvencí pouze od tří zástupců tohoto shodného druhu. Viz příloha *Microsoft Excel* s názvem *Taxonomy-přiřazeno, List 1.*

Tabulka 4: Popis vybraných sekvencí a úspěšnost zařazení mezi vlastní druhy

REFERENČNÍ DRUH	TŘÍDA	ŘÁD	ČELEĎ	ROD	POČET SEKVENCÍ	SPRÁVNĚ PŘIŘAZENO	ÚSPEŠNOST
Bodianus oxycephalus	Actinopterygii	Perciformes	Labridae	Bodianus	7	7	100%
Branchiostegus argentatus	Actinopterygii	Perciformes	Malacanthidae	Branchiostegus	16	16	100%
Caesio caerulaurea	Actinopterygii	Ophidiiformes	Caesionidae	Caesio	8	8	100%
Calotomus carolinus	Actinopterygii	Perciformes	Scaridae	Calotomus	8	8	100%
Carangoides chrysophrys	Actinopterygii	Perciformes	Carangidae	Carangoides	6	6	100%
Cephalopholis boenak	Actinopterygii	Perciformes	Serranidae	Cephalopholis	7	7	100%
Chaetodon auriga	Actinopterygii	Perciformes	Chaetodontidae	Chaetodon	3	3	100%
Chaetodon auripes	Actinopterygii	Perciformes	Chaetodontidae	Chaetodon	4	4	100%
Chaetodon wiebeli	Actinopterygii	Perciformes	Chaetodontidae	Chaetodon	11	11	100%
Choerodon azurio	Actinopterygii	Perciformes	Labridae	Choerodon	7	7	100%
Coryphaena hippurus	Actinopterygii	Perciformes	Coryphaenidae	Coryphaena	3	3	100%
Dactyloptena orientalis	Actinopterygii	Scorpaeniformes	Dactylopteridae	Dactyloptena	8	8	100%
Dendrochirus zebra	Actinopterygii	Scorpaeniformes	Scorpaenidae	Dendrochirus	8	8	100%
Epinephelus amblycephalus	Actinopterygii	Perciformes	Serranidae	Epinephelus	29	29	100%
Epinephelus areolatus	Actinopterygii	Perciformes	Serranidae	Epinephelus	9	9	100%
Epinephelus bleekeri	Actinopterygii	Perciformes	Serranidae	Epinephelus	4	4	100%
Epinephelus spilotoceps	Actinopterygii	Perciformes	Serranidae	Epinephelus	7	7	100%
Hemigymnus melapterus	Actinopterygii	Perciformes	Labridae	Hemigymnus	11	11	100%
Lethrinus haematopterus	Actinopterygii	Perciformes	Lethrinidae	Lethrinus	8	8	100%
Lethrinus lentjan	Actinopterygii	Perciformes	Lethrinidae	Lethrinus	5	5	100%

Tabulka č. 5 ukazuje, jak úspěšné bylo zařazení sekvencí jiných než referenčních druhů k zástupcům stejné čeledi. První sloupec obsahuje názvy čeledí, které jsou zastoupeny druhy v celém souboru *Coral_fishes* a zároveň byly obsaženy u referenčních druhů. Druhý sloupec uvádí, kolik sekvencí v souboru taxonomicky odpovídá dané čeledi. Následující tři sloupce uvádějí počty sekvencí, které byly správně přiřazeny k odpovídající čeledi referenčního druhu. Jsou zde rozdílné výsledky pro různou délku oken W, která se nastavovala ve funkci ***Denzita.***

Nejlepších výsledků bylo dosaženo při zvolené délce okna W = 15 a to 80,56 % úspěšně přiřazených sekvencí ke správné čeledi. Lepších výsledků poté dosahovala analýza s délkou oknem W = 23 před analýzou s délkou okna W = 7.

Tabulka 5: Počty správně přiřazených zástupců do čeledí souboru Coral_fishes

TESTOVANÝ SOUBOR Coral_fishes		Délka okna W=15	Délka okna W=7	Délka okna W=23
ČELEDI obsažené v souboru	Počet sekvencí v souboru	Počet správně přiřazených	Počet správně přiřazených	Počet správně přiřazených
Labridae	2	1	1	1
Malacanthidae	0	0	0	0
Caesionidae	0	0	0	0
Scaridae	9	9	0	9
Carangidae	7	1	7	1
Serranidae	2	2	1	1
Chaetodontidae	0	0	0	0
Coryphaenidae	0	0	0	0
Dactylopteridae	0	0	0	0
Scorpaenidae	16	16	16	16
Celkem:	36	29	25	28
Úspěšnost zařazení (%):		80,56	69,44	77,78

Tabulka 6 analyzuje nový soubor sekvencí s databáze BOLD s názvem Indian_fishes. Tento soubor obsahuje celkem 251 sekvencí, přičemž v analýze je zahrnuto prvních 111. Průměrná délka sekvencí je 680 pb. Obsahuje 75 různých druhů, z nichž 186 zástupců spadá do třídy *Actinopterygii* a řádu *Perciformes*, ze kterého jsou převážně tvořeny referenční sekvence.

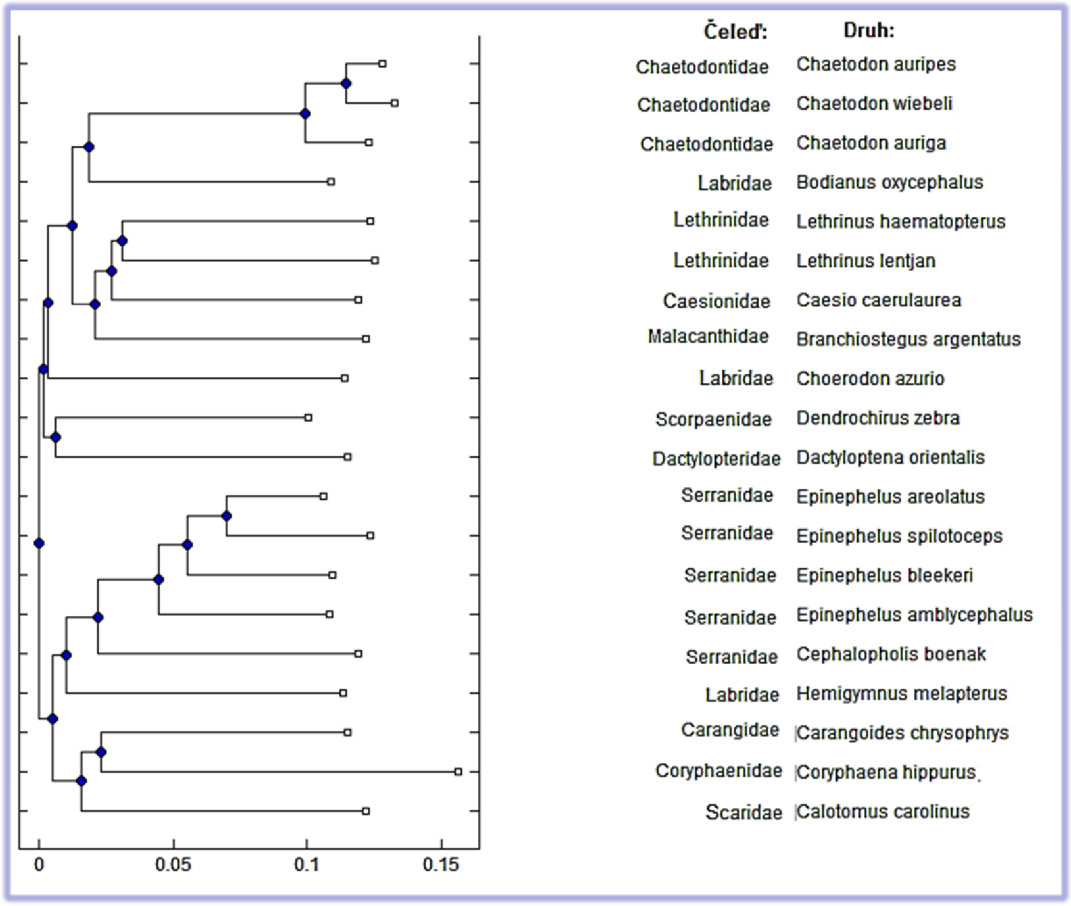
V prvním sloupci jsou opět uvedeny čeledě referenčních sekvencí, ve druhém je počet sekvencí v testovaném souboru, které jsou taxonomicky přiřazeny k dané čeledi, a třetí sloupec obsahuje, kolik z těchto sekvencí bylo správně přiřazeno, ke shodné čeledi z referenčních sekvencí. Zde se dle předchozích výsledků počítala denzita s délkou okna W = 15, kde byla úspěšnost nejvyšší. V tomto souboru se dosáhlo úspěšnosti téměř 95 % přiřazení testovaných sekvencí k referenčním sekvencím stejné čeledě.

Tabulka 6: Úspěšnost zařazení testované souboru Indain_fishes

TESTOVANÝ SOUBOR Indian_fishes	Délka okna W=15	
ČELEDI obsažené v souboru	Počet sekvencí v souboru	Počet správně přiřazených
Labridae	0	0
Malacanthidae	0	0
Caesionidae	0	0
Scaridae	0	0
Carangidae	18	18
Serranidae	19	19
Chaetodontidae	0	0
Coryphaenidae	0	0
Dactylopteridae	2	0
Scorpaenidae	0	0
Celkem:	39	37
Úspěšnost zařazení (%):	94,87	

Pro grafické vyjádření genetické vzdálenosti mezi sekvencemi slouží fylogenetický strom. Na obrázku 16 je fylogenetický strom vytvořený pro dvacet referenčních druhů, je konstruován metodou neighbor-joining, jehož princip spočívá v tom, že se postupně rozkládá hvězdicový strom tak, aby se v každém kroku maximálně snížila celková délka stromu. Distanční vzdálenost byla počítána pomocí metody Jukes-Cantor, která předpokládá, že substituční rychlosti jsou stejné pro všechny typy záměn. Je-li tedy celková rychlost substituce za jiný nukleotid u, pak rychlost změny za konkrétní jeden ze tří odlišných nukleotidů je u/3.[24], [25]

Můžeme si všimnout, že zástupci čeledě *Chaetodontidae* vycházejí z jediné větve a vzdálenost mezi nimi je malá, což vypovídá o genetické blízkosti. Naopak zástupci čeledě *Labridae* jsou umístěné v několika rozdílných větvích, což může znamenat velkou genetickou variabilitu u zástupců této čeledi.



Obrázek 17: Fylogenetický strom referenčních druhů

3 Závěr

Cílem této práce bylo přiblížení metod pro molekulární taxonomii, představení DNA barcodingu a programová realizace vhodné numerické metody pro zpracovávání genomických sekvencí s následnou klasifikací souboru sekvencí s databáze BOLD pomocí distančního porovnání s referenčními sekvencemi.

V práci se mi podařilo přiblížit biologické základy, především buněčný aparát s bližším popisem mitochondrií a mitochondriální DNA. Představila jsem DNA barcoding, jako jeden z nejaktuálnějších projektů zabývající se fylogenetickou a uvedla některé z numerických metod pro zpracování genomických dat.

Z uvedených metod jsem si, jako nejvíce vhodnou, vybrala metodu tvorby denzitních vektorů, která nám nejlépe znázorňuje rozložení nukleotidů v DNA sekvenci a je tedy vhodná pro porovnávání sekvencí mezi sebou. Tuto metodu jsem programově realizovala v prostředí Matlab. Následovala tvorba dvaceti referenčních sekvencí, kdy jsem si ze souboru vybrala vždy tři sekvence jednoho druhu, které jsem zarovnala na stejnou délku, vypočítala jejich denzitní vektory a udělala jejich průměr. Pro klasifikaci testovaných sekvencí jsem vytvořila programy pro načtení celého souboru sekvencí z databáze BOLD a pro výpočet jejich denzitních vektorů s následným distančním porovnáním pomocí euklidovské vzdálenosti mezi jednotlivými denzitními vektory. Na základě vypočítané euklidovské vzdálenosti se k testované sekvenci přiřadila ta referenční sekvence, mezi nimiž byla hodnota euklidovské vzdálenosti nejmenší. Což znamená, že testovaná sekvence se s touto referenční sekvencí nejvíce shoduje s porovnání s devatenácti ostatními a tudíž jsou tyto sekvence k sobě geneticky nejblíže.

Na závěr proběhla analýza celého souboru sekvencí z databáze BOLD s názvem *Coral fishes from South China Sea,* a to s jakou úspěšností se testované sekvence přiřadili ke shodnému referenčnímu druhu, kde se úspěšnost přiřazování pohybovala ve vysokých procentech. U referenčních sekvencí se dosáhlo úspěšnosti 100%, tzn. že k zástupcům stejného druhu se správně přiřadila referenční sekvence. U nereferenčních sekvencí ze souboru *Coral_fishes* bylo dosaženo úspěšnosti přiřazení testované sekvence ke správné čeledi 80,56% při délce okna W=15, 69,44% při délce okna W=7 a 77,78% při délce okna W=23. V testovaném souboru *Indian_fishes* bylo dosaženo úspěšnosti 94,89%, kdy se k testovaným sekvencí přiřadila správná referenční čeleď.

Literatura

[1]. **Rosypal, Stanislav.** *Nový přehled biologie.* Praha : Nakladatelství Scientia, 2003. ISBN 978-80-86960-23-4.

[2]. **Benešová, Marika.** *Odmaturuj z biologie.* Brno : Nakladatelství Didaktis, 2003. ISBN 80-86285-67-7.

[3]. **Hampl, Vladimír.** Molekulární taxonomie: Úvod, taxonomie, molekulární znaky, sekvenace DNA. *Evoluntionary protistology group.* [Online]. [cit. 2012-11-21]. Dostupné z: http://www.protistologie.cz/files/MolTax/Molekularni%20taxonomie1-text.pdf.

[4]. **Raclavský, Vladislav.** Metody molekulární genetiky. [Online]. [cit. 2012-11-23]. Dostupné z: *http://biologie.upol.cz/metody/*

[5]. **Taanman, J.W.** The mitochondrial genome: structure, transcription, translation and replication. *Biochimica et Biophysica Acta 1410.* 21. July 1998.

[6]. **Rosypal, Stanislav.** *Úvod do molekulární biologie.* Brno : Katedra generiky a molekulární biologie PřF MU, 1997.

[7]. *Bioinformatika – Návod do počítačových cvičení: Numerické reprezentace genomických dat.* **Maděránková, Denisa.** 2012.

[8]. **Alberts, Bruce.** *Základy buněčné biologie.*Ústí nad Labem : Espero Publishing s.r.o., 1998. ISBN-80-902906-2-0.

[9]. Barcode of life: What is DNA Barcoding. *Barcode of Life.* [Online] 2010-2012. Dostupné z: http://www.barcodeoflife.org/content/about/what-dna-barcoding.

[10]. Barcode of life: The barcodig landscape. *Barcode of life.* [Online] 2010-2012. Dostupné z: http://www.barcodeoflife.org/content/about/barcoding-landscape.

[11]. International barcode of life:What is DNA Barcoding? *International barcode of life.* [Online] Threestone Studios Inc., 2012. Dostupné z: http://ibol.org/about-us/what-is-dna-barcoding/.

[12]. Barcode of life: What is iBOL. *Barcode of life.* [Online] 2010-2012. Dostupné z: http://www.barcodeoflife.org/content/about/what-ibol.

[13]. Barcode of life: What is cBOL. *Barcode of life.* [Online] 2010-2012. Dostupné z: http://www.barcodeoflife.org/content/about/what-cbol.

[14]. **Cristea, P. D.** Conversion of nucleotides sequences into genomic signals. *Bio-Medical Engineering Center.* 11. March 2002.

[15]. **Fenglan Bai, Tianming Wang.** A 2-D graphical representation of protein sequences based. *Chemical Physics Letters.* 2005.

[16]. **Randic, Milan.** Graphical representations of DNA as 2-D map. *Chemical Physics Letters.* 2004.

[17]. —. On characterization of DNA primary sequences by a condensed. *Chemical Physics Letters.* 1999.

[18]. **Tomaž Pisanski, Jure Zupan, Milan Randic.** On representation of DNA by line distance matrix. *Journal of Mathematical Chemistry.* 2006.
[19]. **Nella Lers, Milan Randic,Dejan Plavsic, Subhash C. Basak.** Four-color map representation of DNA or RNA sequences. *Chemical Physics Letters.* 2005.
[20]. **Xiaofeng Guo, Milan Randic, Subhash C. Basak.** On the Characterization of DNA Primary Sequences by Triplet of Nucleic Acid Bases. *J. Chem. Inf. Comput. Sci. 2001.* 2001.
[21]. **Subhash C. Basak, Milan Randic.** Characterization of DNA Primary Sequences Based on the Average Distances. *J. Chem. Inf. Comput. Sci.* 2001.
[22]. **Jun Wang, Yi Zhang.** Characterization and similarity analysis of DNA sequences grounded. *Chemical Physics Letters.* 2006.
[23]. **Denisa Maderankova, Ivo Provaznik.** Motive Representation in Nucleotide Densities of Bird's Mitochondrial Gene COX1. *ISABEL '11 Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies.* 2011.
[24]. **Hampl, Vladimír.** Molekulární taxonomie: SNP, genetické distance. *Evoluntionary protistology group.* [Online]. [cit. 2013-01-21]. Dostupné z: http://www.protistologie.cz/files/MolTax/Molekularni%20taxonomie5-text.pdf
[25]. **Hampl, Vladimír.** Molekulární taxonomie: Proteinové distance, konstrukce fylogenetických stromů z matice distancí. *Evoluntionary protistology group.* [Online]. [cit. 2013-01-21] Dostupné z: http://www.protistologie.cz/files/MolTax/Molekularni%20taxonomie6-text.pdf