VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

KLASIFIKACE DNA SEKVENCE

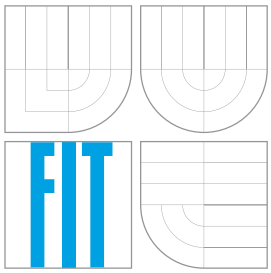
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

PETRA HECZKOVÁ

BRNO 2015

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

KLASIFIKACE DNA SEKVENCE

DNA SEQUENCE CLASSIFICATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

PETRA HECZKOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. TOMÁŠ MARTÍNEK, Ph.D.

BRNO 2015

Abstrakt

Práce se zabývá klasifikací DNA sekvencí. V první části jsou shrnuty informace o existujících metodách a jejich vlastnostech. V druhé části je popsána implementace a experimenty. Průměrná sensitivita metody byla 65% a průmerná specificita 92%.

Abstract

The work deals with DNA sequence classification. The first part summarizes information about existing methods a their characteristics. In the second part there are description of implementation and experiments. Average sensitivity of method was 65% and specificity 92%.

Klíčová slova

molekulární biologie, DNA, klasifikace, křížová validace, jazyk R, taxonomicky závislé metody

Keywords

molecular biology, DNA, classification, cross validation, R language, taxonomy-dependent methods

Citace

Petra Heczková: Klasifikace DNA sekvence, bakalářská práce, Brno, FIT VUT v Brně, 2015

Klasifikace DNA sekvence

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením pana Ing. Tomáše Martínka, Ph.D.

.....
Petra Heczková
17. května 2015

Poděkování

Na tomoto místě bych ráda poděkovala svému vedoucímu Ing. Tomáši Martínkovi, Ph.D. za cenné připomínky a odborné rady. Dále děkuji za přístup k výpočetním a úložným zařízením poskytovaným v rámci programu „Projekty velké infrastruktury pro výzkum, vývoj a inovace," (LM2010005), které vlastní Národní gridová infrastruktura MetaCentrum.

Obsah

1 Úvod **2**

2 Úvod do molekulární biologie **3**

3 Metody pro klasifikaci DNA **7**
3.1 Taxonomicky závislé metody 7
3.2 Taxonomicky nezávislé metody 10
3.3 Validace metod 11
3.4 Hodnocení metod 12

4 Návrh řešení **14**
4.1 Detailní specifikace a formulace cíle 14
4.2 Algoritmus 14
4.3 Trénink 14
4.4 Klasifikace 15

5 Implementace **18**
5.1 Nástroje použité k realizaci 18
5.2 Formát vstupních dat 18
5.3 Popis funkcí 19

6 Výsledky měření **22**
6.1 Zjištěné hodnoty parametru 22
6.2 Experiment 1 23
6.3 Experiment 2 23

7 Závěr **26**

A Obsah CD **29**

B Genom **30**

Kapitola 1

Úvod

V současné době existuje mnoho metod, které se zabývají klasifikací DNA sekvencí. Klasifikace je proces zařazení neznámé DNA sekvence do vhodné taxonomické kategorie. Metody se rozdělují do dvou skupin, taxonomicky závislé a taxonomicky nezávislé. Většina existujících metod je náročná na čas a výpočetní zdroje.

Klasifikace DNA sekvencí je důležitá při identifikaci a třídění organismů, určování evolučního vývoje druhu a tvorbě fylogenetických stromů. To přináší nové informace o genomu a vede k lepšímu pochopení této problematiky.

Cílem této bakalářské práce je návrh, implementace a testování nové metody na reálných datech. Tato metoda klasifikuje sekvence do nadříší. Rozlišuje sekvence eukaryotické, prokaryotické a archea. Metoda připraví sadu známých sekvencí, kterou rozdělí do shluků. Porovnává hledanou sekvenci se všemi centry shluků a k nejpodobnějšímu ji přiřadí. Podle poměru nadříší zastoupených v tomto shluku hledanou sekvenci klasifikuje.

V kapitole 2 se nachází stručný úvod do molekulární biologie. Tato kapitola obsahuje základní popis struktury a vlastností DNA. Popisuje rozdíly mezi jednotlivými nadřísemi, do kterých metody klasifikují. Kapitola 3 charakterizuje rozdělení metod klasifikace DNA sekvencí a podrobněji popisuje jednotlivé metody, kritéria jejich hodnocení a způsoby, jak ověřovat jejich přesnost. Kapitola 4 obsahuje cíle, návrh a detailní popis algoritmu nové metody. Kapitola 5 se zabývá implementací a konkrétními nástroji použitými k realizaci metody a kapitola 6 shrnuje výsledky experimentů, které byly provedeny.

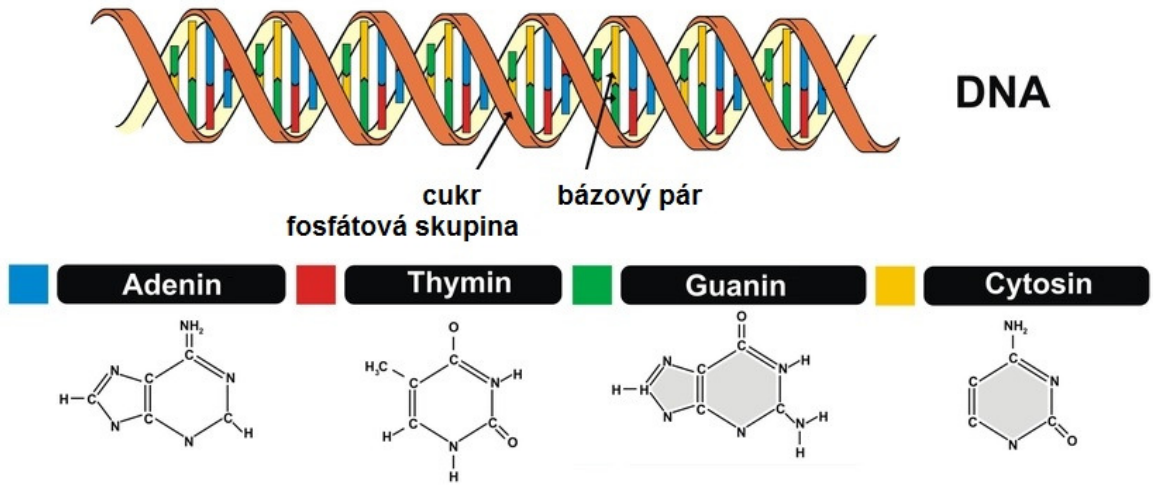
Kapitola 2

Úvod do molekulární biologie

Struktura DNA

DNA (Deoxyribonukleová kyselina [16]) je molekula obsahující informace, které organismy potřebují k vývoji, životu a rozmnožování. Tyto informace jsou uloženy v každé buňce a jsou předávány z generace na generaci.

DNA se skládá z cukru zvaného deoxyribóza, fosfátové skupiny a nukleové báze. V DNA jsou čtyři typy nukleové báze: adenin (A), thymin (T), guanin (G) a cytosin (C). Tyto báze v různém pořadí tvoří geny, které mají specifickou funkci při tvorbě bílkovin. Chemická struktura nukleových bází je vidět na obrázku 2.1. Např. celý lidský genom obsahuje okolo tří miliard bází a dvacet tisíc genů. Nukleotidové báze tvoří páry, A se páruje s T a G se páruje s C, jedná se o tzv. komplementaritu bází. Poměr CG páru je důležitý ukazatel v některých klasifikátorech, např. obsah CG páru u savců je 39–46%, u baterií se pohybuje mezi 25% až 75%.

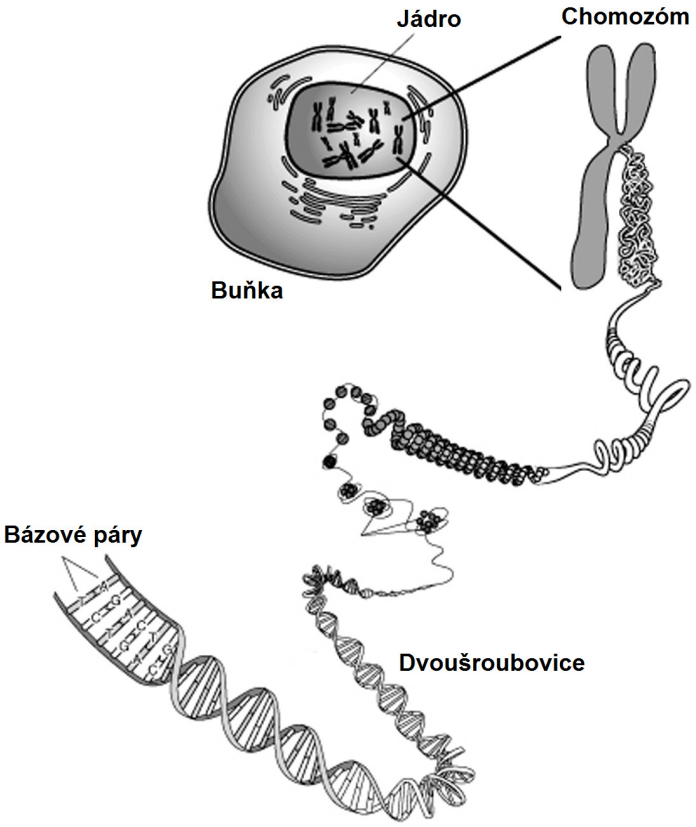


Obrázek 2.1: Chemická struktura nukleových bází a jejich párování v DNA. Modře označený adenin se páruje s červeně označeným thyminen a zelený guanin se žlutým cytosinem. [2]

DNA se uchovává ve formě pravotočivé dvoušroubovice. Ta je tvořena dvěma vlákny

[2] Převzato z http://www.livescience.com/37247-dna.html.

DNA, které se obtáčí kolem společné osy. DNA se dále organizuje do mikroskopicky pozorovatelných útvarů známých jako chromozomy. Ty se u eukaryotických buněk nacházejí v buněčném jádře (to je možné vidět na obrázku 2.2). U buněk prokaryotických a archea se DNA nachází volně v cytoplasmě.



Obrázek 2.2: Vyšší úrovně struktury u eukaryotické buňky: Jednotlivé nukleové báze se párují a tvoří dvoušroubovici. Ta se dále vine až do útvarů, které se označují jako chromozomy. [4]

Archea

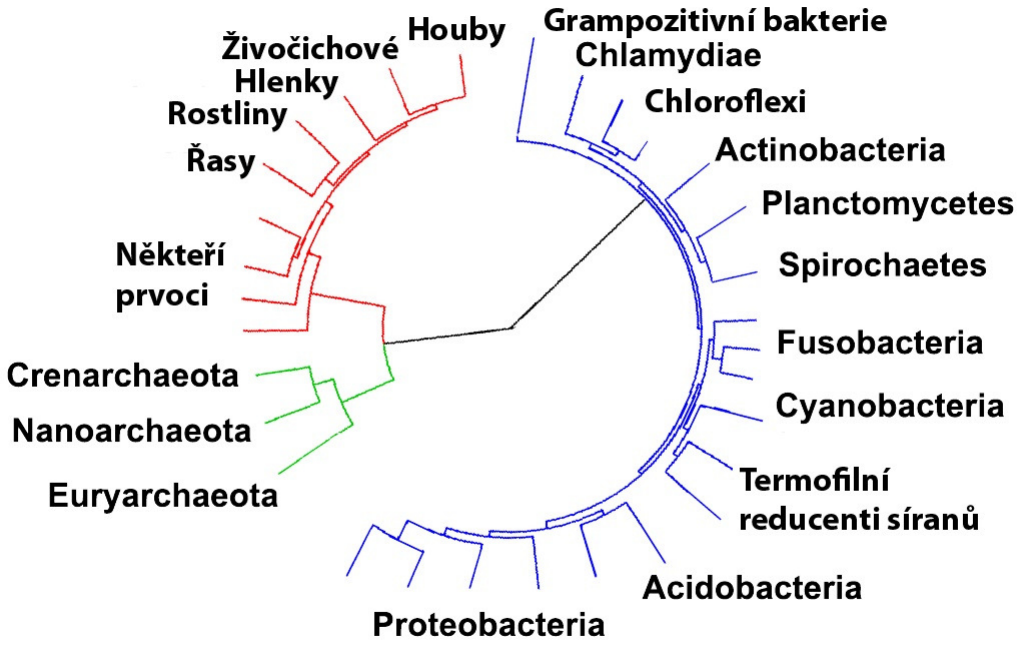
Archea [3] je skupina jednobuněčných mikroorganismů, která se na Zemi objevila před asi třemi a půl miliardami lety. Dále se dělí do čtyř kmenů. Tyto mikroorganismy často žijí v extrémních podmínkach (např. termální prameny a slaná jezera). Některé vlastnosti mají podobné s bakteriemi (např. rozmnožování, stavba buněk) a dohromady tvoří skupinu zvanou prokaryota. Na obrázku 2.3 je fylogenetický strom, který znázorňuje vztahy mezi hlavními taxonomickými doménami.

Prokaryota

Prokaryota [2] (prvojaderní) jsou malé většinou jednobuněčné organismy s poměrně jednoduchou strukturou. Jejich buňky postrádají jádro a další organely. DNA se ve formě jedi-

[4] Převzato z http://creationwiki.org/Chromosome.

ného chromozomu nachází v centrální části buňky a není obklopeno žádnou membránou, tzv. nukleoid. Mohou se zde vyskytovat i plasmidy, což jsou krátké molekuly DNA rozdílné od chromozomální DNA. Plasmidy poskytují organismu genetickou výhodu ve specifickém prostředí (např. obsahují gen pro odolnost vůči antibiotiku). Plasmidy se nacházejí pouze u prokaryotických buněk. Na obrázku 2.4 je vidět struktura prokaryotické buňky.

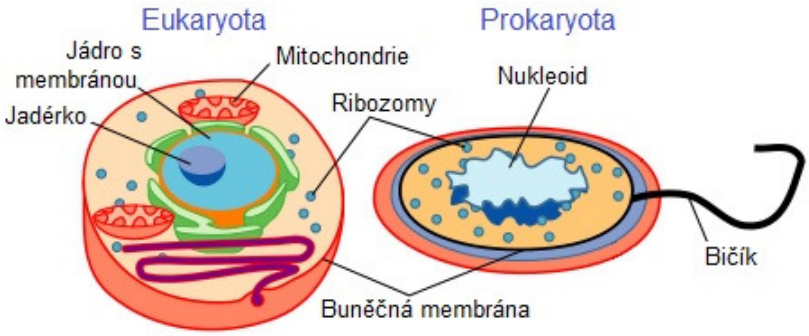


Obrázek 2.3: Fylogenetický strom: Archea jsou označena zeleně, bakterie modře a eukaryota červeně. [6]

Eukaryota

Eukaryota [1] (jaderní) jsou jednobuněčné a mnohobuněčné organismy, které vznikly před jednou a půl miliardami let. Patří zde například živočichové, rostliny, houby nebo prvoci. Jejich buňky obsahují jádro a jiné organely obalené membránou (např. mitochondrie, které pro organismus produkují energii). Genetický materiál se nachází v jádře a je rozdělen do jednotlivých chromozomů. Struktura buňky je na obrázku 2.4.

[6] Převzato z http://cs.wikipedia.org/wiki/Archea.
[8] Převzato z http://cs.wikipedia.org/wiki/Archea.

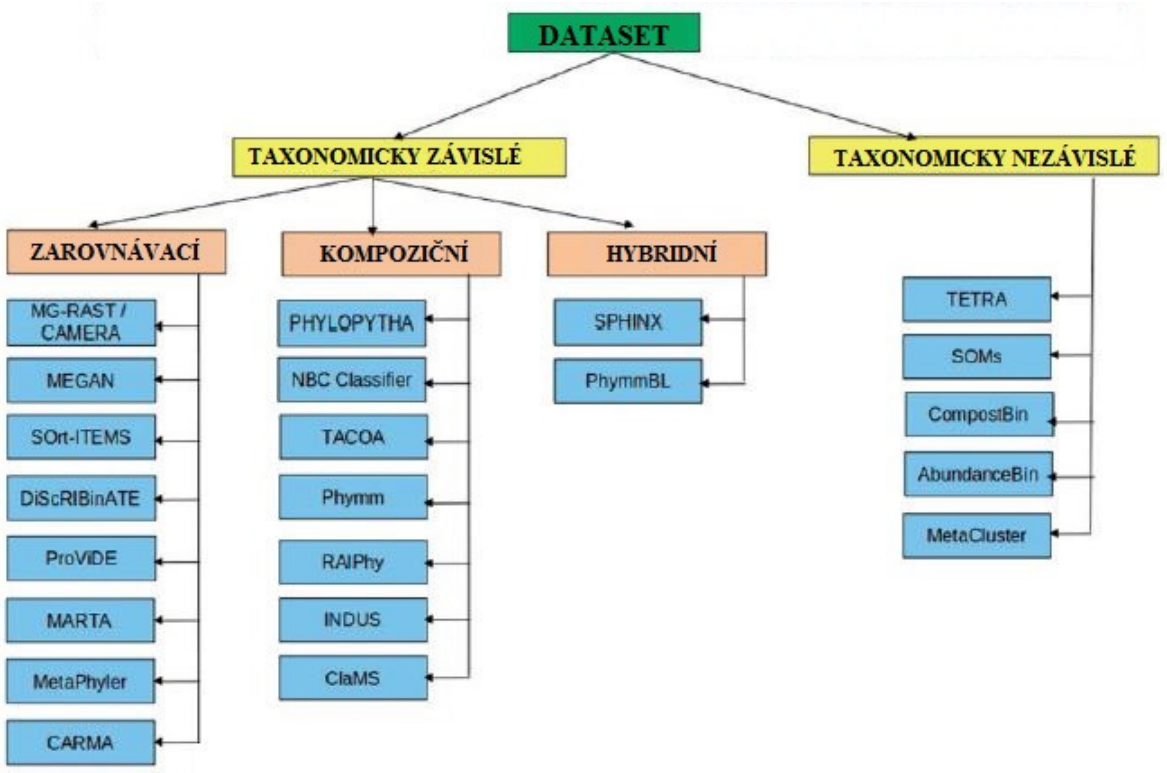


Obrázek 2.4: Porovnání eukaryotické a prokaryotické buňky: V eukarotické jsou organely odděleny od okolí membránou a jádro obsahuje chromozomy. V buňce prokaryotické jsou chromozomy v centrální části, tzv. nukleoid. [8]

Kapitola 3

Metody pro klasifikaci DNA

V oblasti klasifikace DNA sekvencí existuje několik různých přístupů. Tyto přístupy jsou rozděleny do dvou hlavních skupin - taxonomicky závislé a taxonomicky nezávislé. V této kapitole je popsáno rozdělení metod klasifikace DNA a některé konkrétní metody, které jsou vidět na obrázku č. 3.1.



Obrázek 3.1: Metody klasifikace DNA sekvencí a jejich rozdělení. Převzato z [9].

3.1 Taxonomicky závislé metody

Taxonomicky závislé metody porovnávají hledanou sekvenci s již známými sekvencemi. Musí tedy obsahovat referenční knihovnu nebo referenční sadu dat, ve které jsou potřebné informace o těchto známých sekvencích, zejména pak informaci k jaké taxonomické úrovni

sekvence patří. Každá jednotlivá metoda má vlastní způsob, jak ohodnotit podobnost hledané a referenční sekvence. Hledaná sekvence je pak obvykle přiřazena do skupiny s nejlepší shodou.

Důležitou vlastností klasifikátorů je úroveň zařazování. Klasifikátor specifický v kategorii (z ang. rank-specific) je takový klasifikátor, který provádí taxonomické úkoly pouze na určité taxonomické kategorii, např. říše nebo kmenu. V případě, že hledaná sekvence je nezařaditelná v určené kategorii je obvykle ponechána jako neznámá. Opakem je flexibilní v kategorii (z ang. rank-flexible), který klasifikuje na všech úrovních a zároveň usiluje o co nejpřesnější klasifikaci, tzn. pokud není možné určit rod, určuje se čeleď atd.

Podle způsobu porovnání se dále dělí na metody založené na zarovnávání (z ang. alignment-based), kompoziční (z ang. composition-based) a hybridní, které kombinují vlastnosti obojího.

Metody založené na zarovnávání

Metody založené na zarovnávání porovnávají jednotlivé bázové páry hledané a referenční sekvence. Příkladem nástroje, který najde oblasti podobnosti mezi sekvencemi je BLAST [13] (z ang. The Basic Local Alignment Search Tool). Tento program porovnává nukleotidové nebo proteinové sekvence DNA se sekvencemi v databázi a vypočítá statistickou významnost shody. BLAST je často používán jinými metodami, např. metodou PhymmBL.

Kompoziční metody

Kompoziční metody převádí sekvence na příznakové vektory, které obsahují charakteristické vzory složení (např. frekvenci určitých oligonukleotidů). Tyto metody srovnávají vektor hledané sekvence s referenční sadou vektorů. Tato referenční sada se často vytváří jednorázově v předzpracování. Kompoziční metody jsou obvykle rychlejší než metody založené na zarovnávání.

Příkladem kompozičních metod jsou Phylopythia, TACOA, ClaMS a Phymm.

Hybridní metody

Hybridní metody kombinují přístupy kompozičních metod a metod založených na zarovnávání. Například u hybridní metody PhymmBL se metoda Phymm (viz 3.1) provádí pararelně s vyhledáváním BLAST a jejich výsledky se váhově skládají.

Eu-Detect: Algoritmus pro detekci eukaryotních sekvencí

Metoda Eu-Detect [12] je kompoziční metoda, která sekvence nezarovnává (z ang. alignment-free), jejíž učelem je detekce eukaryotních sekvencí. Používá referenční vzory, což jsou frekvence všech možných tetranukleotidů (k-merů o délce 4) v každém fragmentu, které byly vypočítány a uloženy ve formě 256 rozměrného vektoru. Metoda nejprve zjišťuje nejbližší shluky pro danou sekvenci, následně kontroluje poměr prokaryotických do eukaryotických sekvencí v těchto uskupeních.

Jako jednorázové předzpracování je provedeno shlukování metodou k-means. Počet center shluků k, která slouží k zahájení procesu shlukování se spočítá podle vzorce $k \approx \sqrt{\frac{n}{2}}$,kde n je počet objektů, používá se pro stanovení počátečního počtu shluků.

Jsou nalezeny shluky s nejmenší vzdáleností a je spočítán celkový počet sekvencí přiřazených do těchto shluků. Tato hodnota se porovnává s hodnotou souhrného počtu sekvencí.

Souhrnný počet sekvencí (z ang. cumulative sequence count) představuje optimální počet nejbližších sekvencí, pro které se provádí klasifikace. Tato hodnota je zjištěna experimentálně.

Překročí-li procento eukaryotických sekvencí v nejbližší sadě předem stanovený práh pokrytí (z ang. coverage threshold), vstup bude klasifikován jako pravděpodobně eukaryotická sekvence. Nepřekročí-li bude klasifikován jako prokaryotická sekvence, nikdy se nemůže stát, že by sekvence zůstala neklasifikována.

TACOA

TACOA [7] (Taxonomická klasifikace environmentálních genomových fragmentů z ang. Taxonomic Composition Analysis) je algoritmus, který spojuje metodu k-nejbližších sousedů (z ang. k-nearest neighbor) a učení založené na jádře (z ang. kernel-based learning).

Každý fragment DNA sekvence je reprezentován jako vektor GFV (z ang. Genomic Feature Vector). Ten obsahuje poměr mezi hodnotou O a E pro všechny tetra-oligonukleotidy. Hodnota E určuje kolikrát se měl daný tetra-nukleoid v sekvenci nacházet (počítáno pomocí GC obsahu pro daný druh), hodnota O označuje kolikrát ho sekvence skutečně obsahovala. Poté je GFV vstupu porovnáván se sadou referenčních GFV. Shoda je hodnocena funkcí jádra.

ClaMS

Hlavním cílem metody ClaMS [15] (A Classifier for Metagenomic Sequences) je začlenit do procesu shlukování a klasifikace DNA sekvencí informace o přítomnosti a množství sekvencí specifických pro jednotlivé druhy organismů. ClaMS charakterizuje sekvenci pomocí podpisového vektoru, který je odvozen od jeho složení. Podpisový vektor je odvozen pomocí de Bruijnova grafu. Tyto podpisové vektory jsou si velmi podobné až identické v rámci jednoho druhu, zatímco mezi jednotlivými druhy se liší. Tato vlastnost se projevuje i na vyšších taxonomických úrovních.

Pro každou sekvenci, která má být klasifikována je vypočítán podpisový vektor. Ten je srovnáván s podpisovými vektory všech tréninkových sestav, které reprezentují jednotlivé druhy. Poté je přiřazen k druhu s nejlepší shodou.

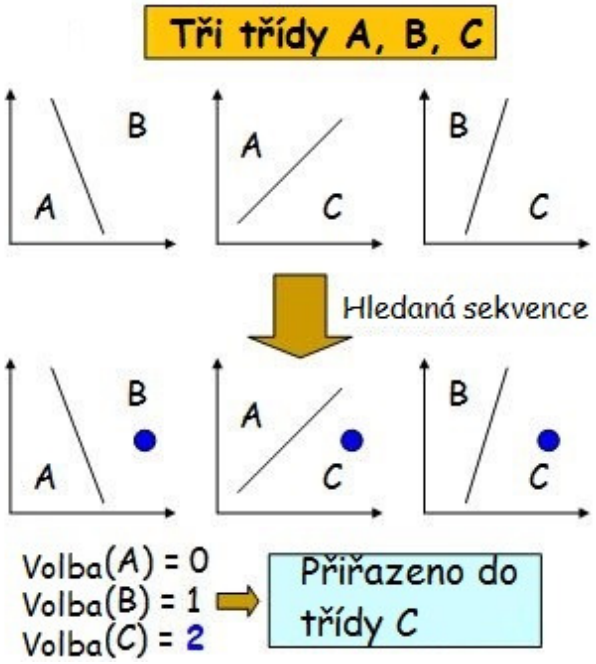
Phylopythia

Metoda Phylopythia [11] používá strojové učení SVM (z ang. support vector machine). Pro zjednodušení výpočtu používá Gausovskou jádrovou transformaci. Během fáze tréninku se metoda na datech různé délky učí rozlišovat mezi položkami dvou tříd. Tato metoda je zaměřena na klasifikaci spíše delších DNA fragmentů.

Při klasifikaci mezi více třídami se vytvoří $\frac{N \cdot (N-1)}{2}$ binárních klasifikátorů, jeden pro každou možnou dvojici tříd. Testovaná sekvence se zpracuje všemi klasifikátory. Sekvence ja pak zařazena do třídy, která byla klasifikátory zvolena nejčastěji. Takové přiřazení můžeme vidět na obrázku č. 3.2.

Phymm a PhymmBL

Phymm a PhymmBL [5] používá při klasifikaci IMMs (interpolované Markovovy modely). Pomocí nich metoda Phymm charakterizuje oligonukleotidy s proměnnou délkou, které jsou typické pro určité fylogenetické skupiny. Pokud hledáme neznámý druh tak hledáme na



Obrázek 3.2: Ukázka přiřazení metodou Phylopythia: Existují tři třídy A, B a C. Byly vytvořeny tři binární klasifikátory pro každou dvojici tříd. Hledaná sekvence byla přiřazena nejčastěji do třídy C, proto je hladaná sekvence klasifikována jako prvek třídy C. Převzato z [11].

vyšší taxanomické pozici (tzv. rank-flexible, viz podkapitola 3.1). PhymmBL je kombinace Phymm a BLAST vyhledávání nejlepší shody, která vede ke zlepšení přesnosti.

Interpolované Markovovy modely

Interpolované Markovovy modely jsou zobecněné modely, které kombinují pravděpodobnosti z kontextů různých délek. Např. v mikobiálním genomu jsou některé 5-mery příliš vzácné, zatímco některé 8-mery se vyskytují mnohem častěji a proto 8-merům přiřadíme větší váhu. Pokud všem k-merům mimo k = 5 přiřadíme váhu 0 vznikne řetězec pevného řádu. IMMs pak počítá lineární kombinaci pravděpodobností podle vah.

IMMs byly poprvé použity k identifikaci kódovacích regionů v metodě GLIMMER (Gene Locator and Interpolated Markov ModelER). Tato metoda se používá k hledání genů v prokaryotní DNA.

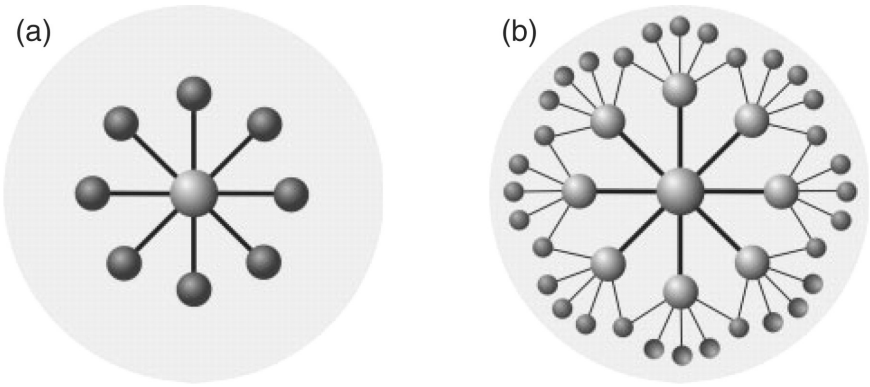
3.2 Taxonomicky nezávislé metody

Taxonomicky nezávislé metody shlukují sekvence na základě podobnosti mezi sebou. Výstupem těchto metod jsou obvykle hierarchické struktury sekvencí, ne konkrétní přiřazení do třídy.

Zde patří například metody využívající SOM (samoorganizující se mapy, z ang. self-organizing map).

H^{2}SOM

Tato metoda [10] využívá hierarchicky rostoucí hyperbolické samoorganizující se mapy (H^{2}SOMs, z ang. Hierarchical Hyperbolic Self-Organizing Map), které slouží ke shlukování fragmentů DNA malé proměnné délky a dělení na konkrétní taxonomické pozice. DNA fragmenty jsou mapovány do tří různých typů příznakových vektorů na základě genomového podpisu (základní vlastnosti, s ohledem na význam oligonukleotidových vzorů, kontrastní prvky). Každá kompletní genomická sekvence je mapována na příznakový vektor. Podle autorů je mnohem pravděpodobnější, že genomové sekvence jsou strukturované hierarchicky (podobně jako jsou seskupeny různé druhy, které jsou reprezentovány Stromem života) a takto organizovaná data rostou exponenciálně, a proto vyžadují mapování do geometrického prostoru s odpovídajícím chováním, H^{2}SOM tedy používá stromovou síť v hyperbolickém prostoru. Ta je zobrazena na obrázku 3.3.



Obrázek 3.3: Ukázka hierarchicky rostoucího schématu metody H^{2}SOMs: Inicializace kořenového uzlu (první podhierarchie) a potomků. Ke každému uzlu jsou přiřazeny příznakové vektory a váha uzlu, je nalezena nejlepší shoda s hledanou sekvencí. Každý potomek se stává kořenovým uzlem pro další potomky. Převzato z [10].

3.3 Validace metod

Validace metody se obvykle provádí na simulovaných datech a databázích. Je možné použít např. FAMeS, soubor dat obsahující 100 000 vzorků délky 650–1000 bázových párů. Tato data pocházejí ze 112 skutečných projektů zabývajících se sekvencováním genomů.

Taxonomicky závislé metody

Pro validaci těchto metod se používá simulovaných metagenomických datasetů a databází. Sekvence v těchto datech by měly simulovat délku a modelovat chyby vznikající při sekvencování. Často se provádí křížová validace. Ta zjišťuje, jak moc budou analýzu ovlivňovat nezávislé vzorky dat. Vstupní sada dat je rozdělena na podmnožiny. Jedna množina je použita k testování (stává se hledanou sekvencí) a zbývající pro trénink (referenční knihovna). Pak validovaná metoda klasifikuje testovací množinu. To se opakuje pro všechny možné testovací množiny. Při variantě leave-one-out se testovací množina skládá pouze z jednoho prvku. Znázornění metody leave-one-out lze vidět na obrázku 3.4. Použití této validace simuluje situaci, kdy je hledaná sekvence z genomu nezastoupeného v referenční knihovně. Je to dáno tím, že testovací a trénovací množiny jsou vždy vybrány tak, aby byly disjunktní.

Taxonomicky nezávislé metody

U těchto metod se provede shlukování na simulovaných datech a přesnost je pak dána parametry jako taxonomická stejnorodost shluků, jejich počet a velikost. V ideálním případě je počet shluků stejný jako počet druhů v množině trénovacích sekvencí.

3.4 Hodnocení metod

Účinnost metody je popsána dvěma parametry, jsou to sensitivita a specificita. Za přesné přiřazení se považuje takové, kdy je sekvence přiřazena ke správné taxonomické větvi.

Sensitivita je poměr správně přiřazených do třídy i a celkového počtu fragmentů ve třídě i. Pro výpočet se používá vzorec 3.1. Uvádí se v procetech.

$$Sn_i = \frac{TP_i}{TP_i + FN_i} \tag{3.1}$$

Specifita se počítá vzorcem 3.2, kde TN_i (z ang. true negative) jsou fragmenty, které nepocházejí z třídy i a ani tam nebyly přiřazeny. FP_i jsou fragmenty patřící do jiné třídy, ale nesprávně přiřazené do třídy i.

$$Sp_i = \frac{TN_i}{FP_i + TN_i} \tag{3.2}$$

Srovnání vybraných metod

Srovnat všechny metody uvedené v této kapitole není možné. Metody byly srovnávány na rozdílných taxonomických úrovních a byly využity jiné sady referenčních dat.

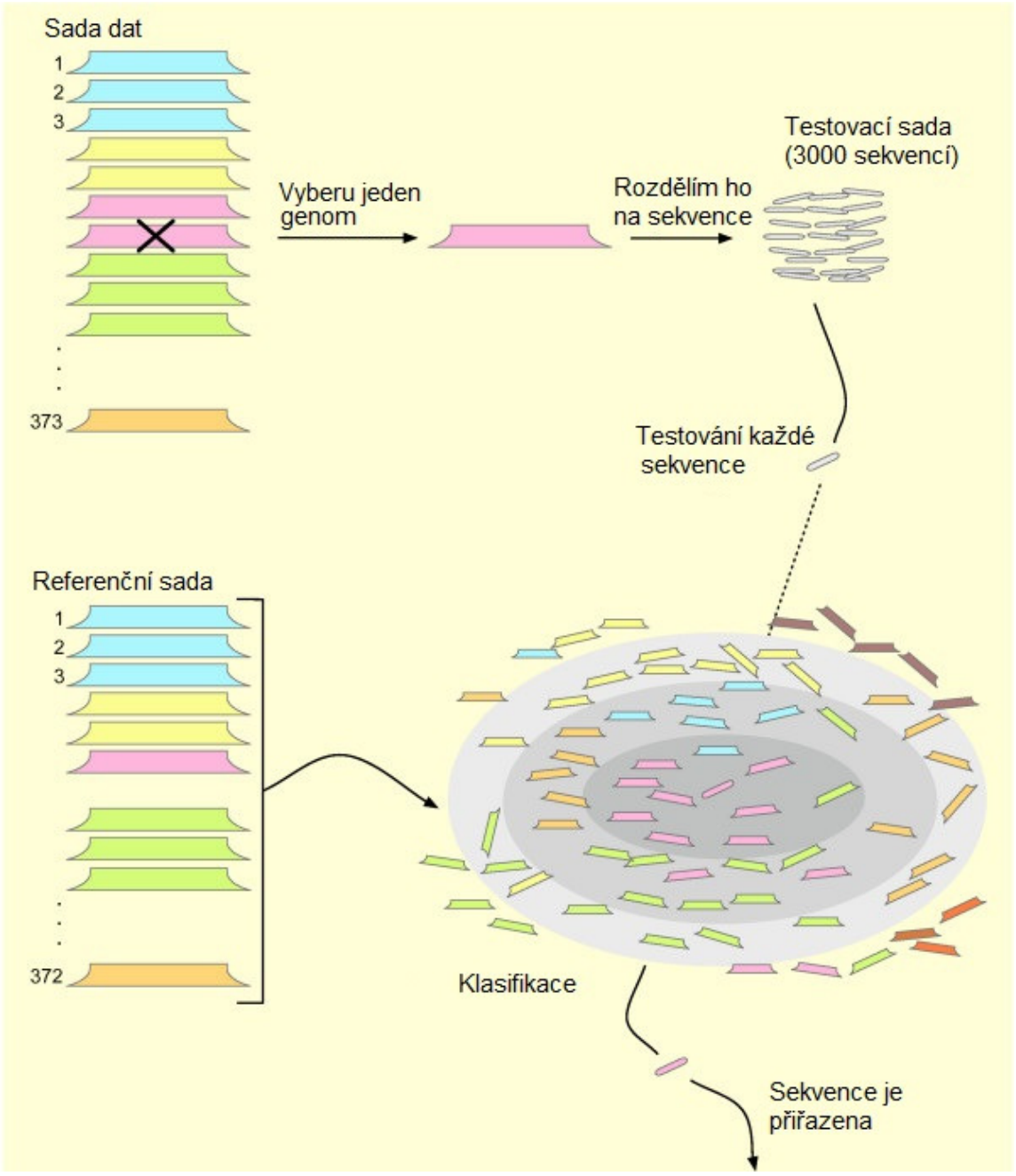
Algoritmus Eu-Detect [12] byl srovnáván s metodou TACOA na úrovni nadtřídy. Bylo použito 237 prokaryotických a 27 eukaryotických genomů, z těchto genomů vytvořily obě metody své referenční sady. Úspěšnost Eu-Detect byla 80–90%, úspěšnost metody TACOA byla 45–53%.

Pro srovnání metod Phymm [5], PhymmBL, Phylopythia a BLAST byly využity sekvence délky 1000 bázových párů z 539 genomů. Sekvence byly klasifikovány na několika taxonomických úrovních. Výsledky jsou zobrazeny v tabulce 3.1.

Tax.úroveň	Phymm	BLAST	PhymmBL	PhyloPythia
Řád	80.6	80.8	86.9	25.1
Třída	85.4	84.1	90.6	30.8
Kmen	89.8	88.0	93.8	50.3

Tabulka 3.1: Srovnání sensitivity metod. Převzato z [5].

Z výše uvedených metod je Eu-Detect nejjednodušší na implementaci a bez velkých požadavků na výpočetní zdroje. Eu-Detect vybrán jako základ pro metodu navrženou v této práci.



Obrázek 3.4: Validace leave-one-out: Je vybrán jeden genom (tzv. testovací množina), který je rozdělen na sekvence a ty jsou pomocí vybrané metody klasifikovány. To se opakuje dokud každý ze 373 genomů není vybrán jako testovací množina. Převzato z [7].

Kapitola 4

Návrh řešení

4.1 Detailní specifikace a formulace cíle

Cílem této práce je vytvořit rychlý klasifikátor, který bude pracovat na taxonomické úrovni nadříše, tzn. archea, prokaryota a eukaryota. Základem pro tuto metodu se stal algoritmus Eu-Detect popsaný v kapitole 3.1.

Metoda navržená v této práci bude rozšířena o klasifikaci do nadříše archea narozdíl od metody Eu-Detect, která rozlišuje pouze eukaryota a prokaryota.

4.2 Algoritmus

Algoritmus je rozdělen do dvou částí. Při první části je vytvořena referenční sada dat a hledány suboptimální hodnoty parametrů. V druhé části je provedena samotná klasifikace do nadříší.

Na obrázku 4.1 na straně 15 je možné vidět vývojový diagram pro tento algoritmus.

4.3 Trénink

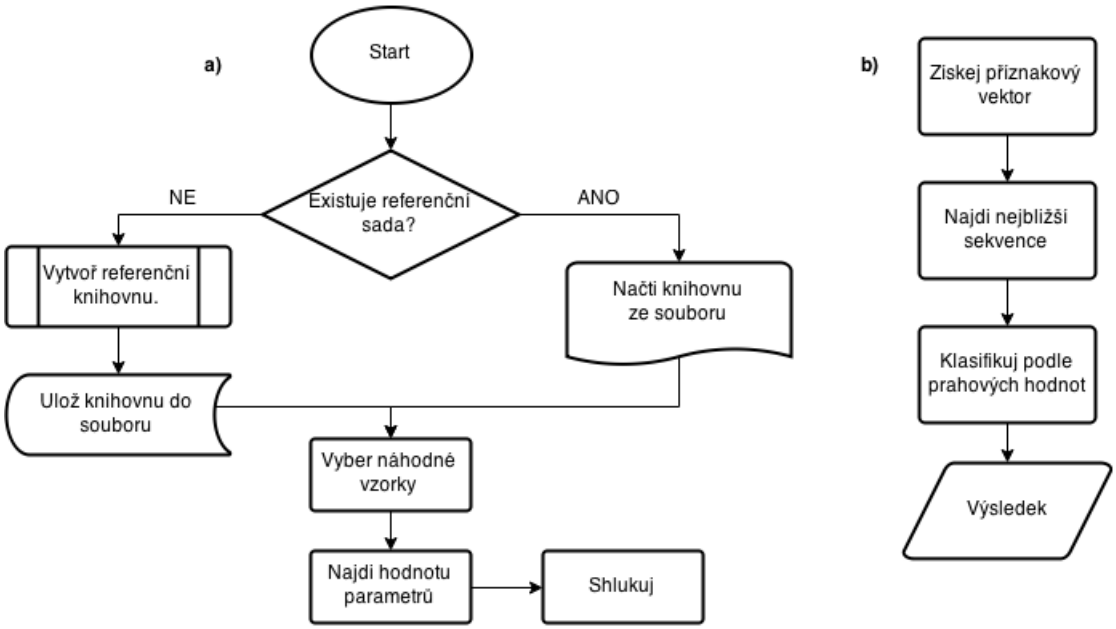
Vstupem této části programu jsou soubory se známými genomy. Výstupem budou shluky referenčních dat, informace o centrech těchto shluků a suboptimální hodnoty parametrů. Trénink se skládá ze tří fázi. V první fázi probíhá příprava příznakových vektorů, ve druhé jsou tyto příznakové vektory vzorkovány a v poslední části se provádí shlukování.

Příprava příznakových vektorů

Vstupem této části jsou známé genomy eukaryotických, prokaryotických a archea organismů. Tyto genomy jsou v souborech typu FASTA. Každý soubor je přečten a rozdělěn na fragmenty po deseti tisících bázových párech. Počet fragmentů je přímoúměrný délce genomu. Z každého fragmentu se vytvoří jeden příznakový vektor (GFV z ang. Genomic Feature Vector), který obsahuje frekvenci jednotlivých tetra-oligonukleotidů. Počet jednotlivých tetra-oligonukleotidů je 4^4 (jednotlivé nukleotidy jsou A, C, G, T a hledají se všechny čtveřice), takže počet dimenzí příznakového vektoru je 256.

$$GVT = (|AAAA|, |AAAC|, |AAAG|, |AAAT|, ...|TTTT|)$$

Výstupem je pole obsahující tyto příznakové vektory, které lze považovat za referenční knihovnu.



Obrázek 4.1: Vývojový diagram: Část a) obsahuje přípravu příznakových vektorů a nalezení vhodných parametrů. Část b) je samotná klasifikace do nadříší podle hodnoty prahů.

Vzorkování

Cílem vzorkování je vytvořit sadu referenčních dat o vhodné velikosti a vyváženosti jednotlivých nadříší. Vstupem je pole příznakových vektorů z předchozí části. Ve fázi vzorkování se vybere určitý počet náhodných fragmentů z každého genomu. Výstupem je menší pole příznakových vektorů se stejným počtem fragmetů pro jednotlivé genomy.

Ve většině případů se vzorkování provádí před počítáním příznakových vektorů. V této situaci je příprava příznakových vektorů časově náročná a při experimentech (viz 6.2) bylo využito velké množství dat, bylo tedy vhodné provést tyto fáze v jiném pořadí.

Identifikace shluků

Cílem této fáze je vytvořit shluky podobných fragmentů. V další fázi se porovnává vzdálenost mezi hledanou sekvencí a těmito centry, to je časově méně náročné než porovnávání vzdálenosti se všemi refenčními vektory.

Vstupem je pole příznakových vektorů. V této fázi jsou spočítány některé důležité hodnoty. Je to počet shluků k, který je dán vzorcem $k \approx \sqrt{\frac{n}{2}}$, kde n je počet fragmentů v referenčním poli. Dále je provedeno shlukování metodou k-means [17], konkrétně variantou Forgy a jsou dány souřadnice středů shluků. Vznikají shluky složené z pouze/většiny sekvencí jedné nadříše nebo shluky, kde je srovnatelný počet sekvencí více nadříší. Výstupem tohoto kroku budou shluky referenčních vzorů a souřadnice center těchto shluků.

4.4 Klasifikace

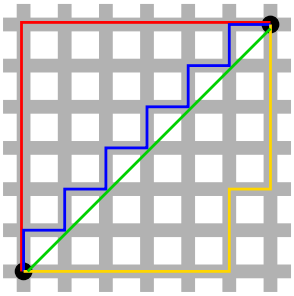
V této fázi nastává samotná klasifikace. Porovnává se vzdálenost hledané sekvence převedené do příznakového vektoru a jednotlivých center shluků, z předchozího kroku. Dále

se hledají nejbližší shluky. V nejbližších shlucích se spočítá poměr fragmentů jednotlivých nadříší. Pokud některá z těchto hodnot překračuje práh pokrytí (z ang. coverage treshold) je hledaná sekvence přiřazena k této nadříši. Existují tři prahy pokrytí pro každou nadříši. Pokud nepřekročí žádný práh pokrytí je označena jako neznámá. Pokud překročí více prahů je přiřazen k nejvyššímu. Jednotlivé prahy pokrytí byly rovněž zjišťovány experimentálně.

Jako měřítko podobnosti pro shlukování mezi jednotlivými vektory byla použita Manhattanská metrika pro měření vzdálenosti od vstupu do těžišť shluků. Ta je dána vzorcem:

$$d(p,q) = \sum_{i=1}^{n} |p_i - q_i|,$$

kde p je příznakový vektor hledané sekvence $p = (p_1, p_2, ...p_{256})$ a q je příznaková vektor středu shluku $q = (q_1, q_2, ...q_{256})$. Vzdálenost je podle tohoto vzorce spočítána postupně pro každý střed. Rozdíl mezi Manhattanskou a Euklidovskou metrikou je možné vidět na obrázku 4.2.



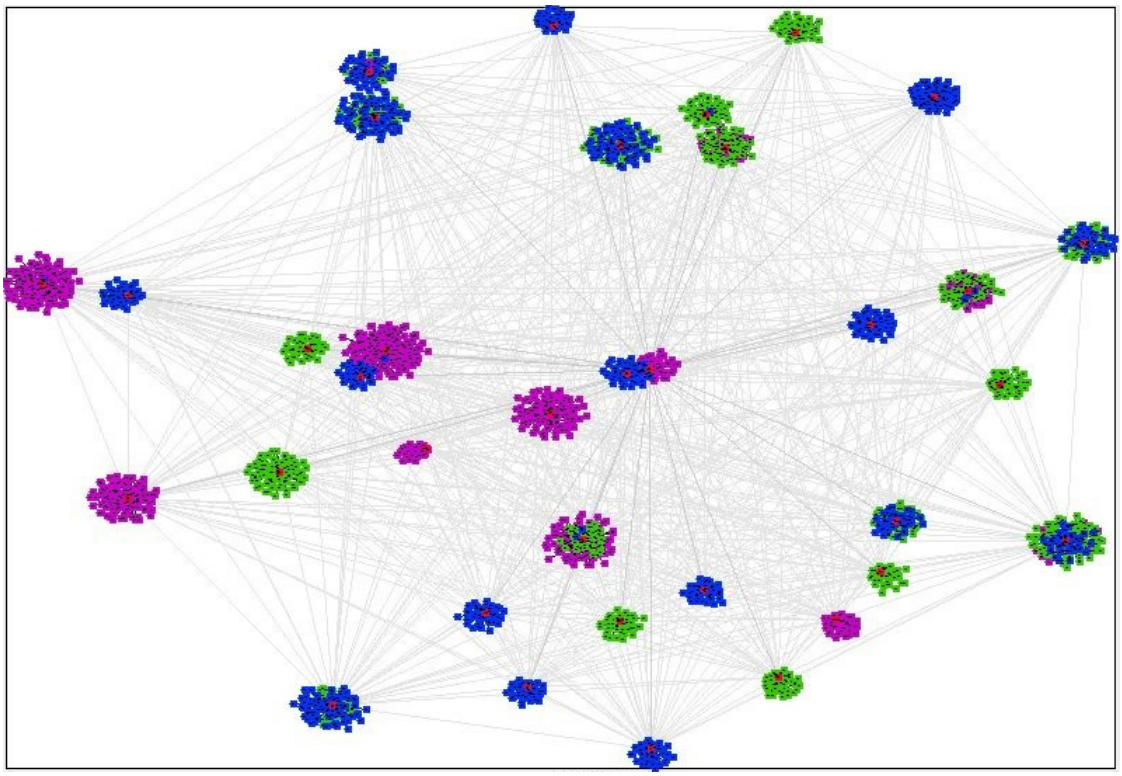
Obrázek 4.2: Manhattanská a Euklidovská metrika: Tři linky označující Manhattanskou metriku (červená, žlutá a modrá) mají všechny vzdálenost 12. Euklidovská metrika je představována zelenou čárou, její délka je $6\sqrt{2} \approx 8,49$ a je vždy nejkratší. [2]

Dostatečný počet nejbližších shluků je určen souhrnným počtem sekvencí *csc* (z ang. cumulative sequence count). Suboptimální hodnota tohoto parametru byla zjišťována experimentálně. Prostor shluků je znázorněn na obrázku 4.3.

Hodnoty prahů jsou dány poměrem jednotlivých nadtříd v referenční sadě dat. Konkrétní hodnoty jsou získany vzorcem 4.1.

$$treshold_i = \frac{pocet\_sekvenci\_nadrise\_i}{celkovy\_pocet\_sekvenci} \tag{4.1}$$

[2] Převzato z http://en.wikipedia.org/wiki/Taxicab_geometry.



Obrázek 4.3: Prostor nashlukovaných fragmentů: Červeně jsou znázorněna centra shluků, fialově fragmenty eukaryotických sekvencí, zeleně prokaryotických sekvencí, modře je archea. Data pocházejí z prvního měření experimentu 1 (viz 6.2). Obrázek byl vytvořen v programu [8].

Kapitola 5

Implementace

5.1 Nástroje použité k realizaci

Tento projekt byl vypracován v programovacím jazyce R. Jazyk R se nejčastěji využívá pro statistické výpočty, datovou analýzu a grafiku. Tento jazyk obvykle používá rozhraní příkazové řádky.

Jako uživatelské rozhraní pro jazyk R bylo použito RStudio Version 0.98.501 - ©2009-2013 RStudio, Inc. RStudio je integrované vývojové prostředí (IDE) pro jazyk R.

K provádění experimentů bylo použito prostředí MetaCentrum VO. Virtuální organizace MetaCentrum VO je jedna z virtuálních organizací české Národní Gridové Iniciativy MetaCentrum NGI, aktivity sdružení CESNET, z.s.p.o. MetaCentrum koordinuje pořizování nákladného programového a aplikačního vybavení, zajišťuje centrální nákup a správu licencí vybraného licencovaného SW. Také se podílí na výzkumu a vývoji nezbytného k zajištění optimální funkcionality, bezpečnosti a vysokého výkonu celé infrastruktury. [6]

Použité knihovny jazyka R

V této praci je používán Bioconductor. Bioconductor je softwarový projekt, který poskytuje nástroje pro analýzu a porozumění genomických dat. Konkrétně je využívána knihovna Biostrings. Tato knihovna pracuje se speciálními paměťově efektivními řetězeci, které představují biologické sekvence. Obsahuje funkce pro srovnávání těchto řetězců a další nástroje pro rychlou manipulaci velkých biologických sekvencí nebo sad sekvencí.[14]

5.2 Formát vstupních dat

Vstupní data jsou ve formátu FASTA. Tento v bioinformatice často používaný formát je podporován knihovnou Biostrings. Každý soubor reprezentuje jednu nebo více nukleotidových sekvencí jednoho organismu. Každá sekvence v souboru FASTA začíná jednořádkovým popiskem. Na začátku tohoto popisu je znak >. Následující slovo je název sekvence, který slouží jako její identifikátor. Následují písmena reprezentující jednotlivé nukleové báze DNA, tzn. adenin (A), cytosin (C), guanin (G) a thymin (T). Může obsahovat i písmeno N nebo X pro neznámou nukleovou bázi. Příklad takového souboru je na obrázku 5.1.

```
>aerPer1i_dna range=chr:1-1669696 5'pad=0 3'pad=0 strand=+
AAATAATAATAAAAATTAAGTGACTCATGCATTATCCTACGAGGTAAAAA
TATGTTATAAATTGTCCCAGACTACCATCAATTTAGGGACAATAGTGTTT
AAGGGATGGCCTTCGGAGCTGGCAGCTCGCGGGTTCAAACTCGCGTAGGG
...
```

Obrázek 5.1: Příklad části FASTA souboru pro organismus Aeropyrum pernix (Archaea). Identifikátor této sekvence je aerPer1i_dna.

5.3 Popis funkcí

Implementace této metody obsahuje čtýři funkce. Tyto funkce a jejich vstupy a výstupy jsou znázorněny na obrázku 5.2.

První funkce se jmenuje `make_lib`. Cílem této funkce je vytvoření referenční knihovny. Jejím vstupem je řetězec označující cestu ke složce, která obsahuje pouze soubory typu FASTA a ze kterých tato referenční knihovna vznikne. Tyto soubory by měly být vhodně pojmenovány. Na začátku musí být řetězec identifikující nadtřídu ar-, ba- nebo eu-. Další znaky mohou být libovolné. Příkladem správného názvu je ar-cr-001.fa.

Tato funkce přečte DNA řetězec ze souboru, rozdělí ho na sekvence po 10 000 bázových párech a nalezne frekvence všech možných tetra-oligonukleotidů, tím vzniknou příznakové vektory. Po úspěšném provedení této funkce vzniknou ve výstupní složce tři soubory. Soubor data, který obsahuje všechny příznakové vektory. V souboru data_name jsou uloženy jména jednotlivých sekvencí, které se skládají z pořadového čísla a názvu původního genomu. V souboru data_list se nachází počet sekvencí vytvořených z každého genomu. Tyto tři soubory se vytvoří ve složce *classDNAlib* a jsou důležité pro další výpočet. Tato funkce je časově nejnáročnější, ale je možné použít předpřipravenou knihovnu z přiloženého CD.

Dále zde existují dvě různé funkce na vzorkování dat `sampl_1` a `sampl_2`. Vstupem každé této funkce je řetězec *classDNAlib*. To je složka obsahující soubory data, data_list a data_name, které vznikly v předchozí části. Dalším parametrem je hodnota samples. Ve funkci `sampl_1` hodnota samples označuje počet vzorků, které jsou náhodně vybrány z každého genomu. Doporučená hodnota pro tuto funkci je 30. Ve funkci `sampl_2` je to procentuální hodnota vzorků, pro vloženou knihovnu je vhodná hodnota 10–20%. Výstupem jsou složky *classDNAlib_sampl1* nebo *classDNAlib_sampl2* podle toho, která vzorkovací funkce byla použita. Tyto složky obsahují podobné soubory jako vznikly v předchozím kroku. Tato funkce vrací jako výstup seznam hodnot vhodných pro prahy.

Ve funkci `train` je prováděno shlukování referenčních vektorů a hledání souřadnic středů těchto shluků. Vstupem této funkce je soubor sampl_data.

Poslední funkcí je `classify`. Vstupem jsou opět soubory vzniklé ve vzorkovacích funkcích, parametr *csc* určující počet nejbližších shluků, středy shluků a jméno FASTA souboru genomu, který bude klasifikován. V této funkci se nachází samotná klasifikace podle hodnot jednolivých prahů. Výstupem je řetězec se zařazením do taxonomické kategorie.

Spuštění

Pro použití knihovny navržené v této práci je nutné mít nainstalovanou knihovnu Biostrings. Instalace knihovny jazyka R se provádí příkazem

```
R CMD INSTALL path/classDNAseq/,
```

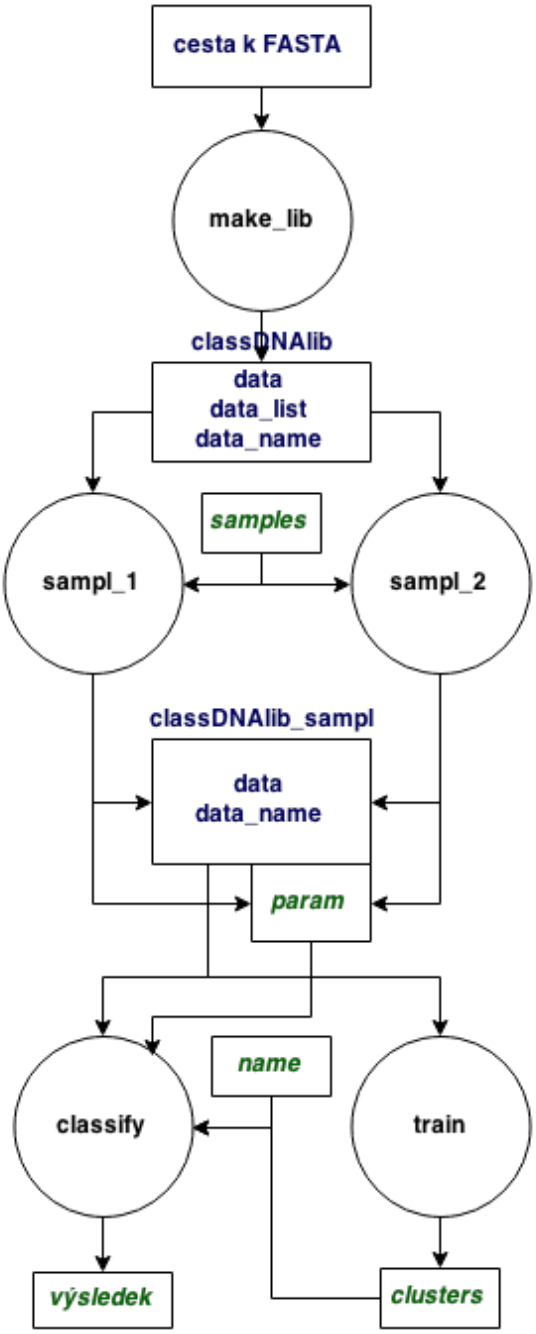
kde `path` označuje cestu ke složce, ve které je knihovna umístěna. v RStudiu je nutné obě knihovny načíst pomocí příkazu `library`.

```
library(Biostrings)
library(classDNAseq)
```

Příklad použití s předpřipravenou knihovnou.

```
p<-sampl_1("knihovna",20)
cl<-train("classDNAlib_saml1")
classify("jmeno_souboru","classDNAlib_saml1",p,10,cl)
```

Na přiloženém CD se nacházejí *skript.R* a *skript2.R*. *Skript.R* obsahuje příklad vytvoření referenční sady příznakových vektorů a vzorkování. Z důvodu časové náročnosti těchto činností pracuje *skript.R* jen s devíti soubory. *Skript2.R* ukazuje trénování a klasifikaci. Používá předpřipravenou referenční sadu, protože sada vytvořená prvním skriptem je příliš malá. Pro další příklady je možné v jazyce R použít např. příkaz `?make_lib`.



Obrázek 5.2: Diagram: Černě jsou označeny funkce, modře soubory a složky, zeleně parametry, červeně knihovna jazyka R.

Kapitola 6

Výsledky měření

Implementace metody navržené v této práci byla testována na reálných sadách dat. Pro vytvoření referenční sady dat obou experimentů bylo použito 20 genomů z organismů každé nadříše, tedy celkem 60 různých genomů. Z nich bylo vypočítáno celkem 96 139 příznakových vektorů.

Tyto genomy byly získány z databáze UCSC Genome Browser a UCSC Microbial Genome Browser. Bližší popis jednotlivých druhů je uveden v příloze B. Tato knihovna se pro další použití nachází na přiloženém CD ve složce *knihovna*.

6.1 Zjištěné hodnoty parametru

Parametr *csc* (z ang. cumulative sequence count) je hodnota určující počet nejbližších fragmentů, které se uvažují při klasifikaci. Vhodná hodnota parametru *csc* byla nalezena mřížkovou metodou. Pro získání byla použita datová sada obsahující 800 eukaryotických fragmentů, 800 prokaryotických fragmentů a 800 fragmentů archea. Bylo zjištěno, že suboptimální hodnota je 100 fragmenů. Tabulka 6.1 udává sensitivitu pro jednotlivé hodnoty *csc*.

csc	Sn	*csc*	Sn
35	0,520	80	0,692
40	0,603	85	0,667
45	0,616	90	0,688
50	0,674	95	0,704
55	0,628	100	0,744
60	0,686	105	0,691
65	0,651	110	0,713
70	0,675	115	0,695
75	0,704	120	0,711

Tabulka 6.1: Tabulka zjištěných hodnot parametru *csc*: Sn je sensitivita, *csc* je počet nejbližších sekvencí.

6.2 Experiment 1

Každý genom byl v tréninkové fázi (viz 4.3) rozdělen na sekvence po 10 000 bázových párech. Počet sekvencí pro jednotlivé genomy se pohybuje od 62 do 20 485. Následně bylo při vzorkování náhodně vybráno 40 sekvencí pro každý genom. To je celkem 2400 příznakových vektorů pro další zpracování. Toto vzorkování a následující měření proběhlo celkem desetkrát. Práh pro klasifikaci do nadtřídy je stejný pro všechny tři nadtřídy a je to 0,33%.

Validace metodou leave-one-out (popsána v kapitole 3.3) trvala průměrně 113 minut, tzn. jedna sekvence byla průměrně klasifikována za 2,8 sekund. Správně zařazených sekvencí bylo 1711, to je přibližně 71,3%. V tabulce 6.2 jsou vidět detailnější výsledky měření, např. počet správně zařazených vzorků typu eukaryota atd.

	TP	TP_A	FN_A	FP_A	TP_P	FN_P	FP_P	TP_E	FN_E	FP_E	U
1	1734	470	330	85	581	219	244	692	108	137	191
2	1693	467	333	94	559	241	258	667	133	183	172
3	1645	444	356	92	556	244	247	645	155	132	284
4	1795	511	289	91	598	202	232	686	114	119	163
5	1681	475	325	112	543	257	223	663	137	115	269
6	1744	469	331	104	535	265	267	740	60	100	185
7	1696	471	329	120	552	248	251	673	127	114	219
8	1735	467	333	126	551	249	263	717	83	123	153
9	1722	485	315	122	539	261	239	698	102	171	146
10	1660	485	315	117	450	350	206	725	75	121	296
P:	1710,5	474,4	325,6	106,3	546,4	253,6	243	690,6	109,4	131,5	207,8

Tabulka 6.2: Tabulka výsledků deseti měření exp. 1: TP (z ang. true positive) označuje počet správně zařazených sekvencí do nadříše, FP (z ang. false positive) je počet sekvencí, které byly do nadříše zařazeny chybně, FN (z ang. false negative) počet sekvencí, které do nadříše patří, ale zařazeny tam nebyly, index označuje nadříši, U (z ang. unknown) představuje počet nezařazených sekvencí.

Sensitivita metody pro tento experiment byla 71,3% a specificita 89,98%. V tabulce 6.3 jsou hodnoty sensitivity Sn a specificity Sp pro jednotlivé nadříše a Matthewsův korelační koeficient [4]. Časová náročnost se nachází v tabulce 6.6.

6.3 Experiment 2

Pro experiment 2 byla použita stejná referenční datová sada. Narozdíl od experimentu 1 byl zachován poměr délek jednotlivých genomů z této sady. Ve vzorkování bylo z každého genomu náhodně vybráno 10% ze všech jeho sekvencí. Vznikla referenční sada obsahujícíc 9615 sekvencí délky 10 000 bázových párů.

Pro druhý experiment je hodnota pro práh eukaryotických sekvencí 88%, práh archea 4%, práh prokaryotických sekvencí 8%.

Naměřená sensitivita byla 57,98% a specificita 93,92% (podrobnější informace o naměřených hodnotách jsou v tabulkách 6.4 a 6.5). Údaje o časové složitosti se nacházejí v tabulce 6.6.

	Sn_A	Sp_A	Sn_P	Sp_P	Sn_E	Sp_E	MCC
1	0,588	0,947	0,726	0,848	0,865	0,914	0,487
2	0,584	0,941	0,699	0,839	0,834	0,886	0,430
3	0,555	0,943	0,695	0,846	0,806	0,918	0,541
4	0,639	0,943	0,748	0,855	0,858	0,926	0,465
5	0,594	0,930	0,679	0,861	0,829	0,928	0,543
6	0,586	0,935	0,669	0,833	0,925	0,938	0,471
7	0,589	0,925	0,690	0,843	0,841	0,929	0,492
8	0,584	0,921	0,689	0,836	0,896	0,923	0,421
9	0,606	0,924	0,674	0,851	0,873	0,893	0,406
10	0,606	0,927	0,563	0,871	0,906	0,924	0,288
P:	0,593	0,934	0,683	0,848	0,863	0,918	0,454

Tabulka 6.3: Tabulka procentuálního zařazení exp. 1: Sn je sensitivita, Sp specificita, MCC [4] je Matthewsův korelační koeficient a index označuje nadříši.

	TP	TP_A	FN_A	FP_A	TP_P	FN_P	FP_P	TP_E	FN_E	FP_E	U
1	7938	57	328	18	494	290	463	7387	1059	151	1045
2	8071	58	327	12	536	248	227	7477	969	166	1139
3	8275	54	331	47	441	343	276	7780	666	177	840
4	8074	50	335	35	555	229	351	7469	977	190	965
5	8323	57	328	16	565	219	440	7701	745	108	728
6	8310	71	314	5	576	208	350	7663	783	153	797
7	8294	72	313	17	554	230	331	7668	778	138	835
8	8571	63	322	37	517	267	326	7991	455	137	544
9	7854	79	306	29	489	295	888	7286	1160	201	643
10	8214	69	316	21	572	212	296	7573	873	160	924
P:	8192,4	63	322	23,7	529,9	254,1	394,8	7599,5	846,5	158,1	846

Tabulka 6.4: Tabulka výsledků deseti měření exp. 2: TP (z ang. true positive) označuje počet správně zařazených sekvencí do nadříše, FP (z ang. false positive) je počet sekvencí, které byly do nadříše zařazeny chybně, FN (z ang. false negative) počet sekvencí, které do nadříše patří, ale zařazeny tam nebyly, index označuje nadříši, U (z ang. unknown) představuje počet nezařazených sekvencí.

	Sn_A	Sp_A	Sn_P	Sp_P	Sn_E	Sp_E	MCC
1	0,148	0,998	0,630	0,948	0,875	0,871	0,760
2	0,151	0,999	0,684	0,974	0,885	0,858	0,838
3	0,140	0,995	0,563	0,969	0,921	0,849	0,769
4	0,130	0,996	0,708	0,960	0,884	0,837	0,765
5	0,148	0,998	0,721	0,950	0,912	0,908	0,726
6	0,184	0,999	0,735	0,960	0,907	0,869	0,759
7	0,187	0,998	0,707	0,963	0,908	0,882	0,773
8	0,164	0,996	0,659	0,963	0,946	0,883	0,702
9	0,205	0,997	0,624	0,899	0,863	0,828	0,565
10	0,179	0,998	0,730	0,966	0,897	0,863	0,765
P:	0,164	0,997	0,676	0,955	0,900	0,865	0,742

Tabulka 6.5: Tabulka procentuálního zařazení exp. 2: Sn je sensitivita, Sp specificita, MCC [4] je Matthewsův korelační koeficient a index označuje nadříši.

	T_1 (s)	T_1 (min)	t_1	T_2 (s)	T_2 (hod)	t_2
1	7392,443	123,207	3,080	45956,461	12,766	4,780
2	6671,729	111,195	2,780	45012,966	12,504	4,682
3	6929,857	115,498	2,887	45355,650	12,599	4,717
4	6661,625	111,027	2,776	43622,370	12,117	4,537
5	6329,689	105,495	2,637	31931,594	8,870	3,321
6	6978,219	116,304	2,908	31234,787	8,676	3,249
7	6431,629	107,194	2,680	27985,138	7,774	2,911
8	6188,175	103,136	2,578	27754,183	7,709	2,887
9	7030,985	117,183	2,930	32074,273	8,910	3,336
10	7320,860	122,014	3,050	31062,572	8,628	3,231
P:	6793,521	113,225	2,831	36198,999	10,055	3,765

Tabulka 6.6: Tabulka časové náročnosti: T je čas, za který byly přiřazeny všechny sekvence (experiment 1 obsahoval 2400 sekvencí, experiment 2 celkem 9615 sekvencí), t je průměrný čas klasifikace jedné sekvence v sekundách. Index označuje experiment.

Kapitola 7

Závěr

Klasifikace DNA sekvencí hraje důležitou roli v bioinformatice a dalších oborech. Využívá se hlavně při identifikaci a třídění organismů. Tato práce obsahuje seznámení se se základními principy molekulární biologie a způsoby uložení informace v DNA sekvencích. Je popsána charakteristika nadříší a hlavní rozdíly mezi nimi. Součástí práce je popis existujících algoritmů pro klasifikaci DNA a jejich hodnocení z hlediska sensitivity.

Byla vytvořena nová metoda pro klasifikaci na základě metody Eu-Detect. Nová metoda byla rošířena o klasifikaci do nadříše archea. Funkčnost metody byla ověřena na dvou sadách reálných dat. V experimentu 1 (viz 6.2) byla použita sada obsahující eukaryotické, prokaryotické a archea genomy v poměru 1:1:1. Zjištěná průměrná sensitivita pro nadříši archea byla 59%, prokaryota 68% a eukaryota 86%. Specificita v tomto experimentu se pohybovala od 83% do 94%. Pro experiment 2 (viz 6.3) byla využita eukaryotická, prokaryotická a archea data v poměru 22:2:1. Tento poměr je dán délkou genomu. Naměřená průměrná sensitivita pro všechny nadříše byla 58% a specificita 94%. Zařazení jedné sekvence trvalo průměrně 3,3 sekundy.

Na základě získaných poznatků by bylo vhodné metodu rozšířit o klasifikaci na další taxonomické úrovně.

Literatura

[1] Eukaryote / eucariote [online]. http://www.nature.com/scitable/definition/eukaryote-eucariote-294, 2014 [cit. 2015-05-03].

[2] Prokaryote / procariote [online]. http://www.nature.com/scitable/definition/prokaryote-procariote-18, 2014 [cit. 2015-05-03].

[3] Ahmed, N.: Archaea [online]. http://www.microbeworld.org/types-of-microbes/archaea, 2014 [cit. 201-05-03].

[4] Bekkar, M.; Djemaa, H.; Alitouche, T.: Evaluation Measures for Models Assessment over Imbalanced Data Sets. ročník 3, č. 10, 2013, ISSN 2225-0506 (online).

[5] Brady, A.; Salzberg, S.: Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. 2009;6(9):673-676, doi: 10.1038/nmeth.135819648916.

[6] CESNET: MetaCentrum NGI [online]. http://www.metacentrum.cz/cs/, 2015-01-30 [cit. 2015-04-25].

[7] Diaz, N.; Krause, L.; Goesmann, A.; aj.: TACOA - Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. 2009;10, doi: 10.1186/1471-2105-10-56.

[8] Frickey, T.; Lupas, A.: CLANS: a Java application for visualizing protein families based on pairwise similarity. 2004;20(18):3702-3704.

[9] Mande, S.; Mohammed, M.; Ghosh, T.: Classification of metagenomic sequences: methods and challenges. 2012;13:669.81, doi: 10.1093/bib/bbs054.

[10] Martin, C.; Diaz, N.; Ontrup, J.; aj.: Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification. 2008;24:1568-1574.

[11] McHardy, A.; Martin, H.; Tsirigos, A.; aj.: Accurate phylogenetic classification of variable-length DNA fragments. 2007;4(1):63-72, doi: 10.1038/nmeth976.

[12] Mohammed, H.; Chadaram, S.; Komanduri, D.; aj.: Eu-detect: an algorithm for detecting eukaryotic sequences in metagenomic data sets. 2011;36(4):709-717, doi: 10.1007/s12038-011-9105-2.

[13] Mount, D.: Using the Basic Local Alignment Search Tool (BLAST) [online]. 2007;2007(7):17.

[14] Pages, H.; Aboyoun, P.; Gentleman, R.; aj.: *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.30.1.

[15] Pati, A.; Heath, L.; Kyrpides, N.; aj.: ClaMS: a classifier for metagenomic sequences. 2011;5:248, doi: 10.4056/sigs.2075298.

[16] Rettner, R.: DNA: Definition, Structure & Discovery. 2013-06-06 [cit. 2015-05-03], `http://www.livescience.com/37247-dna.html`.

[17] Schuetze, H.: K-means [online]. `http://nlp.stanford.edu/IR-book/html/htmledition/k-means-1.html`, 2009-04-07 [cit. 2015-04-25].

Příloha A

Obsah CD

README	návod k instalaci a spuštění skriptů
skript.R	vytváří referenční knihovnu a používá vzorkování
skript2.R	identifikace shluků a klasifikace souborů
classDNAseq	složka knihovny jazyka R, obsahuje zdrojové kódy funkcí
fasta	obsahuje 9 příkladových slouborů, ze kterých se vytváří referenční knihovna ve skriptu 1
knihovna	tato knihovna byla použita k experimentům 1 a 2
knihovna 2 a 3	obsahuje referenční knihovny pro skript 2
priklady	obsahuje soubory určené ke klasifikaci skriptem 2

Příloha B

Genom

Archea	
Aeropyrum pernix	Hyperthermus butylicus
Metallosphaera sedula	Pyrobaculum arsenaticum
Thermoproteus neutrophilus	Sulfolobus acidocaldarius
Vulcanisaeta distributa	Cenarchaeum symbiosum
Nitrosopumilus maritimus	Archaeoglobus fulgidus DSM 4304
Archaeoglobus profundus DSM 5631	Ferroglobus placidus DSM 10642
Halalkalicoccus jeotgali B3	Methanocaldococcus sp. FS406-22
Methanocaldococcus fervens AG86	Methanocaldococcus vulcanius M7
Methanococcus maripaludis C5	Methanococcus vannielii SB
Picrophilus torridus DSM 9790	Pyrococcus abyssi GE5
Prokaryota	
Bifidobacterium longum	Clavibacter michiganensis NCPPB
Rhodococcus RHA1	Gloeobacter violaceus
Prochlorococcus marinus	Trichodesmium erythraeum
Bacillus subtilis	Bacillus halodurans
Listeria innocua	Rhizobium etli CFN 42
Zymomonas mobilis subsp. mobilis ZM4	Xanthobacter autotrophicus Py2
Sinorhizobium meliloti 1021	Roseobacter denitrificans OCh 114
Caulobacter crescentus CB15	Hydrogenobaculum sp. Y04AAS1
Sulfurihydrogenibium sp. YO3AOP1	Porphyromonas gingivalis W83
Gramella forsetii KT0803	Flavobacterium johnsoniae UW101
Eukaryota	
C. remanei	C. japonica
Saccharomyces cerevisiae	Pristionchus pacificus
melanogaster group D. erecta	melanogaster group Drosophila yakuba
C. brenneri	C. briggsae
C. elegans	A. gambiae
elephant shark	D. simulans
guinea pig	Nile tilapia
american alligator	alpaca
atlantic cod	baboon
tree shrew	chimp