VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ
DEPARTMENT OF COMPUTER SYSTEMS

KLASIFIKACE BAKTERIÍ DO TAXONOMICKÝCH KATEGORIÍ NA ZÁKLADĚ VLASTNOSTÍ 16S RRNA

BACTERIA CLASSIFICATION INTO TAXONOMIC CATEGORIES BASED ON PROPERTIES OF 16S RRNA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE **Bc. KATARÍNA GREŠOVÁ**
AUTHOR

VEDOUCÍ PRÁCE **Ing. STANISLAV SMATANA**
SUPERVISOR

BRNO 2020

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačových systémů (UPSY) Akademický rok 2019/2020

Zadání diplomové práce

Studentka: **Grešová Katarína, Bc.**
Program: Informační technologie Obor: Bioinformatika a biocomputing
Název: **Klasifikace bakterií do taxonomických kategorií na základě vlastností 16s rRNA**
Bacteria Classification into Taxonomic Categories Based on Properties of 16s rRNA
Kategorie: Bioinformatika
Zadání:

1. Nastudujte základní principy metagenomiky a její aplikaci na studium mikrobiomu pomocí amplikonového sekvencování genu 16s rRNA.

2. Zpracujte rešerši známých vlastností genu 16s rRNA (sekvenční varianty, strukturní variace atd.) a vyberte z nich ty, které by mohly být vhodné pro klasifikaci bakterií do taxonomických kategorií.

3. Navrhněte novou klasifikační metodu, která bude schopna klasifikovat sekvence genu 16s do taxonomických kategorií za pomoci některé vlastnosti identifikované v bodu 2. Výsledný nástroj by měl být schopen klasifikace na všech úrovních taxonomie.

4. Implementujte navržený nástroj a vyhodnoťte jeho přesnost na vhodné datové sadě.

5. Zhodnoťte dosažené výsledky a diskutujte možná budoucí rozšíření projektu.

Literatura:
- Dle pokynů vedoucího.

Při obhajobě semestrální části projektu je požadováno:
- Splnění bodů 1 až 3 zadání.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce: **Smatana Stanislav, Ing.**
Vedoucí ústavu: Sekanina Lukáš, prof. Ing., Ph.D.
Datum zadání: 1. listopadu 2019
Datum odevzdání: 3. června 2020
Datum schválení: 25. října 2019

Abstrakt

Hlavným cieľom tejto práce bolo navrhnúť a implementovať nástroj, ktorý by bol schopný klasifikovať sekvencie génu 16S rRNA do taxonomických kategórií s využitím vlastností génu 16S rRNA. Vytvorený nástroj analyzuje všetky vstupné sekvencie súčasne, čím sa líši od bežných klasifikačných prístupov, ktoré klasifikujú vstupné sekvencie jednotlivo. Tento nástroj využíva znalosť, že baktérie obsahujú niekoľko kópií génu 16S rRNA, ktoré sa môžu svojou sekvenciou odlišovať. Jedným z hlavných prínosov tejto práce je práve návrh, implementácia a vyhodnotenie schopností tohto nástroja. Experimenty ukázali, že navrhnutý nástroj je pre menšie dátové sady schopný identifikovať odpovedajúce baktérie a určiť správne pomery ich abundancií. Pri väčších dátových sadách sa však prehľadávaný priestor stáva veľmi rozsiahly a členitý, čo vyžaduje ďalšie vylepšenia navrhovaného nástroja, aby bol stavový priestor schopný prehľadávať efektívne.

Abstract

The main goal of this thesis was to design and implement a tool that would be able to classify the sequences of the 16S rRNA gene into taxonomic categories using the properties of the 16S rRNA gene. The created tool analyzes all input sequences simultaneously, which differs from common classification approaches, which classify input sequences individually. This tool relies on the fact that bacteria contain several copies of the 16S rRNA gene, which may differ in sequence. The main contribution of this work is design, implementation and evaluation of the capabilities of this tool. Experiments have shown that the proposed tool is able to identify the corresponding bacteria for smaller datasets and determine the correct ratios of their abundances. However, with larger datasets, the state space becomes very large and fragmented, which requires further improvements in order for it to search the state space in an efficient way.

Kľúčové slová

metagenomika, 16S rRNA, kópie génu 16S rRNA, sekundárna štruktúra RNA, taxonomická klasifikácia, Metropolis-Hastings

Keywords

metagenomics, 16S rRNA, copies of 16S rRNA gene, RNA secondary structure, taxonomic classification, Metropolis-Hastings

Citácia

GREŠOVÁ, Katarína. *Klasifikace bakterií do taxonomických kategorií na základě vlastností 16s rRNA*. Brno, 2020. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Stanislav Smatana

Klasifikace bakterií do taxonomických kategorií na základě vlastností 16s rRNA

Prehlásenie

Prehlasujem, že som túto diplomovú prácu vypracovala samostatne pod vedením pána Ing. Stanislava Smatanu. Uviedla som všetky literárne zdroje a publikácie, z ktorých som čerpala.

.....
Katarína Grešová
10. júna 2020

Poďakovanie

Rada by som poďakovala pánu Ing. Stanislavovi Smatanovi za odborné vedenie práce, trpezlivosť pri konzultáciách, priateľský prístup a pomoc, ktorú mi pri tvorbe práce poskytol.

Obsah

Kapitola 1

Úvod

Ľudské telo je domovom mnohých baktérií, archeí, vírusov a húb. Spoločenstvo týchto dodatočných buniek sa nazýva ľudský mikrobiom a ich počet je aspoň tak veľký, ako počet somatických buniek ľudského tela [98]. Súhrn genetickej informácie obsiahnutej v mikrobiálnych bunkách sa nazýva ľudský metagenóm a pozostáva z väčšieho počtu génov ako samotný ľudský genóm [44].

Prevažná časť týchto mikróbov obýva tráviaci trakt, pričom najviac sa ich nachádza v zostupnom hrubom čreve, kde syntetizujú nevyhnutné aminokyseliny a vitamíny a spracovávajú časti inak nestráviteľných zložiek potravy [21]. Nedávne technologické pokroky a výskumná snaha viedli k nárastu informácií v oblasti ľudského črevného mikrobiómu. Narušený črevný mikrobióm bol spojený s radou zdravotných ťažkostí ako napríklad obezita [74], Crohnova choroba [34], syndróm dráždivého čreva (IBS) [25], CDAD [26], psoriatická artritída [96], atopický ekzém [12], depresia [39] a iné.

Po dlhú dobu bolo možné analyzovať baktérie v ľudskom mikrobióme iba pomocou ich kultivácie. Mnohé druhy baktérií však nie sú kultivovateľné, a preto ich nebolo možné objaviť. Vďaka nedávnemu pokroku vo vysokovýkonnom sekvenovaní je teraz možné efektívne skúmať mikrobiálne spoločenstvá a analyzovať druhy baktérií, ktoré sa v nich nachádzajú [81].

Úlohou tejto práce je vytvoriť novú metódu klasifikácie baktérií, ktorá bude klasifikovať sekvencie génu 16S rRNA získané vysokovýkonným sekvenovaním. Vytvorený nástroj bude využívať vlastnosti zadaných sekvencií 16S rRNA na klasifikáciu baktérií do taxonomických kategórií a určovanie pomerov ich abundancií vo vstupnej vzorke.

Navrhnutý systém sa líši od bežných klasifikačných prístupov, ktoré klasifikujú vstupné sekvencie jednotlivo. Tento systém analyzuje všetky vstupné sekvencie súčasne a využíva znalosť, že baktérie obsahujú niekoľko kópií génu 16S rRNA, ktoré sa môžu svojou sekvenciou odlišovať [66]. Systém obsahuje niekoľko prístupov k ohodnoteniu podobnosti sekvencií, medzi ktoré patrí aj prístup založený na podobnosti sekundárnej štruktúry 16S rRNA. Celková štruktúra nástroja je uvedená v jeho špecifikácii, ktorá sa nachádza v kapitole 5.

Všetky špecifikované časti systému boli úspešne implementované a ich schopnosti boli vyhodnotené pomocou rady experimentov. Implementácia nástroja je prezentovaná v kapitole 6 a analýza výsledkov experimentov je popísaná v kapitole 7. Všetky teoretické informácie, ktoré sú potrebné na pochopenie princípov nástroja, sú uvedené v kapitolách 2, 3 a 4.

Kapitola 2

Genetická informácia živých organizmov

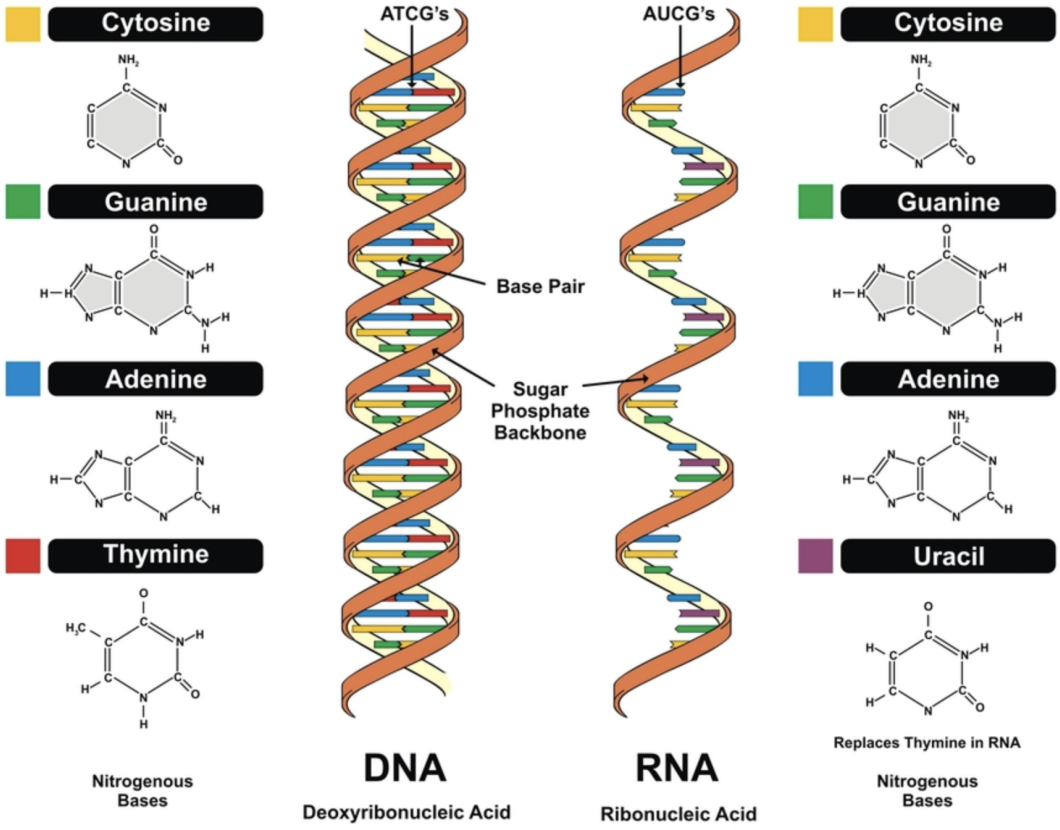
Povrch našej planéty je obývaný živými bytosťami – zložito organizovanými chemickými továrňami, ktoré prijímajú hmotu zo svojho okolia a používajú ju na vytváranie svojich kópií. Objavy minulého storočia čiastočne odhalili povahu týchto živých organizmov. Ukázalo sa, že živé organizmy sú zostavené z buniek, a že tieto jednotky života zdieľajú rovnaký mechanizmus na vykonávanie základných funkcií [16].

2.1 Štruktúra a funkcia nukleových kyselín

Živé bunky na Zemi uchovávajú svoju dedičnú informáciu vo forme dvojvlánkovej molekuly DNA. Oba vlánkna sú vždy vytvorené z rovnakých štyroch nukleotidov – základných stavebných jednotiek nukleových kyselín. Genetická informácia je zakódovaná lineárnou sekvenciou týchto nukleotidov, podobne ako postupnosť núl a jednotiek kóduje informáciu v počítačovom súbore. Každý nukleotid vo vlákne DNA pozostáva z troch častí: cukor (deoxyribóza), fosfátová skupina a dusíkatá báza, ktorá môže byť buď adenín (A), cytozín (C), guanín (G) alebo tymín (T) [16].

Sekvencia molekuly DNA je rozdelená na úseky – gény – ktoré kódujú informáciu na syntézu ďalších molekúl v bunke. Tento proces začína transkripciou, kedy je úsek DNA použitý ako šablóna na vytvorenie kratšej molekuly – RNA. Podobne ako v prípade DNA, kostra každej molekuly RNA je tvorená cukrom a fosfátovou skupinou, avšak je syntetizovaná s použitím iného cukru – ribózy. Ďalším rozdielom je, že RNA zvyčajne pozostáva z jedného vlákna a je podstatne kratšia ako molekula DNA. Navyše, RNA neobsahuje bázu tymín, ale bázu nazývanú uracil (U) [17]. Porovnanie DNA a RNA molekuly je na obrázku 2.1.

Jednou z hlavných funkcií RNA je okopírovať genetickú informáciu z DNA a fyzicky ju preniesť na miesto, kde dôjde k jej preloženiu na výsledný proteín. V procese prekladu je postupnosť nukleotidov RNA preložená do postupnosti aminokyselín – stavebných blokov proteínov. Celý proces je vykonávaný multimolekulárnym strojom – ribozómom. Ribozóm je tvorený dvoma hlavnými vláknami RNA, nazývanými ribozomálna RNA (rRNA), a viac ako 50 rôznymi proteínmi [16].



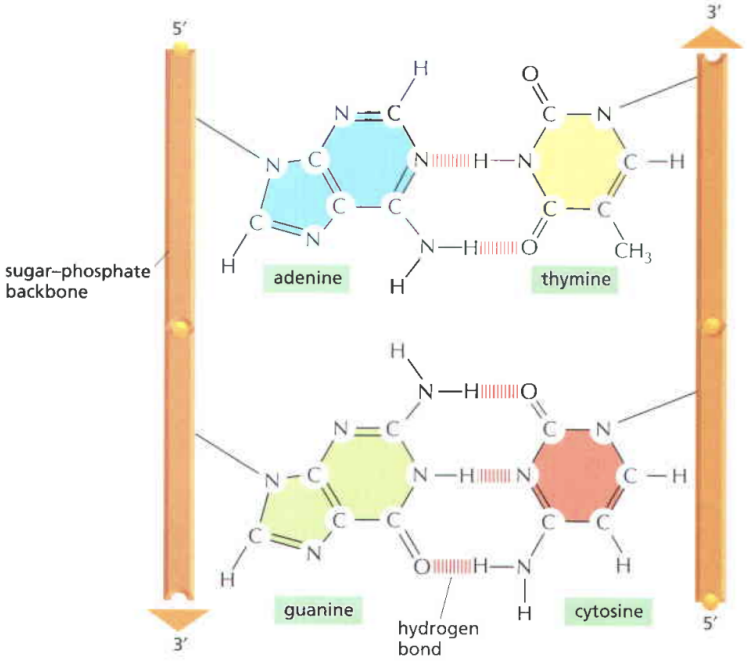
Obr. 2.1: Štruktúra DNA a RNA. DNA je dvojvlánková molekula, ktorá pozostáva z troch častí: cukor (deoxyribóza), fosfátová skupina a dusíkatá báza, ktorá môže byť buď adenín (A), cytozín (C), guanín (G) alebo tymín (T). RNA je jednovláknová molekula, ktorá ako cukor používa ribózu a neobsahuje bázu tymín, ale bázu uracil (U). Obrázok je prevzatý z článku *DNA: Definition, Structure & Discovery* od Rachael Rettnerovej [93].

2.2 Párovanie báz

Spôsoby, akými by molekula DNA mohla špecifikovať tvorbu proteínov, a zároveň ako by táto informácia mohla byť prenášaná z bunky do bunky, boli po dlhý čas záhadou. Riešenie sa objavilo v roku 1953, kedy James Watson a Francis Crick [113] správne predpovedali štruktúru DNA. Dvojvláknovosť molekuly DNA a komplementárne párovanie báz objasnilo problém kopírovania a replikácie informácie uloženej v DNA [16].

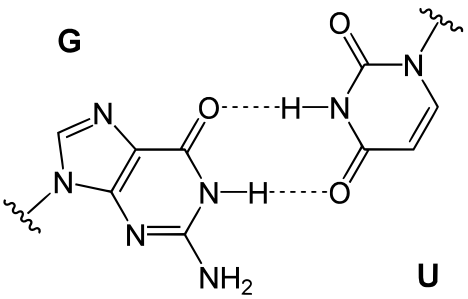
Ako bolo popísané v kapitole 2.1, molekula DNA pozostáva z dvoch dlhých vlákien tvorených štyrmi typmi nukleotidov. Tieto vlákna sú spojené vodíkovými väzbami medzi bázami opačných vlákien – bázy sú vo vnútri molekuly DNA a kostra je na povrchu. Párovanie báz však nie je náhodné. Podľa Watson-Crickovho modelu párovania báz sa A vždy páruje s T (alebo U v prípade RNA), C vždy s G (obrázok 2.2) [16].

Párovanie báz sa neprejavuje len v dvojvláknovej DNA, ale aj v jednovláknových molekulách DNA a RNA. Vlákno RNA sa rôzne stáča, vytvára páry vrámci tohto vlákna, a tým definuje sekundárnu štruktúru RNA. Rozsiahlejší popis sekundárnej štruktúry RNA sa nachádza v kapitole 2.3.



Obr. 2.2: Komplementárne párovanie báz v dvojvláknovej DNA. Tvar a chemická štruktúra báz dovoľuje vytvorenie účinných vodíkových väzieb len medzi A a T a medzi C a G. Obrázok je prevzatý z knihy od Albertsa a kolektívu [16].

Spočiatku sa predpokladalo, že párovanie báz v RNA podlieha výhradne Watson-Crickovým pravidlám. Prvá experimentálne zistená štruktúra tRNA[1] ukázala, že to tak nie je [104]. Následné výskumy zistili, že 60 % párov báz v štruktúrach RNA sú kánonické Watson-Crickove páry, zatiaľ čo zvyšné páry báz sú nekánonické [73]. Najčastejšími nekánonickými pármi vo veľkých RNA molekulách, ako je rRNA, sú kolísavý pár G-U[2] (obrázok 2.3) a pár G-A [42] [43].



Obr. 2.3: Kolísavé párovanie báz G-U, ktoré je geometricky odlišné od kánonického Watson-Crickovho párovania, má však podobnú termodynamickú stabilitu [28]. Obrázok bol získaný z Wikimedia Common a je verejne dostupný bez obmedzení.

Sekundárna a terciárna štruktúra RNA je formovaná a udržiavaná pomocou kánonických aj nekánonických bázových párov [51]. Vďaka mnohým typom nekánonických bázových párov môže RNA nadobúdať mnoho rôznych štruktúr, ktoré jej umožňujú rozmanité funkcie [73].

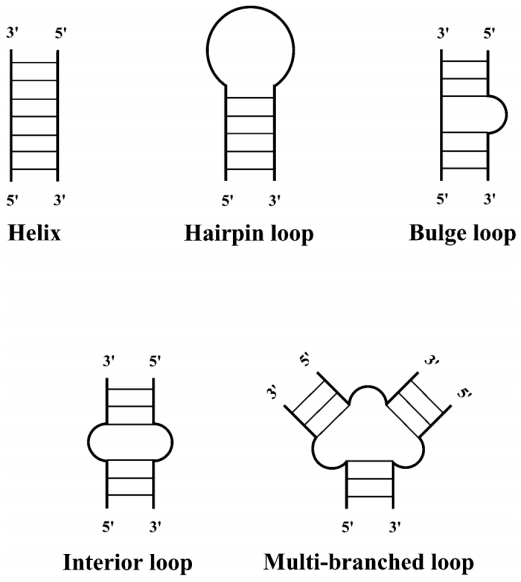
2.3 Sekundárna štruktúra RNA

Väčšina génov obsiahnutých v molekule DNA špecifikuje sekvenciu aminokyselín na tvorbu proteínu; RNA molekuly vykopírované z týchto génov sú následne preložené do proteínov. Finálny produkt niektorých génov je však samotná RNA. Tieto RNA molekuly, podobne

[1] transferová RNA
[2] anglicky: G-U wobble base pair

ako proteíny, slúžia ako enzymatické a štruktúrne komponenty pre širokú škálu procesov v bunke. Jednou z takýchto nekódujúcich RNA molekúl je práve rRNA, ktorá tvorí jadro ribozómu [16].

Funkcia molekuly je určená jej štruktúrov. Molekula RNA vytvára rôzne bázové páry, a tým umožňuje vznik rozmanitým motívom v sekundárnej štruktúre RNA, ktorá popisuje interakciu v rámci jedného vlákna nukleových kyselín [53]. Základné sekundárne štruktúry RNA sú uvedené na obrázku 2.4.



Obr. 2.4: Základné motívy sekundárnej štruktúry RNA: stonka, vlásenka, vydutá slučka, vnútorná slučka a slučka s viacerými vetvami. Obrázok je prevzatý z článku *A statistical sampling algorithm for RNA secondary structure prediction* od Dinga a Lawrenca [33].

Predikcia sekundárnej štruktúry molekuly RNA je intenzívne študovaný problém. Prvotné metódy predikcie sekundárnej štruktúry z jednej sekvencie boli založené na prístupoch ako termodynamické skladanie [83] [123] a pravdepodobnostné modelovanie [35]. Molekula RNA však môže nadobúdať veľké množstvo teoreticky možných sekundárnych štruktúr. Zuker a Sankoff [122] odhadli, že počet možných sekundárnych štruktúrnych modelov je väčší než 1.8^n, kde n je počet nukleotidov v danej sekvencii.

Odlišným prístupom je komparatívna analýza sekvencií [24], ktorá je založená na jednoduchom princípe: jedna sekundárna štruktúra RNA môže byť získaná z rôznych RNA sekvencií. Komparatívna metóda bola validovaná v roku 1976, keď bola získaná kryštálová štruktúra tRNA [64] a všetky predpovedané sekundárne interakcie sa v nej nachádzali.

2.4 Formáty reprezentácie genetickej informácie v počítači

Sekvencia nukleotidov môže byť chápaná ako reťazec nad abecedou obsahujúcou štyri symboly. Z tohto predpokladu vychádzajú formáty na reprezentáciu nukleotidových sekvencií `fasta` a `fastq`. Sekundárna štruktúra ribonukleovej kyseliny môže byť popísaná párovaním báz – na čom je založená zátvorková notácia.

Formát `fasta` je textový formát, ktorý slúži na reprezentáciu sekvencií nukleových kyselín alebo aminokyselín. V prípade kyseliny ribonukleovej sa jedná o reťazce nad abecedou $\{A, C, U, G\}$. Záznam jednej sekvencie začína riadkom obsahujúcim popis sekvencie, po ktorom nasledujú riadky so samotnou sekvenciou (obrázok 2.5). Riadok obsahujúci popis vždy začína symbolom „>“ [8].

```
>URS00004AF543 Homo sapiens rRNA
GCUAAACCUAGCCCCAAACCCACUCCACCUUACUACCAGACAACCUUAGCCAAACCAUUUACCCAAAUAAAGUAUAGGCG
AUAGAAAUUGAAACCUGGCGCAAUAGAUAUAGUACCGCAAGGGAAAGAUGAAAAAUUAUAACCAAGCAUAAUAUAGCAAG
GACUAACCCCUAUACCUUCUGCAUAAUGAAUUAACUAGAAAUAACUUUGCAAGGAGAGCCAAAGCUAAGACCCCCGAAAC
```

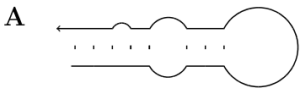
Obr. 2.5: Časť sekvencie ľudskej 16S ribozomálnej RNA získanej z databázy RNAcentral [9].

Formát `fastq` je rozšírením formátu `fasta` o hodnoty kvality jednotlivých báz v sekvencii. Hodnota kvality bázy je odvodená od odhadu pravdepodobnosti, že daná báza bola určená chybne [59]. Záznam jednej sekvencie je vo formáte `fastq` zakódovaný pomocou štyroch typov riadkov (obrázok 2.6). Prvý riadok každého záznamu začína symbolom „@“ a pokračuje názvom záznamu s jeho popisom. Nasledujúce riadky obsahujú samotnú sekvenciu. Po sekvencii nasleduje riadok, ktorý začína symbolom „+“ a voliteľne môže obsahovať zopakovaný názov záznamu. Posledné riadky obsahujú hodnoty kvality jednotlivých báz, zakódované pomocou ASCII symbolov [29].

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Obr. 2.6: Ukážka dát vo formáte `fastq` získaných z databázy SRA [72].

Zátvorková notácia[3] je textová reprezentácia sekundárnej štruktúry RNA. Spárované zátvorky symbolizujú párovanie báz a voľné bázy sú zobrazené bodkami (obrázok 2.7) [7].



Obr. 2.7: Sekundárna štruktúra molekuly RNA je zobrazená (A) spolu so sekvenciou (B) a štruktúrov v zátvorkovej notácii (C). Obrázok je prevzatý z prednášky CS Duke University [3].

2.5 Porovnávanie sekvencií nukleových kyselín

Porovnávanie dvoch sekvencií nukleových kyselín a určenie ich podobnosti je bežnou úlohou bioinformatiky. Ak je problém porovnávania zovšeobecnený na problém porovnávania reťazcov, potom je možné použiť Hammingovu vzdialenosť – počet pozícií, na ktorých sa zodpovedajúce znaky líšia. Toto zovšeobecnenie sa však nepoužíva. Hammingova vzdialenosť predpokladá, že i-ty znak jednej sekvencie je zarovnaný s i-tym znakom druhej sekvencie. Tento predpoklad však v prípade sekvencií nukleových kyselín neplatí, pretože v DNA nastávajú mutácie, čo je evolučný proces, ktorý môže viesť k substitúcii, vloženiu či vymazaniu nukleotidu. Aj keď sú reťazce ATATATAT a TATATATA veľmi odlišné z pohľadu Hammingovej vzdialenosti, ich vzdialenosť sa stane podstatne menšou, ak sa zarovnajú posunutím druhého reťazca o jednu pozíciu doprava (obrázok 2.8) [60].

[3] anglicky: dot-bracket notation

```
A T A T A T A T -
  : : : : : : :
- T A T A T A T A
```

Obr. 2.8: Zarovnanie sekvencie ATATATAT a TATATATA. Obrázok je prevzatý z knihy *An introduction to bioinformatics algorithms* [60].

Vhodnejším prístupom na porovnávanie sekvencií nukleových kyselín je editačná vzdialenosť, ktorá vyjadruje minimálny počet editačných operácií potrebných na transformovanie jednej sekvencie na sekvenciu druhú. Bežne sa používajú tri typy editačných operácií: vloženie znaku, vymazanie znaku a nahradenie znaku iným [60]. Pomocou týchto operácií sa dá popísať zarovnanie sekvencií. Zarovnanie z obrázku 2.8 vznikne vložením medzery na prvú pozíciu druhej sekvencie a vložením medzery na deviatu pozíciu prvej sekvencie.

2.5.1 Algoritmy zarovnania sekvencií

Algoritmus Needleman-Wunsch bol predstavený v roku 1970 Needlemanom a Wunschom a stal sa bežne používaným prístupom na výpočet optimálneho globálneho zarovania dvoch sekvencií [82]. Jedná sa o využitie dynamického programovania na získanie globálneho zarovnania, kde sa optimálne zarovnanie dosiahne po celej dĺžke oboch sekvencií. Zarovnanie musí siahať od začiatku do konca oboch sekvencií, aby bolo dosiahnuté najvyššie celkové skóre [119].

Algoritmus môže byť demonštrovaný na dvoch sekvenciách CGTGAATTCAT a GACTTAC. Pre zistenie najvyššieho skóre podobnosti dvoch sekvencií algorimus využíva maticu, ktorá reprezentuje všetky možné kombinácie párov, ktoré môžu byť získané zo vstupných sekvencií pri vložení medzier [4].

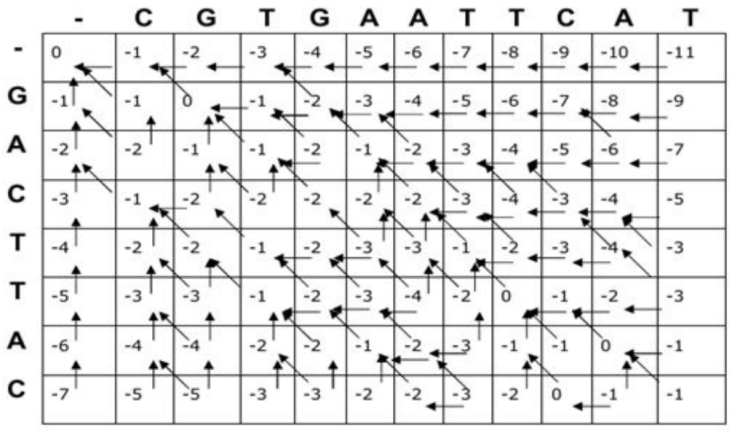
Ďalším krokom je výpočet najväčšieho možného skóre pre každú bunku matice. Začína sa z ľavého horného rohu a smeruje sa k pravému dolnému rohu, pričom sa postupuje po riadkoch. Pre každú bunku i, j sa počíta maximálne skóre podľa nasledujúceho vzorca:

$$M_{i,j} = max \begin{cases} M_{i-1,j-1} + S_{i,j} \\ M_{i,j-1} + W \\ M_{i-1,j} + W \end{cases}$$

kde $S_{i,j}$ reprezentuje penalizáciu i-teho a j-teho znaku zo zarovnávaných sekvencií a W je penalizácia za vloženie medzery. Do matice sú ešte vložené spätné ukazovatele určujúce, z ktorej bunky bolo získané maximálne skôre aktuálnej bunky [4].

Výsledné skóre zarovnania sa nachádza v pravom dolnom rohu matice. Pomocou spätných ukazateľov je hľadané optimálne zarovnanie z pravého dolného rohu do ľavého horného rohu. Získaná cesta maticou predstavuje optimálne zarovnanie – ukazuje, ktoré bázy sa majú zarovnať a na ktoré miesta majú byť vložené medzery. Týchto zarovnaní môže byť všeobecne viac a všetky môžu mať rovnaké skóre. Pre sekvencie z uvedeného príkladu existujú dve optimálne globálne zarovnania, zobrazené na obrázku 2.10 [4].

Modifikáciou algoritmu Needleman-Wunsch je algoritmus Smith-Waterman, ktorý bol predstavený Smithom a Watermanom v roku 1981 [101] a používa sa na riešenie problému lokálneho zarovnania sekvencií. Tento algoritmus je vhodné použiť pri vyhľadávaní kratšej sekvencie v dlhšej sekvencii.



Obr. 2.9: Vyplnená matica dynamického programovania pre výpočet zarovnania sekvencií CGTGAATTCAT a GACTTAC. Hodnoty v bunkách určujú najväčšie skóre zarovnania prefixov sekvencií po danú pozíciu a spätné ukazovatele určujúce, z ktorej bunky bolo získané maximálne skôre aktuálnej bunky. Obrázok je prevzatý z článku *Global alignment of two sequences – Needleman-Wunsch Algorithm* [4].

```
C G T G A A T T C A T     C G T G A A T T C A T
- - - G A C T T - A C     - G - - A C T T - A C
```

Obr. 2.10: Optimálne globálne zarovnania sekvencií CGTGAATTCAT a GACTTAC získané algorimom Needleman-Wunsch.

Algoritmus Smith-Waterman vychádza z algoritmu Needleman-Wunsch s niekoľkými modifikáciami. Bunky v prvom riadku a prvom stĺpci matice dynamického programovania sú inicializované na hodnotu 0. Bunky matice dynamického programovania neobsahujú záporné hodnoty. Ak pri výpočte skóre zarovnania vznikne záporná hodnota, do bunky sa uloží hodnota 0. Výsledné optimálne zarovnanie sa nehľadá len z pravého dolného rohu matice; maximálne hodnoty skóre sa hľadajú v celej matici a od týchto miest sa hľadá optimálne zarovnanie pomocou spätných ukazateľov.

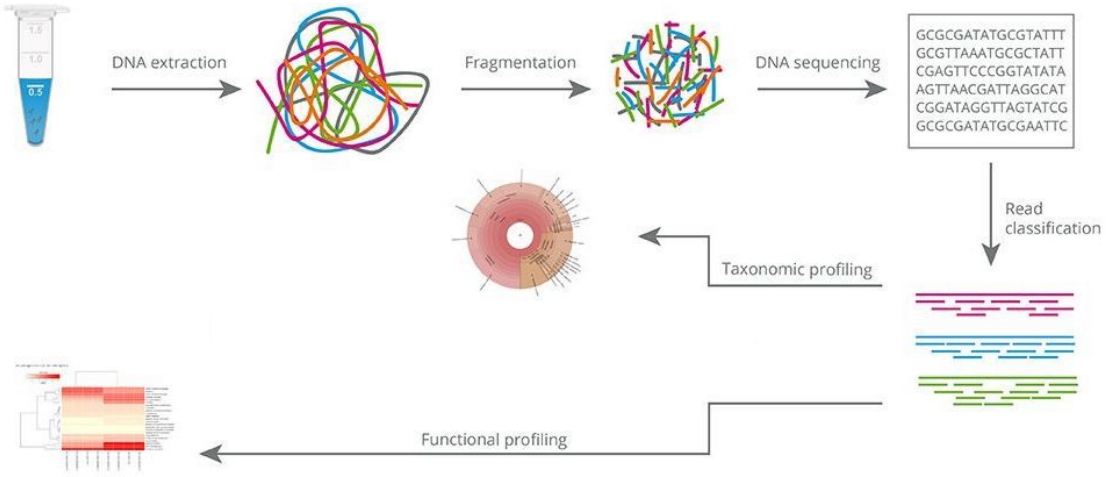
V prípade potreby zarovnávať veľké množstvo sekvencií môže byť kvadratická časová zložitosť dynamického programovania nedostačujúca. V tom prípade je možné použiť heuristický nástroj `blast`[4] [19], ktorý nájde oblasti lokálnej podobnosti medzi sekvenciami. Program porovnáva nukleotidové alebo proteínové sekvencie so sekvenčnými databázami a identifikuje tie sekvencia v databáze, ktorých podobnosť je štatisticky významná.

[4]Basic Local Alignment Search Tool

Kapitola 3

Metagenomická analýza s využitím vlastností 16S rRNA

Metagenomika je štúdium genetického materiálu získaného z organizmov, ktoré obývajú spoločené prostredie. Mikrobiológia (štúdium mikroorganizmov) bola tradične založená na kultivovaní vzoriek v laboratóriu. Väčšina mikroorganizmov však nemôže byť takto vypestovaná, a preto ich existencia nebola zistená. Metagenomika ponúka objektívny pohľad nie len na štruktúru komunity (bohatosť a distribúcia druhov), ale aj na funkčný (metabolický) potenciál komunity (obrázok 3.1) [57].



Obr. 3.1: Metagenomická analýza s využitím sekvenovania druhej generácie. Obrázok je prevzatý z článku *Metagenome Analysis* [6] a upravený.

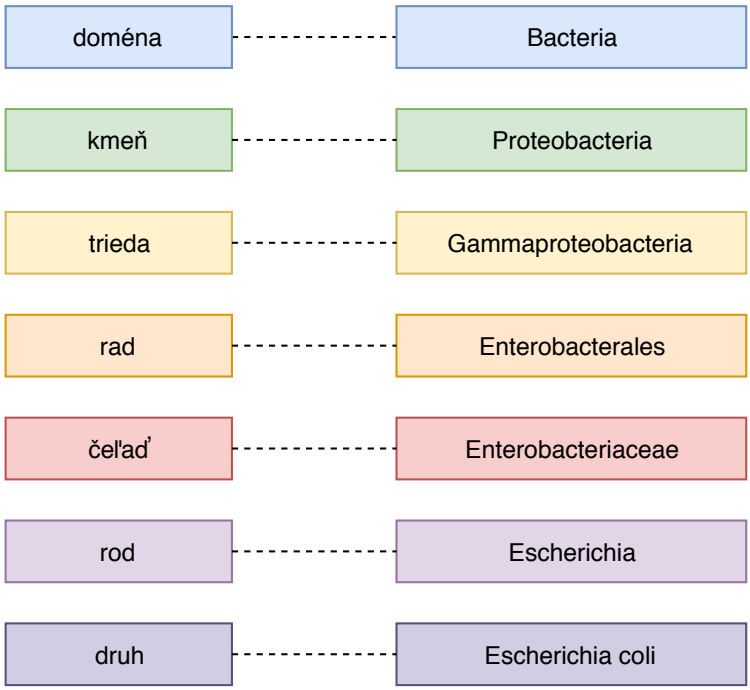
Metagenomika teda pokrýva dva hlavné typy analýzy. Prvým typom je taxonomická analýza, ktorá sa zaoberá zisťovaním, aké baktérie sú prítomné v danom vzorku. Druhým typom je funkčná analýza, ktorá skúma, čoho sú baktérie vo vzorke schopné [81]. Táto práca je zameraná na taxonomickú analýzu.

3.1 Taxonómia

Taxonómia je veda zaoberajúca sa pomenovávaním a klasifikáciou organizmov na základe ich charakteristík [116]. Taxonómia uplatnená na všetky živé organizmy poskytuje konzistentný nástroj na klasifikáciu a identifikáciu organizmov. Táto konzistentnosť umožňuje biológom na celom svete používať spoločné označenie. Spoločný jazyk, ktorý taxonómia poskytuje, minimalizuje nejasnosti týkajúce sa pomenovaní a umožňuje sústrediť pozornosť na dôležité vedecké otázky a javy [22].

Taxonomický systém bol prvýkrát zavedený v 18. storočí švédskym prírodovedcom Carolusom Linnaeusom, ktorý je taktiež vynálezcom princípu pomenovávania organizmov ich rodovým a druhovým názvom. Vyvinul hierarchický klasifikačný systém, ktorý sa s určitými zmenami používa dodnes – taxonomickú hierarchiu. Organizmy sú v tomto systéme organizované do skupín podľa ich morfologických, behaviorálnych, genetických a biochemických vlastností. Každá úroveň klasifikácie sa nazýva taxón [23].

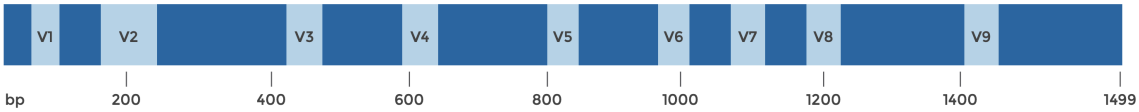
V súčasnosti taxonomická hierarchia obsahuje osem úrovní, ktorými sú (od najvšeobecnejšej po najšpecifickejšiu): doména, ríša, kmeň, trieda, rad, čeľaď, rod a druh [23]. Niektoré moderné hierarchie nepoužívajú úroveň ríša a v tejto práci s ňou taktiež nebude pracované. Ukážka zaradenia organizmu do jednotlivých taxonomických úrovní je na obrázku 3.2.



Obr. 3.2: Taxonomické kategórie Escherichia coli. Vytvorené na základe dát z NCBI Taxonomy Browser [37].

3.2 Štúdium mikrobiómu pomocou génu 16S rRNA

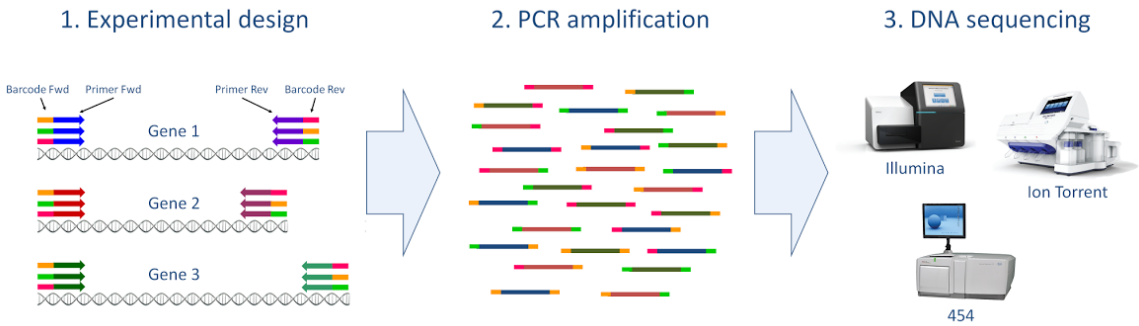
Gén 16S rRNA je sekvencia DNA, ktorá kóduje RNA komponentu malej podjednotky ribozómu. Nachádza sa v genóme všetkých bakteriálnych druhov a môže byť nájdený v každej bunke. Gén 16S rRNA obsahuje dva typy regiónov: regióny, ktoré sa počas evolúcie menili veľmi pomaly a regióny, ktoré podstúpili rapídne zmeny (obrázok 3.3) [16].



Obr. 3.3: Vizualizácia sekvencie génu 16S rRNA s konzervovanými (tmavé) a viariabilnými (svetlé) regiónmi. Obrázok je prevzaný z článku *16S rRNA Amplicon Sequencing Offers Enhanced Metagenomic Detection* od Bioo Scientific [2].

Použitie sekvencií génov 16S rRNA na štúdium baktérialnej fylogénézy[1] a taxonómie je zďaleka najbežnejším prístupom z mnohých dôvodov. Tieto dôvody zahŕňajú (i) prítomnosť v takmer všetkých baktériách, často vo forme multigénovej rodiny alebo operónov; (ii) funkcia génu 16S rRNA sa v priebehu času nezmenila, čo naznačuje, že náhodné zmeny sekvencie sú presnou mierov času (vývoja) a (iii) gén 16S rRNA (1500 bázových párov) je dostatočne veľký, aby obsahoval štatisticky relevantné informácie [86].

Použitie génu 16S rRNA vo fylogenetike bolo v roku 1973 propagované Carlom Woesom a Georgom E. Foxom [117]. Počet štúdií skúmajúcich mikrobióm explodoval od technologického pokroku sekvenačných metód druhej generácie, ktoré uľahčili analýzu nezávislú na kultivovaní a klonovaní [67]. Najbežnejším sekvenačným prístupom na analýzu mikrobiómu je amplikónová analýza génu 16S rRNA [58] [91]. V tejto metóde je región 16S rRNA amplifikovaný pomocou PCR[2] s primermi[3], ktoré rozpoznávajú vysoko konzervované oblasti génu (obrázok 3.4) [95].



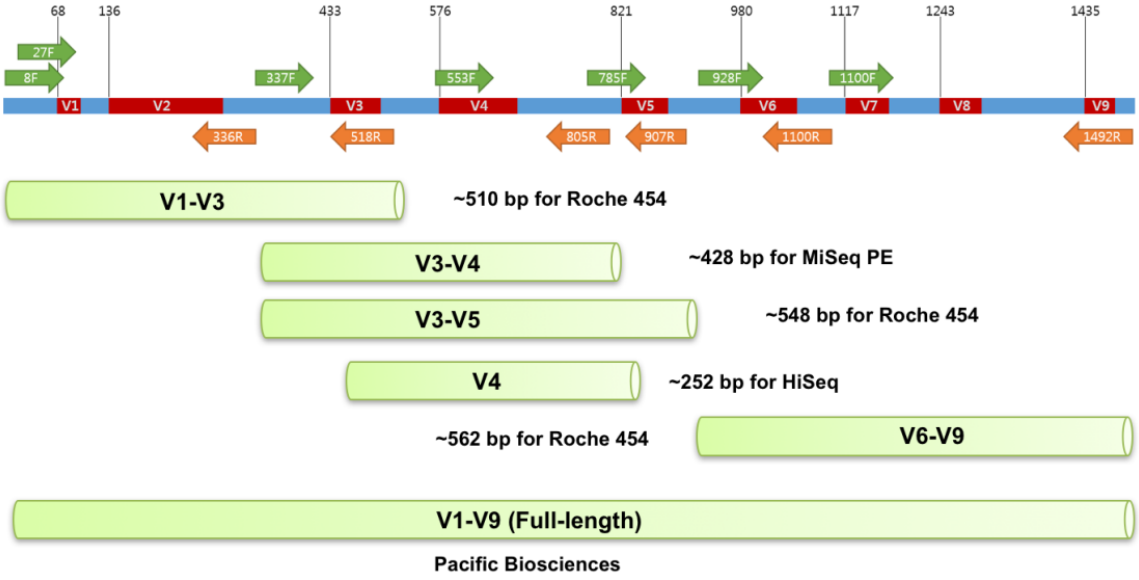
Obr. 3.4: Kroky amplikónového sekvenovania druhej generácie. 1. Experimentálny návrh sekvencií primerov na amplifikáciu požadovaných oblastí génov (markerov) a návrh značiek na identifikáciu vzorkov alebo jednotlivcov. 2. PCR amplifikácia markerov. 3. Sekvenovanie výsledkov aplifikácie. Obrázok je prevzatý z článku *Amplicon sequencing and high-throughput genotyping – Basics* [97] a upravený.

Okrem vysoko konzervovaných miest, kde sa viažu primery, sekvencia génu 16S rRNA obsahuje variabilné oblasti, ktoré môžu poskytnúť špecifické informácie na identifikáciu baktérií aj na úrovni druhov [89]. Dôsledkom je, že v lekárskej mikrobiológii prevládlo sekvenovanie génu 16S rRNA ako rýchla a lacná alternatíva k fenotypovým metódam bakteriálnej identifikácie [27].

[1] vývoj biologického druhu počas celej jeho histórie
[2] polymerázová reťazová reakcia
[3] krátka molekula, na ktorej sa začína polymerázová reťazová reakcia DNA

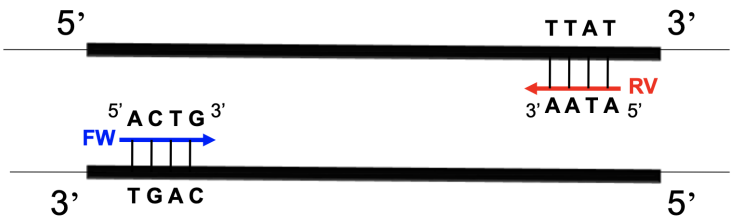
3.2.1 Cielený výber oblastí génu 16S rRNA pomocou primerov

Bakteriálny gén 16S rRNA obsahuje deväť hypervariabilných oblastí (V1 – V9) s dĺžkou od asi 30 do 100 párov báz, ktoré sú zapojené do sekundárnej štruktúry malej ribozomálnej podjednotky [48]. Aj keď sa hypervariabilné oblasti génu 16S môžu dramaticky odlišovať medzi baktériami, gén 16S ako celok si zachováva veľkú homogénnosť dĺžky, čo uľahčuje zarovnanie [87]. Gén 16S ďalej obsahuje vysoko konzervované oblasti medzi hypervariabilnými regiónmi, čo umožňuje navrhnúť univerzálne primery, ktoré môžu spoľahlivo produkovať rovnaké sekcie génu 16S v rôznych taxónoch (obrázok 3.5) [111].



Obr. 3.5: Zhrnutie populárnych sád primerov pre jednotlivé platformy sekvenovania druhej generácie. Deväť hypervariabilných regiónov identifikovaných medzi baktériami, pomenované V1 až V9, je zobrazených červeno. Dopredné primery sú zobrazené zeleno, spätné oranžovo. Obrázok je prevzatý z článku *16S rRNA and 16S rRNA Gene* [10].

Pár primerov sa skladá z dopredného[4] a spätného[5] primeru. Dopredný primer je navrhnutý tak, aby bol komplementárny k skúmanému vláknu. Spätný primer je komplementárny k opačnému vláknu (obrázok 3.6) [67].



Obr. 3.6: Ukážka návrhu dopredného (FW) a spätného (RV) primeru. Obrázok je prevzatý z prednášky *Genetics & Genotyping* [45].

[4] anglicky: forward
[5] anglicky: reverse

3.3 Analýza hypervariabilných oblastí génu 16S rRNA

Aj keď žiadna hypervariabilná oblasť nedokáže presne klasifikovať všetky baktérie od domény k druhu, niektoré môžu spoľahlivo predpovedať konkrétne taxonomické úrovne. Stupeň konzervovanosti sa značne líši medzi hypervariabilnými regiónmi, pričom viac konzervované regióny korelujú s taxonómiou vyššej úrovne a menej konzervované regióny s nižšími úrovňami, ako sú rod a druh [120].

Zatiaľ čo celá sekvencia génu 16S umožňuje porovnanie všetkých hypervariabilných oblastí, pri dĺžke 1500 párov báz to môže byť neúmerne nákladné pre štúdie, ktoré sa snažia identifikovať alebo charakterizovať rôzne bakteriálne spoločenstvá. Mnohé komunitné štúdie si vyberajú čiastočne konzervované hypervariabilné oblasti, ako je V4, pretože môžu poskytnúť rozlíšenie na úrovni kmeňa rovnako presne ako úplný gén 16S. Kombinácia regiónov V4-V6 bola stanovená ako optimálna podoblasť pre fylogenetické štúdium nových kmeňov [120].

Aj keď je analýza hypervariabilných regiónov génu 16S silným nástrojom na štúdium bakteriálnej taxonómie, má problém s rozlíšením úzko príbuzných druhov [111]. Čeľade Enterobacteriaceae, Clostridiaceae a Peptostreptococcaceae môžu mať sekvenčnú podobnosť celého génu 16S až 99 % na úrovni druhov. Výsledkom je, že sekvencie regiónu V4 sa môžu líšiť iba o niekoľko nukleotidov, takže klasifikácia pomocou referenčnej databázy nie je schopná spoľahlivo určiť tieto baktérie na nižších taxonomických úrovniach [61].

Štúdie, ktoré využívajú analýzu obmedzenú na hypervariabilné regióny génu 16S, nie sú schopné pozorovať rozdiely v úzko príbuzných taxónoch a zoskupia ich do jednej taxonomickej jednotky, čím podhodnotia celkovú rozmanitosť vzorky [111]. Navyše, bakteriálne genómy môžu obsahovať niekoľko kópií génu 16S – najmenšia zistená podobnosť týchto kópií bola 98,74 % [30]. Aj keď sa nejedná o najpresnejšiu metódu klasifikácie bakteriálnych druhov, analýza hypervariabilných oblastí zostáva jedným z najužitočnejších dostupných nástrojov pre štúdie bakteriálnych spoločenstiev [61].

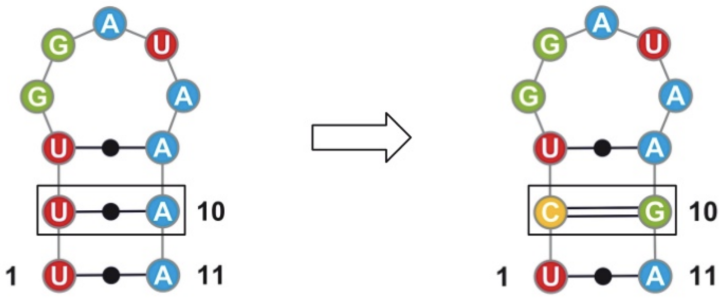
Sekvenčná diverzita medzi kmeňmi je presnejšie meraná pomocou DNA-DNA hybridizácie. Taxonómovia definujú druh ako kmene, ktoré majú podobnosť pri DNA hybridizácii aspoň 70 % [92]. Druhy s touto úrovňou podobnosti typicky majú gény 16S rRNA zhodné aspoň na 97 % [103]. Kmene s identitou génu 16S rRNA menšou ako 97.5 % pravdepodobne nebudú príbuzné na úrovni druhov. Existuje však niekoľko kmeňov, ktoré majú hybridizačnú podobnosť DNA menšiu ako 50 %, a preto sú klasifikované ako odlišné druhy, ale identitu génu 16S rRNA majú 99 % až 100 % [70] [102].

3.4 Rozdiely v rýchlosti vývoja medzi štruktúrnymi prvkami rRNA

Molekuly RNA sa skladajú do definovaných štruktúr, ktoré sú kritické pre ich biologické funkcie. Počas vývoja RNA je štruktúra oveľa viac konzervovaná než sekvencia [41] [49]. Varianty sekvencií, ktoré prispievajú k rozdielom medzi druhmi, sú tie, ktoré zachovávajú štruktúru a funkciu molekuly RNA.

rRNA je prítomná vo všetkých existujúcich druhoch a pravdepodobne sa datuje od najskorších foriem života. Odráža teda evolučnú históriu samotného života a môže byť použitá na vytvorenie evolučných vzťahov medzi všetkými druhmi na Zemi [85]. Rekonštrukcia fylogenézy závisí od predpokladaného evolučného modelu, a preto je dôležité pochopiť, ako sa rRNA skutočne vyvíja.

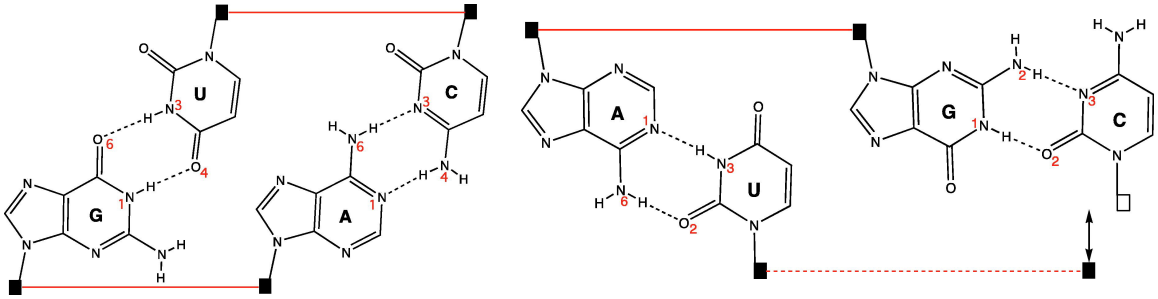
Najrozšírenejším modelom rRNA evolúcie je model „rýchlosti naprieč pozíciami[6]“, pri ktorom sa na priradenie rýchlosti evolúcie každej pozícii používa viacnásobné zarovnanie [109]. Očakáva sa, že sekundárna štruktúra ovplyvňuje rýchlosť vývoja hlavne pomocou kompenzačných mutácií v stonkách (obrázok 3.7). Predpokladá sa, že dôležitosť stoniek spočíva v ich štruktúre, a preto by substitúcia jedného páru báz za iný mala byť prijatá.



Obr. 3.7: Mutácie môžu viesť na kompenzačné mutácie kvôli zachovaniu komplementarity párovania báz. Obrázok je prevzatý z prednášky *RNA 2nd structure prediction based on multiple alignments* [105].

Na rozdieľ od toho sa predpokladá, že nespárované oblasti sú dôležité svojou sekvenciou. Tento názor bol podporený paradoxným zistením, že väčšina vysoko konzervovaných oblastí (oblastí so žiadnou alebo len malou variabilitou na úrovni sekvencie) v rRNA bakteriálnej malej podjednotke bola skôr nespárovaná, než spárovaná [118] [94] [84].

Predpoklad, že vývoj RNA je prevažne založený na kompenzačných mutáciách v párových oblastiach, naznačuje, že na popis párových oblastí vo vývojových štúdiách by mali byť použité matice špecifických rýchlostí. RNA porušuje predpoklad nezávislosti pozície, ktorý je základom mnohých vývojových modelov, pretože udržiavanie párovania báz vyžaduje, aby sa bázy na dvoch interagujúcich miestach menili korelačným spôsobom. V súčasnosti mnoho modelov vývoja RNA pracuje so závislosťou pozícií v párovaných oblastiach tým, že umožňuje korelované mutácie [94] [106] [107], vrátane nekánonických interakcií párov báz, reprezentovaných pomocou izosterických matíc [73]. Páry báz, ktoré majú podobné geometrické parametre, sú označované ako izosterické (obrázok 3.8).



Obr. 3.8: Nekánonické párovanie G-U a A-C je izosterické v trans konfigurácií. Watson-Crickovo párovanie G-C a A-U nie je izosterické v trans konfigurácii. Obrázok je prevzatý z článku *Isostericity and tautomerism of base pairs in nucleic acids* od Westhofa [114].

[6]anglicky: rates across sites

Aj keď je štandardný model rýchlo sa vyvýjajúcich stoniek všeobecne akceptovaný [107] [115] [56], existujú dva dobré dôvody domnievať sa, že rozdelenie RNA oblastí do dvoch skupín – párované a nepárované – poskytuje obmedzený pohľad na vývoj RNA.

Po prvé, aj keď mnoho bázových párov v mnohých molekulách môže byť experimentálne zmenených bez narušenia funkcie, to isté platí pre nepárové oblasti. Napríklad, nahradenie veľkých, alebo zle štrukturovaných, slučiek trojitými slučkami[7] sa bežne používa na zlepšenie kryštalizácie RNA [46]. Preto nie je jasné, či sú v priemere mutácie v stonkách tolerované častejšie ako mutácie v nepárových oblastiach.

Po druhé, pozorovania toho, že mnoho vysoko konzervovaných báz v rRNA je nepárových [118], nemusí znamenať, že väčšina nepárových báz v rRNA je vysoko konzerovaných. Napríklad, mapy konzervovanosti z webovej stránky komparatívnej RNA (CRW) [24] ukazujú, že 44-35 % neukleotidových pozícií v baktériách a eukaryotách (v oboch veľkých podjednotkách a malej podjednotke ribozómu) je konzervovaných vo viac ako 98 % sekvenciách v zarovnaní. Z týchto viac ako 98 % konzervovaných pozícií je iba 50-54 % nespárovaných. Pretože v rRNA existuje viac spárovaných pozícií ako nespárovaných, v priemere asi 50 % nespárovaných pozícií a 30 % spárovaných pozícií je vysoko konzerovaných. Druhá polovica nespárovaných pozícií sa teda môže voľne meniť [99].

3.5 Fylogenetická analýza s ohľadom na sekundárnu štruktúru 16S rRNA

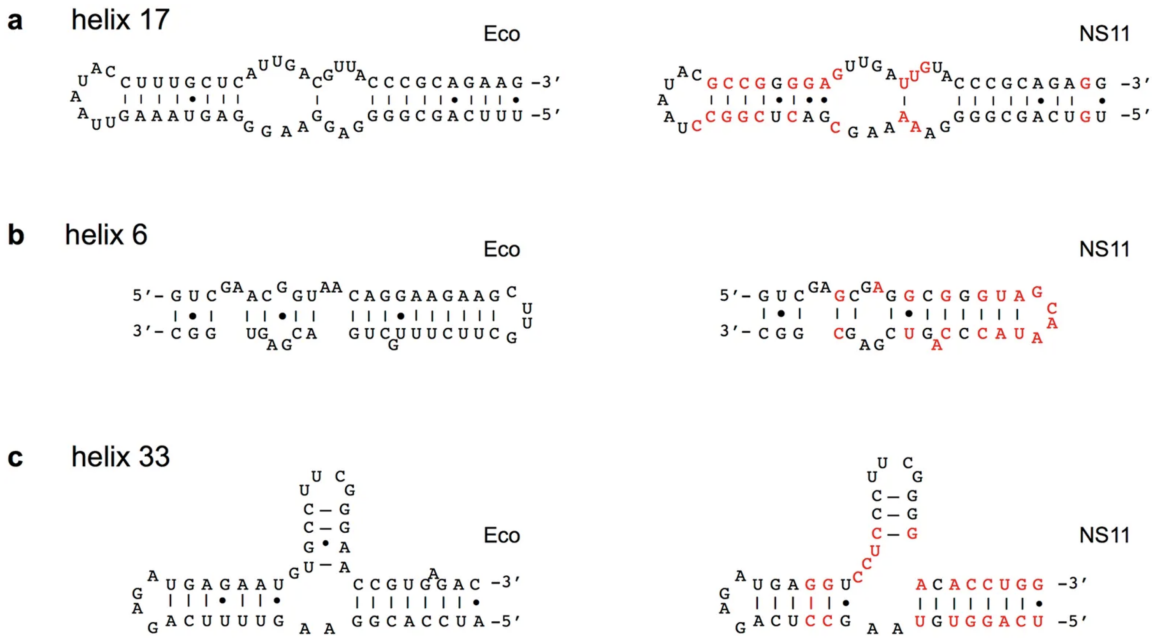
Jednoduché zarovnanie izolovanej sekvencie 16S rRNA je menej informatívne ako analýza tejto sekvencie obmedzenej modelmi sekundárnej štruktúry [100]. Vývoj sekundárnej štruktúry molekuly 16S rRNA je obmedzený prirodzeným výberom [50], keďže funkcionalita je priamo spojená s rôznymi hierarchickými úrovňami molekulárnej štruktúry [40] [49]. Napríklad, variácie v sekvencii sú pozorované v stonkách, ale sekundárna štruktúra nie je zvyčajne narušená touto variáciou, a to vďaka kompenzačným mutáciám, ktoré zachovávajú párovanie báz v stonke [79].

Sekundárna štruktúra 16S rRNA už bola použitá v rôznych fylogenetických štúdiach. Li a kolektív [75] použili informácie o sekundárnej štruktúre na vylepšenie modelu tvorby fylogenetického stromu pri skúmaní kaprovitých rýb. Podobný prístup použili Smith a Bond [100] pri skúmaní fylogenézy pavúkov. Obe štúdie využili sekundárnu štruktúru na priradenie váh jednotlivým oblastiam sekvencie, ktoré následne použili na vytvorenie presnejšieho fylogenetického stromu.

3.6 Komparatívna analýza funkcie 16S rRNA

Escherichia coli a Acidobacteria, ktoré sú fylogeneticky odlišné na úrovni kmeňu, majú sekvencie 16S rRNA zhodné na 78 % (líšia sa v 334 nukleotidoch z približne 1500). Bolo zistené, že z týchto 334 odlišných pozícií bol iba jeden pár baz škodlivý a zvyšných 332 (99,4 %) nukleoditov malo podobnú funkcionalitu (obrázok 3.9) [108].

[7] anglicky: tetraloop



Obr. 3.9: Porovnanie sekundárnych štruktúr Escherichia coli a acidobakteriálnej 16S rRNA: (a) stonka 17, (b) stonka 6 a (c) stonka 33. NS11 16S rRNA, ktoré bolo funkčné v Escherichia coli, má sekvenciu zhodnú na 78.4 % s Eco 16S rRNA. Nukleotidy, ktoré sa líšia medzi sekvenciami Eco a NS11, sú v štruktúre NS11 znázornené červenou farbou. Obrázok je prevzatý z článku *Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16S rRNAs* od Tsukuda a kolektívu [108].

3.7 Kópie génu 16S rRNA v bakteriálnom genóme

Napriek širokému použitiu 16S rRNA na taxonomickú klasifikáciu baktérií existuje niekoľko aspektov, ktoré obmedzujú interpretáciu získaných výsledkov. Najdôležitejšia je skutočnosť, že počet kópií génu 16S rRNA na genóm sa líši od 1 do 15, alebo aj viac kópií [66]. Zdá sa, že počet kópií je do určitej miery špecifický pre taxóny, zaznamenali sa však aj rozdiely medzi kmeňmi toho istého druhu [14]. Počet kópií génu 16S rRNA bol daný do súvislosti s bakteriálnou schopnosťou prežiť, pretože počet kópií niektorých taxónov koreluje s ich schopnosťou reagovať na priaznivé podmienky pre rast. Predpokladá sa, že taxóny s nízkym počtom kópií sú viac oligotrofné [36] [65].

Ďalej sa predpokladá, že kópie génov rRNA v organizme sú homogenizované prostredníctvom génovej konverzie [54]. Avšak sekvencie 16S z toho istého druhu, alebo dokonca z rovnakého genómu sa často líšia. V dôsledku toho je odhadované, že množstvo variánt 16S rRNA je 2,5-krát väčšie ako počet bakteriálnych druhov [14] a u niektorých bakteriálnych taxónov [112] [121] sa pozorovali vysoko odlišné 16S rRNA sekvencie. Bakteriálne druhy so sekvenciami, ktoré sa líšia o viac ako 1%, sú celkom bežné [88]. Ešte väčšia variabilita 16S rRNA sekvencií bola zistená u termofilných baktérií, kde sa ako potenciálna príčina navrhol vyšší výskyt horizontálneho prenosu génov [14].

Kapitola 4

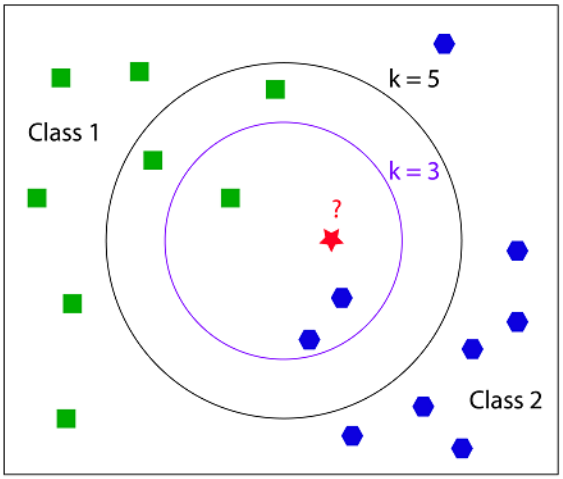
Klasifikácia mikrobiómu

Klasifikácia je proces priradenia objektu do učitej triedy na základe jeho vlastností. Základom je pochopenie štruktúry dát z dátovej sady, ktorá už je rozdelená do skupín – tried. Toto odvodenie je typicky dosiahnuté pomocou štatistického modelu, ktorý je následne použitý na odhad tried dát, ktoré zatiaľ neboli videné [15].

Klasifikačné algoritmoy teda typicky obsahujú dve fázy. Prvá fáza je trénovacia, v ktorej je z trénovacích dát vytvorený model. V druhej – testovacej – fáze je vytvorený model použitý na určenie tried testovacích dát, ktoré klasifikačný algoritmus zatiaľ nevidel [15].

Algoritmom, ktorého príncíp je základom väčšiny metód klasifikácie baktérií, je algoritmus k-najbližších susedov (skrátene k-NN[1]), ktorý je jeden z najstarších a najjednoduchších algoritmov na riešenie klasifikačných a regresných problémov, navrhnutý v roku 1951 [38].

Je to neparametrický algortimus, čo znamená, že obsahuje fixný počet parametrov, bez ohľadu na veľkosť dát [62]. Patrí medzi strojové učenie s učiteľom a trieda neznámeho vzorku je určená na základe jeho k najbližších susedov – vzorky s najmešou vzdialenosťou od neznámeho vzorku – a objekt je priradený do triedy, ktorá je najviac zastúpená medzi jeho susedmi (obrázok 4.1) [52].

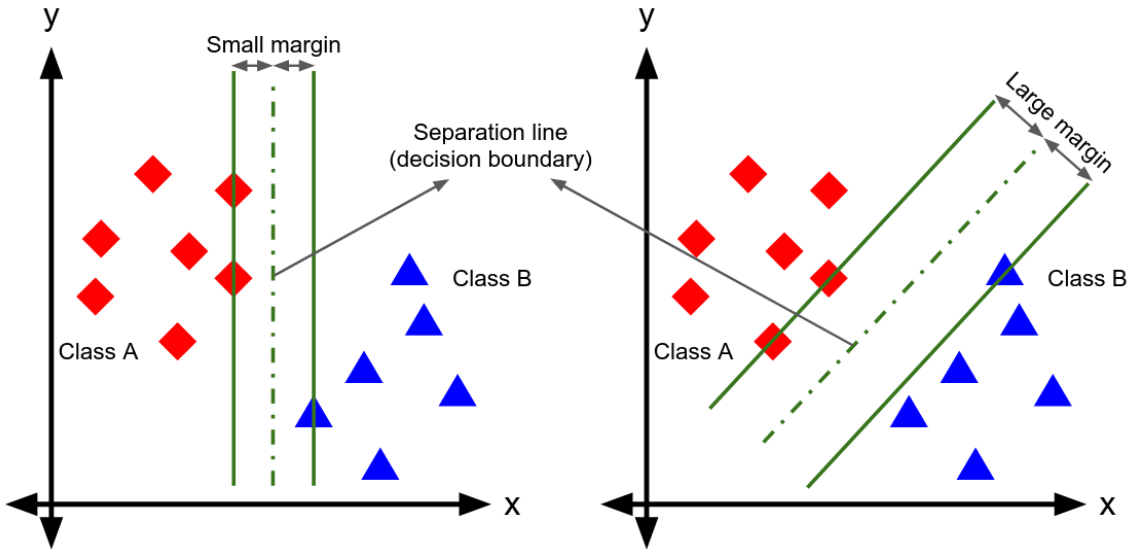


Obr. 4.1: Klasifikácia objektu „červená hviezda“ do jednej z dvoch tried. Pri $k = 3$ bude objekt klasifikovaný do triedy 2, pri $k = 5$ bude klasifikovaný do triedy 1. Obrázok je prevzatý z článku *k-Nearest Neighbors* [11].

Algoritmus k-nejbližších susedov jednotlivo analyzuje vstupné bakteriálne sekvencie. Ku každej vstupnej sekvencii určí sekvencie s najväčšou podobnosťou a z nich vytvorí výsledné taxonomické zaradenie skúmanej bakteriálnej sekvencie.

[1] anglicky: k-Nearest Neighbors

Ďalší algoritmus používaný na klasifikáciu mikrobiómu je metóda podporných vektorov (skrátene SVM[2]). SVM je metóda strojového učenia s učiteľom, v ktorej je model zostavený na základe trénovacej dátovej sady. Metóda je založená na princípe rozhodovacích hyperrovín, kde každá hyperrovina umožňuje oddeliť dve triedy podľa rozloženia ich znakov [31]. SVM sa snaží nájsť hyperrovinu, ktorá najlepšie rozlišuje dve triedy objektov, to znamená nájsť hyperrovinu, ktorá vytvára najväčší okraj medzi týmito dvoma triedami. Princíp je znázornený na obrázku 4.2.



Obr. 4.2: Vizualizácia hľadania hyperroviny, ktorá vytvára najväčší okraj medzi dvoma triedami. Obrázok je prevzatý z článku *Support Vector Machine: Classification* [13].

Algoritmy k-nejbližších susedov, metóda podporných vektorov a im podobné klasifikačné algoritmy sú vhodné na problémy, kde je možné vstupné objekty klasifikovať jednotlivo. V prípade, kedy je vstupné objekty potrebné analyzovať ako celok, je potrebné použiť iný prístup. Ak je možné definovať priestor potenciálnych riešení, potom môže byť použitý niektorý z algoritmov prehľadávania stavového priestoru.

4.1 Algoritmus Metropolis Hastings

Metropolis Hastings [55] nepatrí medzi klasifikačné algoritmy, ale jeho schopnosti je možné na klasifikáciu využiť. Patrí do triedy MCMC[3] vzorkovacích algoritmov a najbežnejšie sa používa na optimalizáciu vzorkovania z posteriórneho rozloženia.

Monte Carlo [80] (alebo simulácia Monte Carlo) je trieda algoritmov, ktorá využíva náhodnosť a vzorkovanie na riešenie matematických problémov. Obzvlášť sa používa, keď je funkcia zložitá na analytický zápis (neexistuje žiadna uzavretá forma). Tieto metódy sa používajú na riešenie problémov, ako je integrácia, optimalizácia alebo generovanie vzorkov z pravdepodobnostného rozdelenia.

Cieľom Monte Carlo simulácie je získať sadu vzoriek $\{x^{(i)}\}_{i=1}^{N}$ z cieľového rozloženia $p(x)$ definovaného na vysokorozmernom priestore. Týchto N vzoriek je možné použiť na aproximáciu cieľového rozloženia s využitím nasledujúcej funkcie:

[2]anglicky: Support Vector Machines
[3]z anglického Markov Chain Monte Carlo

$$p_N(x) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x^{(i)}}(x)$$

kde $\delta_{x^{(i)}}(x)$ označuje hodnotu detla-Dirac v bode $x^{(i)}$ [20].

Markovova vlastnosť [78] hovorí, že v postupnosti experimentov poznanie súčasnosti robí minulosť a budúcnosť nezávislou. Experimenty sú bez pamäti. Presnejšie, pravdepodobnosť nového stavu závisí len na aktuálnom stave:

$$P\left(X_{t+1}|X_1, X_2 \ldots X_t\right) = P\left(X_{t+1}|X_t\right)$$

Algoritmus Metropolis Hastings teda vzorkuje prehľadávaný priestor s využitím Markovovej vlastnosti. Nový navrhovaný stav je generovaný len na základe aktuálneho stavu a pravdepodobnosť prijatia navrhovaného stavu závisí na ohodnotení daného stavu. Pre stavový priestor q je postup zobrazený v algoritme 1 [63].

Algoritmus 1: Algoritmus Metropolis Hastings

Inicializuj: $x^{(0)} \sim q(x)$;
for *iterácia i = 1,2,...* **do**
  Navrhni: $x^{\text{cand}} \sim q\left(x^{(i)}|x^{(i-1)}\right)$;
  Pravdepodobnosť prijatia: $\alpha\left(x^{cand}|x^{(i-1)}\right) = \min\left\{1, \frac{q\left(x^{(i-1)}\right)x^{\text{cand}}\right)\pi\left(x^{\text{cand}}\right)}{q\left(x^{\text{can}\,d}|x^{(i-1)}\right)\pi\left(x^{(i-1)}\right)}\right\}$;
  $u \sim$ Uniform $(u; 0, 1)$;
  if $u < \alpha$ **then**
    Príjmi návrh: $x^{(i)} \leftarrow x^{\text{cand}}$;
  else
    Odmietni návrh: $x^{(i)} \leftarrow x^{(i-1)}$;
  end
end

Kapitola 5

Návrh klasifikácie baktérií na základe variánt 16S rRNA

Hlavnou úlohou tejto práce je navrhnúť nástroj, ktorý bude schopný klasifikovať sekvencie génu 16S do taxonomických kategórií na základe vlastností 16S rRNA. Ako jedna z vhodných vlastností génu 16S rRNA, ktorá môže byť použitá pri klasifikácii baktérií, je výskyt viacerých kópií tohto génu. Ako bolo uvedené v kapitole 3.7, baktérie obsahujú niekoľko kópií génu 16S rRNA a počty a sekvencie týchto kópií sa medzi baktériami líšia. Tento poznatok je použitý na vytvorenie systému, ktorý sa na neznáme sekvencie pozerá ako na celok a určuje, aké baktérie a v akom pomere sa nachádzajú na vstupe.

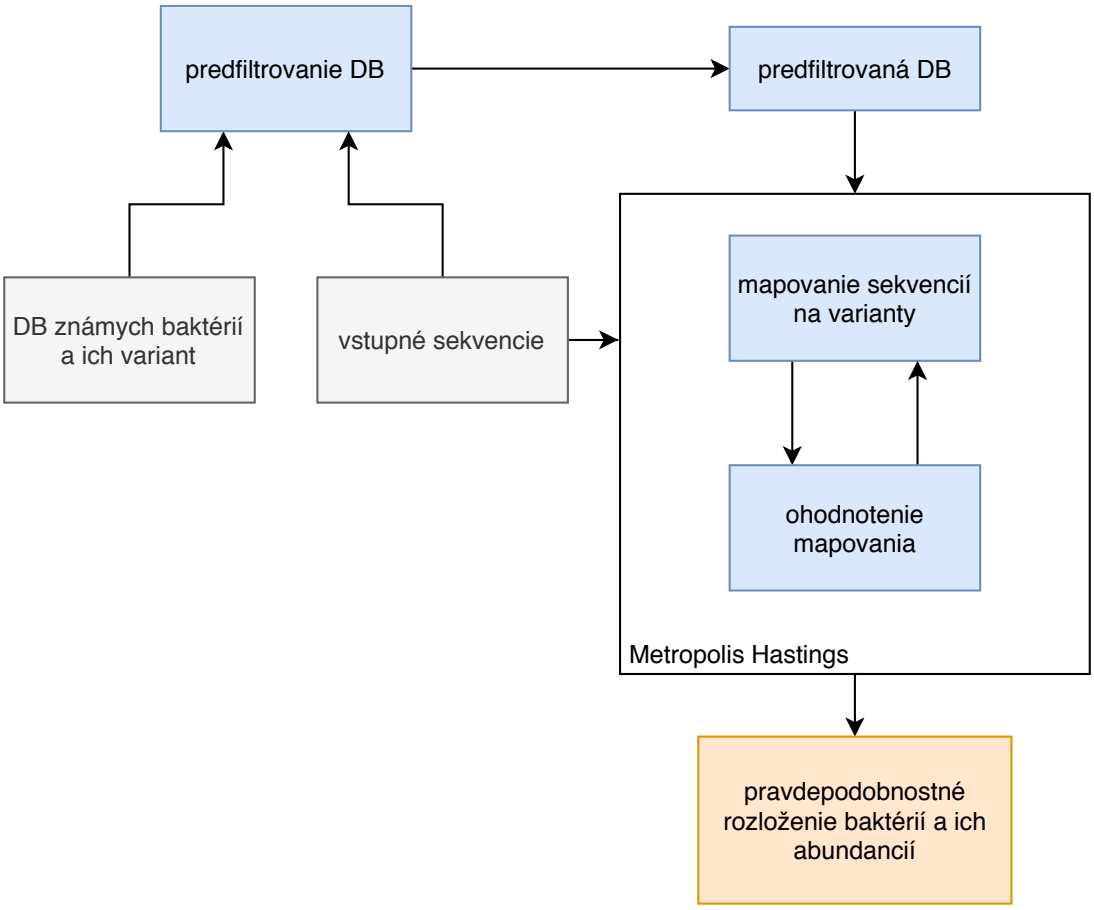
Základom tohto prístupu je hľadanie optimálneho mapovania medzi vstupnými sekvenciami a variantami známych baktérií. Navrhovaný prístup určuje nie len baktériu, ktorej vstupná sekvencia odpovedá, ale aj konkrétnu variantu 16S rRNA danej baktérie. Toto mapovanie nemusí byť jednoznačné a ukážky rôznych mapovaní sú uvedené v kapitole 5.2 spolu s dôvodmi, prečo je potrebné využiť prehľadávanie stavového priestoru.

Použitý algoritmus Metropolis Hastings dovoľuje prehľadávanie priestoru možných mapovaní s tým, že frekvencia stavu na výstupe odpovedá hodnote, ako pravdepodobný je daný stav z pohľadu zvoleného štatistického modelu. Prehľadávanie stavového priestoru je popísané v kapitole 5.3. Pri prehľadávaní je možné použiť rôzne prístupy na ohodnotenie stavu systému. Navrhované prístupy sú popísané v kapitole 5.4. Výstupom systému teda nie je presná identifikácia baktérií, ale pravdepodobnostné rozloženie nad možnými baktériami a ich abundanciami. Schéma návrhu tohto systému je na obrázku 5.1.

Kvôli komplexnosti prístupu prehľadávania priestoru možných mapovaní je najskôr potrebné predfiltrovať databázu známych baktérií a ďalej pracovať len s tými baktériami, ktoré majú potenciál sa vyskytovať vo vstupných dátach, čím sa zníži výpočetná náročnosť následného prehľadávania stavového priestoru. Tento krok je popísaný v kapitole 5.1.

5.1 Predfiltrovanie databázy známych baktérií

Pri skúmaní jednotlivých vstupných sekvencí nevieme určiť, z akej baktéria bola daná sekvencia získaná. Môžeme však vytvoriť zoznam potenciálnych baktérií, ktoré budú následne podrobené dôkladnejšiemu skúmaniu. Získanie kandidátnych baktérií prebieha pomocou porovnávania vstupných sekvencií s variantami baktérií v databáze. Ak má baktéria variantu, ktorá je dostatočne podobná s niektorou vstupnou sekvenciou, potom je daná baktéria zaradená medzi kandidátne baktérie. Pri baktériach, ktorých varianty sú príliž vzdialené od



Obr. 5.1: Návrh architektúry nástroja na klasifikáciu baktérií na základe variánt 16S rRNA. Vstupy nástroja sú zobrazené sivou farbou, hlavné kroky zú zobrazené modrou farbou a finálny výstup oranžovou farbou.

vstupných sekvencií, je vysoká pravdepodobnosť, že sa nenachádzajú vo vstupných dátach a preto nie sú predmetom ďalšieho skúmania. Vďaka tomu sa výrazne zníži počet analyzovaných baktérií, čo bude mať za následok zníženie výpočetnej náročnosti následného prehľadávania stavového priestoru.

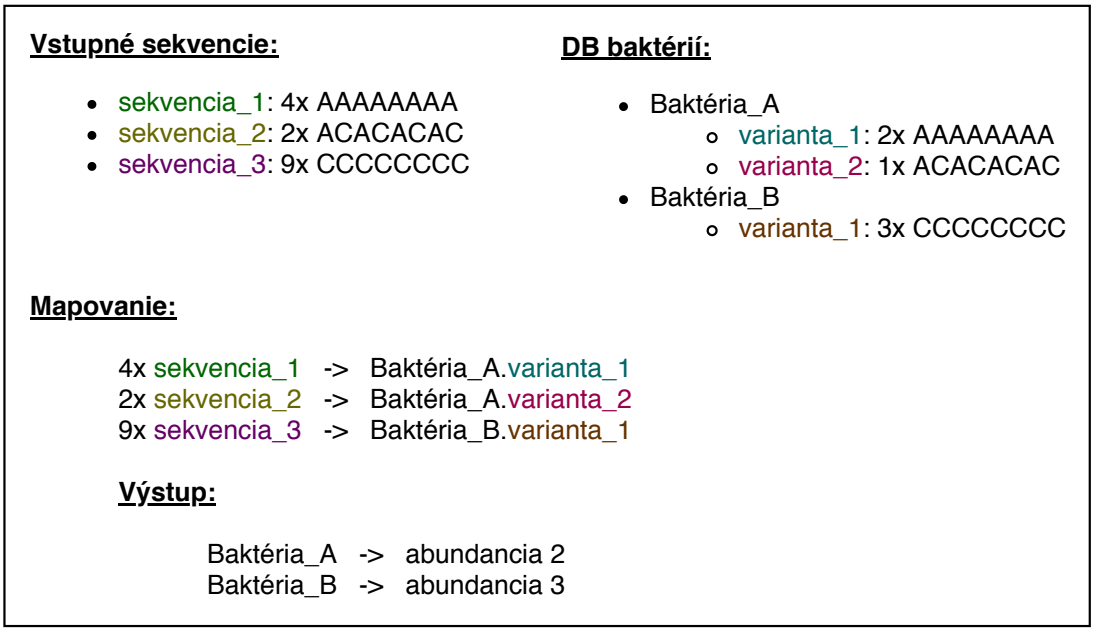
Na meranie podobnosti sekvencií môže byť použitých niekoľko prístupov. Ak majú sekvencie rovnakú dĺžku, potom je na predfiltrovanie možné použiť k-merové spektrum. K-mer označuje podsekvenciu dĺžky k nájdenú vo vstupnej sekvencii. K-merové spektrum potom označuje abundancie jednotlivých k-merov v sekvencii [71]. Po transformácii do k-meroveho spektra sú všetky sekvencie reprezentované vektormi rovnakej dĺžky a na ich porovnanie je možné použiť niektorú zo vzdialenostných metrík. Parametrami metódy je hodnota k, vzdialenostná metrika a prah určujúci, ako veľmi vzdialené sekvencie ešte majú byť predmetom skúmania.

Ďalším možným prístupom je použitie nástroja `blast` [19]. Tento nástroj vyhľadáva zhody na základe lokálneho zarovnania sekvencí a môže byť použitý aj pri klasifikácií sekvencií, ktoré nepokrývajú celý gén 16S rRNA. Vstupom nástroja sú všetky vstupné sekvencie a všetky sekvencie variánt baktérií. Výstupom sú všetky nájdené lokálne zhody. Tieto zhody je potrebné následne vyfiltrovať a určiť tie baktérie, s ktorými mali vstupné sekvencie dostatočnú zhodu. Parametrom tohto prístupu predfiltrovania je maximálny počet nezhôd

medzi vstupnou sekvenciou a sekvenciou varianty baktérie. K-merové spektrum a nástroj `blast` určujú podobnosť na základe sekvenčnej podobnosti 16S rRNA.

5.2 Mapovanie vstupných sekvencií na varianty baktérií

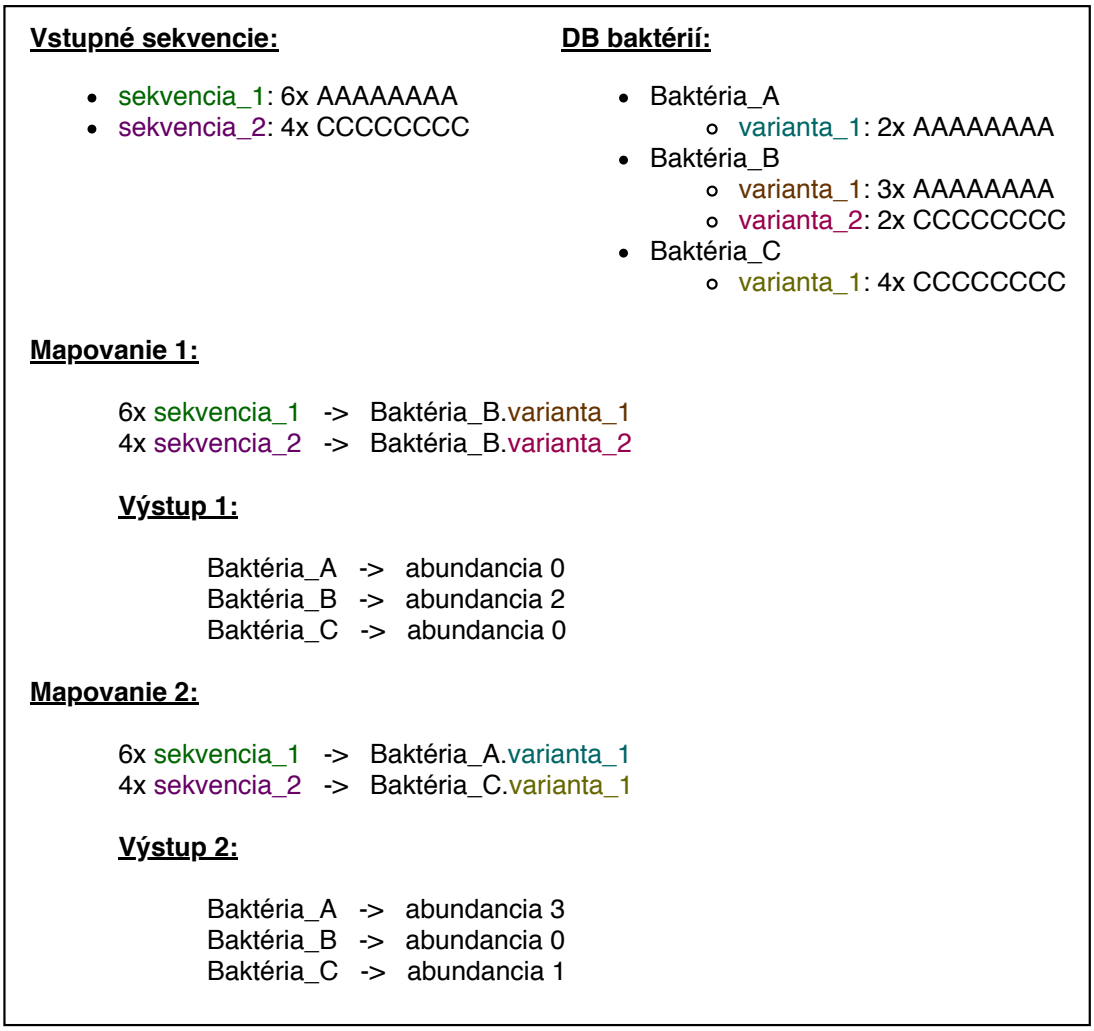
Po získaní množiny kandidátnych baktérií môžeme pristúpiť k mapovaniu vstupných sekvencií na varianty baktérií a určovaniu výsledných baktérií a ich abundancií. Základný princíp mapovania je znázornený na obrázku 5.2, ktorý predstavuje ideálny prípad, kde sekvencie na vstupe presne zodpovedajú variantám baktérií v databáze. Po namapovaní sekvencií na správne varianty je potom výstupom systému, že na vstupe bola baktéria A s abundanciou 2 a baktéria B s abundanciou 3.



Obr. 5.2: Ukážka mapovania vstupných sekvencií na varianty baktérií v ideálnom prípade, kedy počty aj sekvencie presne zodpovedajú. Výsledné abundancie baktérií sú vypočítané pomocou podielu priradených sekvencií k variante a abundancie danej varianty. Detajlnejší popis sa nachádza v kapitole 5.3.

Pri reálnych dátach však nemôžeme očakávať takéto jednoznačné mapovanie medzi sekvenciami a variantami. Samotné sekvenovanie zanáša do vstupných dát rôzne chyby. Sekvenačné techniky stále nedosahujú 100 % presnosť, a teda niektoré bázy vstupných sekvencií môžu byť pozmenené. Ďalej sa môže stať, že pomery abundancií jednotlivých sekvencií nebudú pri sekvenovaní zachované, alebo že nejakú variantu 16S rRNA sa nepodarí osekvenovať vôbec.

Nejednoznačnosť mapovania nastáva aj v prípade zdieľania rovnakých variánt medzi viacerými baktériami. Jednoduchá ukážka tejto nejednoznačnosti je na obrázku 5.3, kde už na prvý pohľad môžu nastať dve rôzne mapovania, ktoré vedú k odlišnej výslednej identifikácii baktérií a ich abundancií.

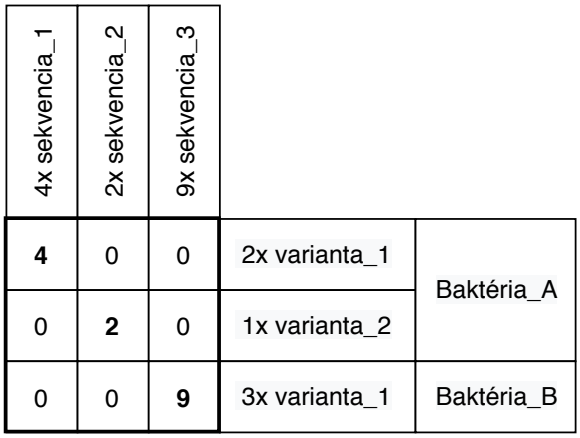


Obr. 5.3: Ukážka mapovania vstupných sekvencií na varianty baktérií v prípade, kedy mapovanie nie je jednoznačné. Na prvý pohľad existujú dva úplne odlišné mapovania, pri ktorých sekvencie aj pomery ich abundancií sedia, ale výsledné abundancie baktérií sú odlišné.

5.3 Prehľadávanie stavového priestoru

Cieľom navrhovanej klasifikácie je nájsť optimálne mapovanie medzi vstupnými sekvenciami a variantami baktérií, ktoré v ideálnom prípade odpovedá skutočnosti. Po získaní tohto mapovania je možné určiť taxonomické zaradenia baktérií vo vsupnej vzorke a pomery ich abundancií. Ako však bolo ukázané, nájdenie optimálneho mapovania je z mnohých dôvodov problematické a preto je potrebné použiť prístup prehľadávania stavového priestoru. Možnosťou je použiť algoritmus Metropolis Hastings, ktorý bol popísaný v kapitole 4.1, a ktorý je vhodný pre vzorkovanie z pravdepodobnostného rozloženia s mnohými dimenziami.

V tomto prípade si jeden stav prehľadávaného priestor môžeme predstaviť ako dvojrozmenú maticu, kde riadky predstavujú varianty známych baktérií a stĺpce predstavujú vstupné sekvencie. Čísla v bunkách matice určujú, koľko sekvencií z daného stĺpca je priradených variante v danom riadku. Prevedenie mapovania z obrázku 5.2 do maticového tvaru je zobrazené na obrázku 5.4.



Obr. 5.4: Maticové zobrazenie mapovania medzi vstupnými sekvenciami a variantami baktérií. Stĺpce zodpovedajú vstupným sekvenciám, riadky variantám baktérií a čísla v bunkách matice určujú, koľko sekvencií z daného stĺca je priradených variante v danom riadku.

Na použitie algoritmu Metropolis Hasting, potrebujeme byť schopní vygenerovať náhodný stav systému, vygenerovať nový stav systému na základe aktuálneho stavu a ohodnotiť stav systému. Pseudokód navrhovaného generovania náhodného stavu systému je uvedený v algoritme 2.

Algoritmus 2: Generovanie náhodného stavu systému

```
for každú unikátnu sekvenciu na vstupe do
    while celá abundancia aktuálnej sekvencie nie je priradená do
        vygeneruj náhodné celé číslo t z intervalu <0, nepriradená abundancia>;
        vyber náhodnú variantu z kandidátnych baktérií;
        priraď t vstupných sekvencí vybratej variante;
        odčítaj hodnotu t od nepriradenej abundancie;
    end
end
```

Pseudokód navrhovaného generovania nového stavu na základe aktuálneho stavu je uvedený v algoritme 3. Vstupom algoritmu je aktuálny stav systému, teda aktuálne mapovanie medzi vstupnými sekvenciami a variantami baktérií. Kľúčové pre tento algoritmus je, aby nový stav systému bol v stavovom priestore blízko aktuálneho stavu. Tým je zabezpečené, že stavový priestor bude vzorkovaný systematicky.

Algoritmus 3: Generovanie nového stavu systému na základe aktuálneho stavu

```
for každú unikátnu sekvenciu na vstupe do
    for požadovaný počet zmien do
        náhodne vyber variantu, ktorej je priradený nenulový počet kópií aktuálnej
         sekvencie;
        vygeneruj náhodné celé číslo t z intervalu <1, priradená abundancia>;
        vyber náhodnú variantu z kandidátnych baktérií;
        presuň t sekvencí od aktuálnej varianty k náhodne zvolenej variante;
    end
end
```

Po vygenerovaní nového stavu systému je potrebné vyhodnotiť, ako pravdepodobný je daný stav z pohľadu zvoleného štatistického modelu. Ohodnotenie stavu môže byť vykonané z dvoch pohľadov: podobnosť vzájomne namapovaných sekvencií a dodržanie pomerov abundancií variánt pri jednotlivých baktériách. Konkrétne prístupy hodnotenia stavu systému sú popísané v kapitole 5.4 a v navrhovanom nástroji môžu byť na ohodnotenie použité ľubovolné kombinácie týchto prístupov.

Ohodnotenia stavu získané pomocou rôznych pístupov môžu nadobúdať hodnoty z rôznych intervalov. Preto je vhodné použiť normalizáciu skóre podľa vzťahu:

$$score' = a + \frac{(score - score_{\min})\,(b - a)}{score_{\max} - score_{\min}}$$

kde $< score_{\min}, score_{\max} >$ je pôvodný interval hodnôt skóre a $< a, b >$ je požadovaný interval hodnôt skóre, v tomto prípade $< 0, 1 >$. Pôvodný interval hodnôt je potrebné získať pred spustením samotnej optimalizácie algoritmom Metropolis Hastings. Pre každý prístup ohodnotenia je zvlášť spustené prehľadávanie stavového pristoru, pričom ohodnocovanie stavu je vykonávané len pomocou daného prístupu. Počas prehľadávania sa zaznamená najnižšia a najvyšia hodnota skóre, ktoré sú následne uložené ako $< score_{\min}, score_{\max} >$.

Po ohodnotení navrhovaného stavu je potrebné určiť, či má byť prijatý ako aktuálny stav sytému. Ak je ohodnotenie navrhovaného stavu lepšie ako ohodnotenie aktuálneho stavu, navrhovaný stav je prijatý. V opačnom prípade je navrhovaný stav prijatý s určitou pravdepodobnosťou, ktorá sa odvíja od toho, ako veľmi je navrhovaný stav horší od aktuálneho stavu. Navrhovaný stav je potom prijatý, ak platí nasledujúci vzťah:

$$exp\left(\frac{proposed\_score - current\_score}{accept\_factor}\right) > Uniform(0.3, 1.0)$$

kde *accept_factor* reguluje silu selekčného tlaku.

Výsledkom prehľadávania stavového priestoru je zoznam stavov, ktoré algoritmus považoval za dostatočne pravdepodobné. V ideálnom prípade, algoritmus konverguje k jednému optimálnemu stavu, ktorý má najlepšie skóre a reprezentuje skutočnú identifikáciu taxonomického zaradenia baktérií vo vzorke a pomery ich abundancií. V prípade viacerých, podobne dobre ohodnotených stavov, algoritmus bude oscilovať medzi týmito lokálnymi extrémami a vo výsledkoch budeme môcť túto neistotu pozorovať.

Stav prehľadávaného systému obsahuje informáciu o mapovaní vstupných sekvencií na varianty baktérií, ale vo výsledku nás zaujímajú odhadované abundancie baktérií. Pre každú baktériu sa získajú počty priradených sekvencií k jednotlivým variantám. Odhadovaná abundancia baktérie je potom vypočítaná podľa vzťahu:

$$n = average(n_i / n_{v_i})$$

kde n_i je počet vstupných sekvencií priradených k variante v_i, n_{v_i} je abundancia varianty v_i a i iteruje cez indexy variánt danej baktérie.

5.4 Ohodnotenie stavu systému

Na ohodnotenie vygenerovaného stavu systému boli navrhuté dve triedy prístupov, a to z pohľadu podobností sekvencí a z pohľadu dodržania pomerov abundancií variant. Na vyhodnotenie podobnosti sekvencií boli zvolené prístupy založené na algoritme lokálneho zarovnania sekvencií Smith–Waterman, heuristickom nástroji `blast` a prístup vyhodnotenia

komplementarity nukleotidov na väzbových pozíciach. Na ohodnotenie dodržania pomerov abundancií variánt pri jednotlivých baktériách boli navrhnuté prístupy založené na binomiálnom rozdelení, Kullback–Leiblerovej divergencii a Jensen–Shannonovej divergencii.

5.4.1 Metóda `water`

Evaluátor `water` využíva algoritmus Smith-Waterman popísaný v kapitole 2.5.1 na určenie sekvenčnej podobnosti sekvencií génu 16S rRNA. Tento algoritmus vypočíta lokálne zarovnanie zadaných sekvencií a preto je možné ho použiť aj na klasifikáciu vstupných sekvencií, ktoré nepokrývajú celý gén 16S rRNA.

Jedným z výstupov zarovnania dvojice sekvencií je počet zhodných báz v zarovnaní. Táto hodnota je použitá na výpočet skóre podobnosti:

$$matching\_score = \frac{alignment[identical\_matches]}{alignment[query\_length]}$$

Vypočítané podobnosti sekvencií všetkých dvojíc (vstupná sekvencia, varianta baktérie) je pre urýchlenie výpočtu vhodné uložiť v maticovom tvare, ktorý je znázornený na obrázku 5.5. Táto matica môže byť predpočítaná pri všetkých evaluátoroch, ktoré hodnotia podobnosť sekvencie, pretože podobnosti sa počas prehľadávania stavového priestoru nemenia.

sekvencia_1	sekvencia_2	sekvencia_3		
1	**0.5**	0	varianta_1	Baktéria_A
0.5	**1**	0	varianta_2	
0	0	**1**	varianta_1	Baktéria_B

Obr. 5.5: Maticové zobrazenie uloženia podobnosti medzi vstupnými sekvenciami a variantami baktérií. Stĺpce zodpovedajú vstupným sekvenciám, riadky variantám baktérií a čísla v bunkách matice určujú podobnosť zodpovedajúcich sekvencií.

Hodnoty podobnosti uvedené v obrázku 5.5 sú pre názornosť zjednodušené. V skutočnosti sa ešte využíva poznatok, že sekvencie patriace jednému organizmu musia mať sekvenčnú podobnosť aspoň 97% [103]. Z toho dôvodu sú hodnoty podobnosti menšie ako 97% orezané na hodnotu 0% a podobnosť z intervalu $< 97\%, 100\% >$ je preškálovaná do intervalu $< 0\%, 100\% >$.

5.4.2 Metóda `blast`

Ak bol na predfiltrovanie použitý nástroj `blast`, potom je možné uložiť si hodnoty podobnosti do vhodnej štruktúry a použiť ich počas ohodnocovania systému. Toto ukladanie spočíva v troch krokov. V prvom kroku je vytvorený slovník všetkých kombinácií dvojíc

(vstupná sekvencia, varianta baktérie), pričom sa pracuje už s predfiltrovanou databázou baktérií.

V druhom kroku sa prehľadávajú výsledky nástroja `blast` a indentifikujú sa zarovnania, ktoré zodpovedajú dvojiciam vo vytvorenom slovníku. Prehľadávanie pomocou slovníka je potrebné z toho dôvodu, že výsledky `blast`-u obsahujú aj zarovnania, ktoré zodpovedajú baktériám, ktoré už boli odfiltrované. Zo získaného zarovnania je vypočítané skóre podobnosti podľa vzťahu:

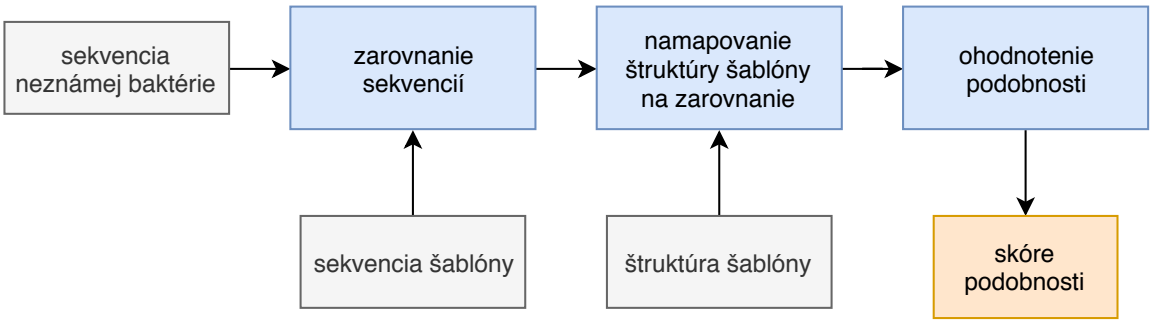
$$matching\_score = \frac{result[identical\_matches]}{result[query\_length]}$$

Ku každej dvojici sekvencií v slovníku je priradené najlepšie nájdené skóre. Vytvorený slovník sekvencií s priradeným normalizovaným skóre obsahuje všetky potrebné informácie pre ďalšiu prácu. Podobne ako pri použití metódy `water`, pre urýchlenie výpočtu ohodnotenia stavu systému je tento slovník v treťom kroku ešte prevedený do maticového tvaru znázorneného na obrázku 5.5.

5.4.3 Metóda `structure`

Evaluátor `structure` je navrhnutý na základe vlastnosti sekundárnej štruktúry 16S rRNA z toho dôvodu, že štruktúra je evolučne konzervovanejšia ako sekvencia. 16S rRNA je súčasťou ribozómu, ktorý svoju funkciu plní vďaka svojej štruktúre. Z toho dôvodu je zachovanie štruktúry dôležitejšie ako zachovanie sekvencie a môže sa stať, že dve veľmi podobné štruktúry môžu vzniknúť z pomerne odlišných sekvencií. Základom tejto metódy je transformácia všetkých používaných sekvencií do priestoru známych baktérií. Tento priestor obsahuje informáciu o sekundárnej štruktúre 16S rRNA a umožňuje jednoduchý výpočet podobnosti medzi jednotlivými sekvenciami.

Navrhovaný priestor známych baktérií si môžeme predstaviť ako N-rozmerný priestor, kde každá dimenzia predstavuje jednu známu baktériu, ktorá je špecifikovaná šablónov. Hodnoty každej dimenzie sú z intervalu $< 0, 1 >$ a reprezentujú skóre podobnosti štruktúry skúmanej 16S rRNA so štruktúrou 16S rRNA baktérie danej dimenzie. Schéma zistenia hodnoty jednej dimenzie pre jednu neznámu sekvenciu je znázornená na obrázku 5.6. Výsledkom transformácie sekvencie jednej neznámej baktérie je N-rozmerný vektor obsahujúci hodnoty z intervalu $< 0, 1 >$.



Obr. 5.6: Návrh procesu transformácie sekvencie 16S rRNA neznámej baktérie na skóre podobnosti jednej dimenzie v priestore známych baktérií. Vstupy procesu sú zobrazené sivou farbou, hlavné kroky zú zobrazené modrou farbou a finálny výstup je zobrazený oranžovou farbou.

Šablóny známych baktérií pozostávajú zo sekvencie a sekundárne štruktúry 16S rRNA baktérií, pre ktoré bola sekundárna štruktúra zistená (obrázok 5.7). V šablóne je štruktúra reprezentovaná zátvorkovou notáciou. Sekvencia aj štuktúra šablóny pokrýva celý gén 16S rRNA.

```
Filename:_d.16.b.S.pyogenes.GEN.bracket;Organism:_Streptococcus_pyogenes
UACGUAGGUCCCGAGCGUUGUCCGGAUUUAUUGGGCGUAAAGCGAGCGCAGGCGGUUUUUUAAGUCUGAAGUUAAAG
..)))))))))).)))))))))).........(((.....(.((((.(.(((.(((((((.((((((((((.....
```

Obr. 5.7: Navrhovaný formát šablóny známej baktérie. Šablóna obsahuje indentifikátor baktérie, jej sekvenciu a sekundárnu štruktúru v zátvorkovej notácii.

Ak sekvencia neznámej baktérie aj sekvencia v šablóne reprezentujú rovnakú oblasť 16S rRNA, potom je na ich zarovnanie vhodné použiť globálne zarovnanie sekvencií – algoritmus Needleman-Wunsch. Výstupom algoritmu je skóre zarovnania a samotné zarovnanie. Skóre zarovnania hovorí o vhodnosti zarovnania sekvencií, čo v tomto prípade nie je relevantné a vypovedajúcim bude až ohodnotenie podobnosti štruktúr. Dôležitým výsledkom algoritmu je teda samotné zarovnanie sekvencií, a hlavne pozície, na ktoré boli vložené medzery.

Aby bolo možné ďalej jednotne pracovať so sekvenciami aj štruktúrou, je potrebné premietnuť informácie získané zo zarovnania do štruktúry 16S v šablóne – vložiť medzery na miesta, kde boli vložené medzery pri globálnom zarovnaní sekvencií (obrázok 5.8).

Sekvencia neznámej baktérie

```
AUACUGUACCCAAAGCGUA
```

Šablóna

```
Example_template
AUACCGUACGGAAACGUA
.(((.((......)))))
```

Globálne zarovnanie sekvencie šablóny a neznámej sekvencie

```
AUACCGUACGGAAA-CGUA
||||.|||||..|||  ||||
AUACUGUACCCAAAGCGUA
```

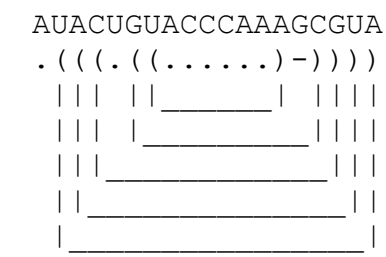
Vloženie medzier do štruktúry šablóny podľa zarovnania

```
.(((.((......)-))))
```

Obr. 5.8: Ukážka procesu úpravy štruktúry 16S v šablóne podľa globálneho zarovnania sekvencií. Podľa zarovnania sekvencie neznámej baktérie a sekvencie zo šablóny sú do štruktúry 16S v šablóne vložené medzery.

Upravená štruktúra 16S zo šablóny je následne namapovaná na zarovnanú sekvenciu neznámej baktérie. Ďalej je vyhodnotené, či párovanie zo šablóny môže nastať aj v neznámej sekvencii. Šablóna na obrázku 5.9 obsahuje 5 párov báz a všetky z nich môžu nastať na

daných miestach aj v sekvencii neznámej baktérie. Preto je normalizované skóre podobnosti pre tento príklad rovné 1.



Obr. 5.9: Ukážka namapovania upravenej štruktúry 16S zo šablóny na zarovnanú neznámu sekvenciu a vyhodnotenie možných párovaní.

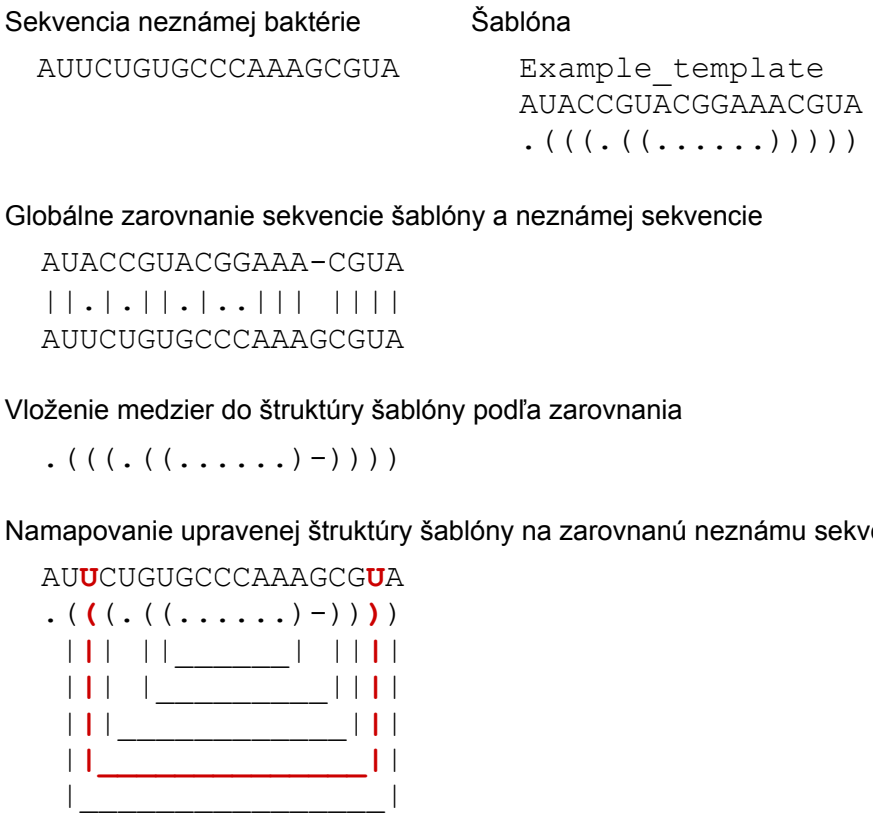
Iný výsledok nastáva v prípade, kedy sú sekundárne štruktúry 16S rRNA porovnávaných baktérií viac odlišné. Ak je na vstupe rovnaká šablóna ako na obrázku 5.9, ale odlišná neznáma sekvencia, možný výsledok je zobrazený na obrázku 5.10. Šablóna stále obsahuje päť párovaní, ale na zodpovedajúcich miestach v sekvencii je možné vytvoriť len štyri. Preto výsledné normalizované skóre pre tento príklad je 0,8, čo znamená, že porovnávané baktérie sú si menej podobné v sekundárnych štruktúrach, ako baktérie z obrázku 5.8. V príkladoch uvedených na obrázkoch 5.9 a 5.10 sa na vyhodnotenie, či môže párovanie nastať, používalo Watson-Crickovo párovanie.

Pseudokód výpočtu normalizovanej podobnosti štruktúr je uvedený v algoritme 4. Na vstupe je zarovnaná sekvencia neznámej baktérie a štruktúra 16S zo šablóny. Po zarovnaní majú oba vstupy rovnakú dĺžku, preto sú súbežne prechádzané po jednotlivých znakoch. Zátvorková notácia štruktúry označuje pozície, na ktorých by mali byť zodpovedajúce bázy spárované. Na zaručenie párovania zodpovedajúcich báz je použitá dátová štruktúra *zásobník*[1]. V prípade otváracej zátvorky sa pomocou operácie `stack.push(b)` vloží príslušná báza na vrchol zásobníka. Po nájdení najbližšej zatváracej zátvorky sa pomocou operácie `stack.pop()` získa posledná uložená báza a tá sa porovná s aktuálnou bázou.

Bázy na daných pozíciách sú získané zo skúmanej sekvencie a ich schopnosť vytvoriť párovanie je vyhodnotená funkciou `match()`. Výsledky tejto funkcie sú z intervalu $< 0, 1 >$ a vyjadrujú pravdepodobnosť, že daná dvojica báz vytvorí validný bázový pár. Výsledkom analýzy celej sekvencie neznámej baktérie je hodnota, ktorá vyjadruje súčet týchto pravdepodobností. Na záver je ešte vykonaná normalizácia počtom párov báz, ktoré sa nachádzali v štruktúre 16S zo šablóny. Táto normalizácia zaručí, že každá výsledná podobnosť štruktúr bude z intervalu $< 0, 1 >$. Podobnosť môže byť vyjadrená vzťahom:

$$\text{podobnosť} = \frac{\text{súčet pravdepodobností}}{\text{počet možných párov}}$$

[1] anglicky: stack



Obr. 5.10: Ukážka procesu úpravy štruktúry 16S zo šablóny podľa globálneho zarovnania sekvencie zo šablóny a sekvencie neznámej baktérie. Následné namapovanie upravenej štruktúry 16S zo šablóny na zarovnanú neznámu sekvenciu a vyhodnotenie možných párovaní. V uvedenej sekvencii neznámej baktérie nie je možné vytvoriť všetky párovania, ktoré sa nachádzajú v sekundárnej štruktúre zo šablóny. Nevytvorený pár U-U je zobrazený červenou farbou.

Algoritmus 4: Výpočet normalizovanej podobnosti štruktúr

```
Input: query_seq_align (aligned query sequence), templ_struct_align (aligned
       template structure)
Result: normalized_score
score = 0;
base_pairs = 0;
stack = empty_stack;
for (s, b) in (templ_struct_align, query_seq_align) do
    if s == "(" then
        stack.push(b);
    else
        if s == ")" and !stack.empty() then
            left_base = stack.pop();
            right_base = b;
            score += match(left_base, right_base);
            base_pairs += 1;
        end
    end
end
normalized_score = score / base_pairs;
```

Po transformácií sekvencií do priestoru známych baktérií bude pre všetky dvojice (vstupná sekvencia, varianta baktérie) vypočítaná ich podobnosť podľa vzťahov:

$$distance = euclidean(sequence, variant)$$

$$similarity = \frac{1}{1 + distance}$$

Pre urýchlenie výpočtu ohodnotenia stavu systému sú tieto hodnoty ešte uložené do maticového tvaru, kde stĺpce zodpovedajú vstupným sekvenciám a riadky zodpovedajú variantám baktérií. Bunky matice potom obsahujú hodnoty, ktoré určujú podobnosť vstupnej sekvencie v danom stĺpci so sekvenciou varianty v danom riadku. Ukážka tohto uloženia podobnosti, ktoré vychádza z obrázku 5.2, je znázornená na obrázku 5.5

5.4.4 Metóda `binomial`

Evaluátor `binomial` ohodnocuje priraďovanie sekvencií k variantám z pohľadu binomiálneho rozdelenia. Na sekvenovanie sa môžeme pozerať ako na náhodný výber sekvencie zo súboru všetkých možných sekvencií. Pri každom čítaní skúmame, či bola vybraná daná sekvencia, prípadne iná sekvencia. Tento výber môžeme chápať ako Bernoulliho proces, kde sledovaným javom je osekvenovanie zvolenej varianty. Počet sekvencií priradených k zvolenej variante bude teda podliehať binomiálnemu rozdeleniu.

Pseudokód tohto prístupu je uvedený v algoritme 5. Počty priradených sekvencií k jednotlivým variantám zodpovedajú súčtom hodnôt v riadkoch matice stavu systému. Odhadovaná abundancia je potom vypočítaná podľa vzťahu:

$$n = average(n_i / n_{v_i})$$

kde n_i je počet vstupných sekvencií priradených k variante v_i a n_{v_i} je abundancia varianty v_i. Následne môžeme vypočítať očakávaný počet priradených sekvencií k variante podľa vzťahu:

$$n'_i = n_{v_i} * n$$

a pravdepodobnosť vytiahnutia varianty pri náhodnom vzorkovaní vstupných sekvencií podľa vzťahu:

$$p(v_i) = \frac{n'_i}{N}$$

kde N je celkový počet vstupných sekvencií. Následne môžeme pomocou binomiálneho rozdelenia vypočítať ohodnotenie aktuálneho priradenia vstupných sekvencií k variantám:

$$score = Binom(n_i, N, p(v_i))$$

Binomiálne rozdelenie popisuje početnosť výskytu náhodného javu v nezávislých pokusoch, kde jav má stále rovnakú pravdepodobnosť. Počítame pravdepodobnosť, že pri náhodnom výbere N vstupných sekvencií priradíme n_i sekvencií variante v_i, ak poznáme pravdepodobnosť vytiahnutia danej varianty – $p(v_i)$.

Algoritmus 5: Ohodnotenie stavu na základe binomiálneho rozdelenia

```
for každú baktériu do
    získaj počty priradených sekvencií k jednotlivým variantám;
    vypočítaj odhadovanú abundanciu baktérie;
    for každú variantu do
        vypočítaj očakávaný počet priradených sekvencií;
        vypočítaj pravdepodobnosť vytiahnutia varianty pri náhodnom vzorkovaní
         vstupných sekvencií;
        vypočítaj skóre pomocou binomiálneho rozdelenia;
    end
end
```

5.4.5 Metóda `kl`

Metóda `kl` pracuje s abundanciami variánt ako s pravdepodobnostným rozložením a počíta Kullback–Leiblerovu divergenciu [68] medzi očakávaným a získaným pravdepodobnostným rozložením.

Ak máme baktériu s tromi variantami, kde jednotlivé abundancie sú 4, 1 a 3, potom to môžeme reprezentovať pravdepodobnostným rozložením [0.5, 0.125, 0.375]. Podobne je vypočítané rozloženie nad počtami priradených sekvencií k jednotlivým variantám baktérie v aktuálnom stave systému. Kullback–Leiblerova divergencia potom určuje, koľko informácie sa stratí, keď nahradíme očakávané pravdepodobnostné rozloženie rozložením v aktuálnom stave systému.

Kullback–Leiblerova divergencia vychádza z výpočtu entropie, ktorá je pre pravdepodobnostné rozloženie p definovaná nasledovne:

$$H = -\sum_{i=1}^{N} p\left(x_i\right) \cdot \log p\left(x_i\right)$$

Ak na výpočet použijeme log_2, môžeme entropiu interpretovať ako minimálny počet bitov, ktorý potrebujeme na zakódovanie informácie. Kullback–Leiblerova divergencia je len malou modifikáciou výpočtu entropie. K očakávanému pravdepodobnostnému rozloženiu p pridáme odhadované pravdepodobnostné rozloženie q a divergenciu vypočítame podľa vzťahu:

$$D_{KL}(p\|q) = \sum_{i=1}^{N} p\left(x_i\right) \cdot \left(\log p\left(x_i\right) - \log q\left(x_i\right)\right)$$

Pomocou Kullback–Leiblerovej divergencie môžeme presne vypočítať, koľko informácie sa stratí, keď aproximujeme jedno rozloženie iným [69]. Pri ohodnocovaní stavu systému nás však zaujíma podobnosť, preto je divergenciu ešte potrebné transformovať podľa vzťahu:

$$score = \frac{1}{1 + D_{KL}(p\|q)}$$

5.4.6 Metóda `js`

Metóda `js`, podobne ako metóda `kl`, sa na abundancie variánt pozerá ako na pravdepodobnostné rozloženie a následne počíta Jensen–Shannonovu divergenciu [32] medzi očakávaným a získaným pravdepodobnostným rozložením. Je založená na Kullback–Leibler divergencii [68] s niektorými významnými (a užitočnými) rozdielmi, vrátane toho, že je symetrická a má vždy konečnú hodnotu. Jensen–Shannonova divergencia $\mathrm{D_{JS}(p\|q)}$ je definovaná ako:

$$\mathrm{D_{JS}}(p\|q) = \frac{1}{2}D(p\|m) + \frac{1}{2}D(q\|m)$$

kde $D(p\|q)$ je Kullback–Leiblerova divergencia a $m = \frac{1}{2}(p + q)$. Aj túto divergenciu je potrebné transformovať podľa vzťahu:

$$score = \frac{1}{1 + D_{JS}(p\|q)}$$

Kapitola 6

Implementácia navrhnutého nástroja

Nástroj navrhnutý v kapitole 5 bol úspešne implementovaný s využitím programovacích jazykov Python 3, R a Bash. Celá implementácia pozostáva z niekoľkých modulov umiestnených v podzložke `src`. Kľúčovými modulmi sú modul `inference`, ktorý obsahuje implementáciu klasifikácie na základe variánt 16S rRNA a modul `structure`, ktorý obsahuje implementáciu transformácie sekvencií do priestoru známych baktérií. Zvyšné moduly obsahujú rôznu pomocnú funkcionalitu. Moduly `lib` a `taxassign` boli poskytnuté vedúcim práce, Stanislavom Smatanom.

6.1 Klasifikácia na základe variánt 16S rRNA

Klasifikácia na základe variánt 16S rRNA vyžaduje databázu baktérií, ktorá bude obsahovať všetky kópie génu 16S rRNA baktérií a ich taxonomické zaradenie. Táto databáza bola vytvorená vedúcim tejto práce, Stanislavom Smatanom, z dát dostupných v databáze *NCBI Refseq* [90].

Databáza známych baktérií a sekvencií ich variánt génu 16S rRNA je reprezentovaná triedou `WholeDB`. Táto trieda spracuje vstupnú databázu baktérií vo formáte `csv` a baktéri uloží ako list objektov `KnownBacteria`. Tento objekt obsahuje informáciu o taxonomickom zaradení danej baktérie a ďalej obsahuje sekvencie variánt baktérie a ich abundancie. Trieda `WholeDB` obsahuje metódu `close_bacteria`, ktorá môže byť využítá pri predfiltrovaní databázy. Vstupnými parametrami tejto metódy sú sekvencia génu 16S rRNA a prah. Výstupom metódy je zoznam baktérií, ktoré obsahujú aspoň jednu variantu dostatočne podobnú vstupnej sekvencii. Podobnosť je v tomto prípade počítaná pomocou k-merového spektra.

Neznáme vstupné sekvencie, ktoré sú predmetom klasifikácie, sú spravované triedou `InputSequences`. Trieda ponúka metódu `parse_from_fasta`, ktorá ako vstupný parameter očakáva cestu k súboru so vstupnými sekvenciami vo formáte `fasta`. Identifikátory sekvencií vo vstupnom súbore su ignorované, uložené sú len sekvencie. Zhodné sekvencie sú zlúčené do jedného objektu, ktorý okrem sekvencie obsahuje aj informáciu o abundancii.

Nasledujúcim krokom v klasifikácii baktérií na základe variánt 16S rRNA je predfiltrovanie databázy známych baktérií, na čo slúži trieda `ConstrainedDB`. Vstupnými parametrami je databáza známych baktérií reprezentovaná objektom `WholeDB`, neznáme vstupné sekven-

cie reprezentované objektom `InputSequences`, prah a metóda, pomocou ktorej sa majú vyhodnocovať podobnosti sekvencií. Implementované boli dva prístupy: *k-mer* a *blast*.

Prístup *k-mer*, implementovaný v metóde `create_kmer`, využíva k-merové spektrum vstupných sekvencií a sekvencií variánt známych baktérií. Tieto spektrá sú pre sekvencie nemenné a preto sú predpočítané už pri úvodnom ukladaní sekvencií. Implementácia výpočtu k-merového spektra bola prevzatá z diplomovej práce Nikoly Valešovej [110]. Na predfiltrovanie sa využíva popísaná metóda `close_bacteria` z triedy `WholeDB`.

Prístup *blast*, implementovaný v metóde `create_blast`, využíva lokálne zarovnanie sekvencií. Na získanie týchto zarovnaní bolo implementované vlastné rozhranie pre unixový nástroj `blast` od *NCBI* [77]. Taktiež bola vytvorená lokálna indexovaná databáza známych baktérií, ktorú nástroj `blast` používa pri vyhľadávaní.

Predfiltrovanie databáze baktérií využíva informáciu o identických znakoch zarovnania. K ďalšiemu spracovaniu sú vybrané len tie baktérie, ktorých varianta je dostatočne podobná vstupnej sekvencii. Výsledné zarovnania z nástroja `blast` sú ďalej uložené pre zjednodušenie následného ohodnotenia stavu systému. Toto ukladanie je implementované v metóde `create_alignments_array`.

Mapovanie medzi vstupnými sekvenciami a variantami baktérií, teda stav systému, je reprezentované triedou `AssignmentMatrix`. Trieda je inicializovaná vstupnými sekvenciami reprezentovanými triedou `InptuSequences` a predfiltrovanou databázou známych baktérií reprezentovanou triedou `ConstrainedDB`. Trieda ďalej obsahuje kľúčové metódy pre použitie algoritmu Metropolis Hastings:

- `init_state`, ktorá vygeneruje náhodný stav systému,
- `generate_proposed_state`, ktorá vygeneruje nový stav systému na základe aktuálneho stavu,
- `evaluate_current_state` a `evaluate_proposed_state`, ktoré ohodnotia aktuálny a novo vygenerovaný stav systému,
- `move_to_proposed_state`, ktorá priradí novo vygenerovaný stav do aktuálneho stavu,
- `get_current_state_summary`, ktorá slúži na získanie aktuálnych čiastkových ohodnotení a na získanie odhadovaných abundancií jednotlivých skúmaných baktérií.

Samotná trieda `AssignmentMatrix` neobsahuje implementáciu na ohodnotenie stavu. Implementácia sa nachádza v module `evaluators`, ktorý obsahuje triedy pre jednotlivé prístupy k ohodnoteniu stavu systému. Každá trieda ponúka rovnaké rozhranie, čím je umožnené, aby bolo možné prístupy k ohodnoteniu ľubovolne zamieňať bez potreby úpravy triedy `AssignmentMatrix`. Použité evaluátory však musia byť inicializované rovnakými vstupnými sekvenciami a predfiltrovanou databázou baktérií ako trieda `AssignmentMatrix`, pre ktorú majú byť použité na ohodnotenie stavu.

Každý evaluátor poskytuje globálnu premennú `NAME`, kde je uložený jeho jedinečný identifikátor. Ďalej poskytuje metódu `evaluate`, ktorá na vstupe prijíma stav systému, ktorý má byť ohodnotený a vracia vypočítané ohodnotenie stavu. Počas výpočtu sú jednotlivé pravdepodobnosti prevádzané do logaritmického merítka, čím je zvýšená rýchlosť a numerická stabilita výpočtu. Každý evaluátor ešte umožňuje použitie normalizácie výsledného skóre do intervalu $< 0, 1 >$. V tom prípade je potrebné nastaviť pôvodný rozsah skóre do atribútu `normalize_interval` a atribútu `normalize` nastaviť hodnotu `True`. Pôvodný rozsah skóre je potrebné získať pred spustením samotnej optimalizácie algoritmom Metropolis

Hastings. Implementácia sa nachádza v metóde `find_normalization_intervals` v triede `MetropolisHastings`. Normalizácia je implementovaná v metóde `normalize_score`. Evaluátor `water` využíva na výpočet lokálneho zarovnania sekvencií unixový nástroj `water` z balíku *EMBOSS* [76].

Trieda `MetropolisHastings` je inicializovaná dvoma parametrami: objektom, ktorý reprezentuje stav systému (v tomto prípade `AssignmentMatrix`) a zoznamom názvov evaluátorov, ktoré majú byť následne použité na ohodnotenie stavu. Podľa názvov sú vytvorené zodpovedajúce objekty evaluátorov. Kľúčovou metódou triedy `MetropolisHastings` je metóda `walk`, ktorá vykoná prehľadávanie stavového priestoru. Vyhodnotenie, či má byť navrhovaný stav prijatý, teda či je navrhovaný stav dostatočne dobrý, sa nachádza v metóde `is_proposed_state_good_enough`.

Výaledkom prehľadávania stavového priestoru je tabuľka, ktorá obsahuje informácie o každom prijatom stave. Zaznamenávajú sa tieto údaje:

- celkové ohodnotenie stavu,
- čiastkové ohodnotenia stavu získané jednotlivými evaluátormi,
- abundacie jednotlivých baktérií.

6.2 Transformácia sekvencií do priestoru známych baktérií

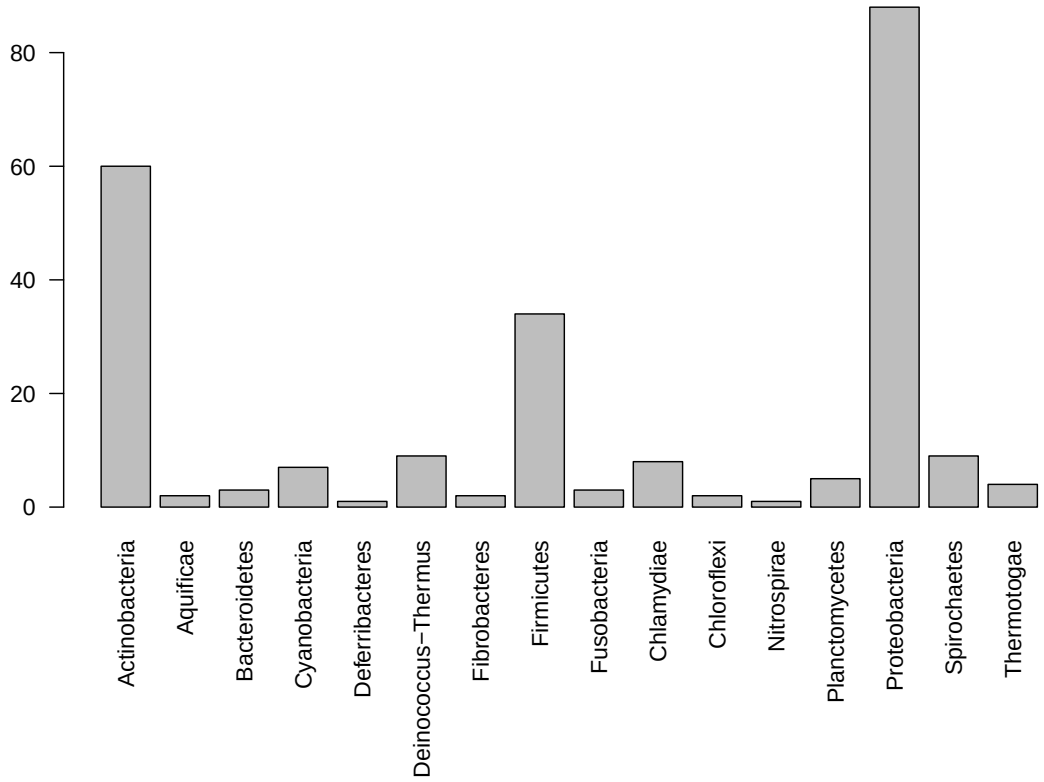
Evsluátor `structure` používa transformáciu sekvencií do priestoru známych baktérií. Táto transformácia je implementovaná v triede `ScoreTransformation`. Trieda je inicializovaná dvoma súbormi, ktoré obsahujú šablóny. Jeden súbor je vo formáte `templ`, ktorý bol navrhnutý v kapitole 5.4.3. Druhý súbor obsahuje rovnaké baktérie, ale je vo formáte `fasta` a využíva sa na zjednodušenie následného zarovnania sekvencií.

Inicializovaná trieda môže byť opakovane volaná s rôznymi súbormi vo formáte `fasta`, ktoré obsahujú sekvencie určené na transformáciu. Transformácia spočíva vo výpočte normalizovanej podobnosti štruktúr medzi vstupnými sekvenciami a šablónami. Výpočet tej to podobnosti bol implementovaný podľa algoritmu 4. Na získanie globálneho zarovnania sekvencií bol použitý nástroj `needle` z balíku *EMBOSS* [76].

Transformácia sekvencií do priestoru známych baktérií vyžaduje šablóny, ktoré obsahujú sekvencie aj sekundárne štruktúry 16S rRNA známych baktérií. Sekvencie aj štruktúry boli získané z Comparative RNA Web Site and Project [24]. Nejedná sa síce o štruktúry získané experimentálne (napríklad rentgenovou kryštalografiou), ale pomocou komparatívnej analýzy sekvencií. Jedná sa o pomerne kvalitnú predikciu sekundárnej štruktúry, ktorá ponúka dostatočné množstvo štruktúr na vytvorenie šablón na transformáciu do priestoru známych baktérií. Dostupné štruktúry rozdelené podľa kmeňov sú zobrazené na obrázku 6.1.

Z CRW stránky boli získané súbory dvoch formátov: `bracket`, ktorý obsahuje sekundárnu štruktúru v zátvorkovej notácii, ale neobsahuje sekvenciu a formát `rnaml`, ktorý obsahuje informáciu o sekvencii aj štruktúre. Získanie týchto súborov je implementované v skripte `crw_retriever.py` v module `template.common`. Súbory boli následne analyzované, pričom sekvencia bola získaná zo súboru vo formáte `rnaml` a štruktúra z zodpovedajúceho súboru vo formáte `bracket`.

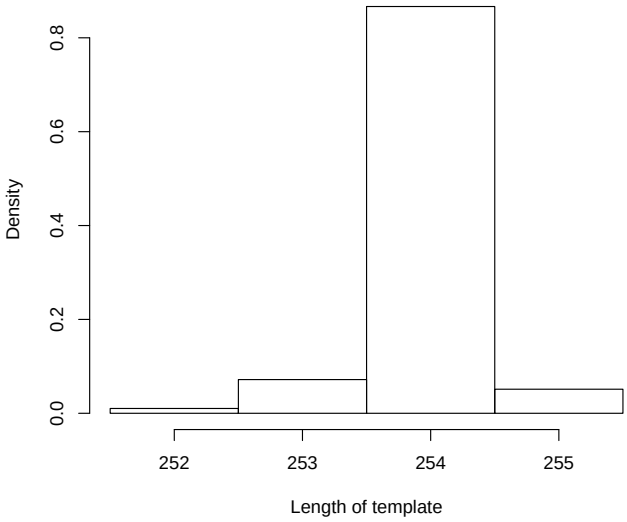
Získané súbory zodpovedajú celej dĺžke 16S rRNA, ale pre niektoré aplikácie je vhodné orezať sekvencie na požadovaný región, napríklad V4. Primery používané pri sekvenácií regiónu V4 boli získané zo stránky 16S Illumina Amplicon Protocol [1]. Pomocou týchto primerov boli orezané získané sekvencie známych baktérií a im zodpovedajúce sekundárne



Obr. 6.1: Rozloženie výskytu jednotlivých kmeňov baktérií v sekundárnych štruktúrach dostupných na stránke CRW.

štruktúry. Implementácia sa nachádza v skripte `join_templates.py` v module `template`. Tento skript obsahuje aj implementáciu vytvorenia šablón, ktoré pokrývajú celý gén 16S rRNA. Na orezanie bol použitý skript `v_ripper.py` z nástroja SPINGO [18].

Korektnosť orezania šablón bola overená porovnaním získaných dĺžok regiónov V4. V článku od Illuminy [5] je uvedené, že región V4 väčšiny mikrobiálnych druhov je dlhý približne 254 párov báz, prípadne sa od tejto dĺžky líši iba niekoľkými pármi. Rozloženie dĺžok získaných šablón je zobrazené na obrázku 6.2.



Obr. 6.2: Rozloženie dĺžok šablón orezaných na región V4. Získané rozloženie odpovedá zisteniam, že región V4 väčšiny mikrobiálnych druhov je dlhý približne 254 párov báz.

Kapitola 7

Vyhodnotenie navrhnutého nástroja

Po implementácii navrhnutého klasifikátoru bolo potrebné vyhodnotiť jeho výkonnosť a analyzovať jeho schopnosti. Kvôli komplexnosti klasifikátoru na základe variánt 16S rRNA bol tento klasifikátor vyhodnotený na niekoľkých úrovniach. Na vyhodnotenie bol vytvorený skript, ktorý inicializuje všetky potrebné objekty a spustí klasifikáciu. Tento skript pracuje s konfiguračným slovníkom, v ktorom sú definované všetky potrebné parametre pre klasifikáciu. Požadované parametre sú uvedené v tabuľke 7.1.

Parameter	**Dátový typ**	**Popis**
whole_DB	reťazec	Cesta k `csv` súboru, ktorý obsahuje taxonómie a varianty známych baktérií.
DB_method	reťazec	Metóda, ktorá sa má použiť na predfiltrovanie databázy.
threshold	celé číslo	Prah vzdialenosti použitý pri predfiltrovaní databázy baktérií.
input_samples	reťazec	Cesta k `fasta` súboru so vstupnými sekvenciami.
evaluation_methods	pole reťazcov	Názvy požadovaných evaluátorov.
normalization	boolean	Zapnutie/vypnutie určenia normalizačných intervalov pre jednotlivé evaluátory.
normalization_steps	celé číslo	Počet krokov algoritmu Metropolis Hastings, ktoré sa majú použiť na určenie normalizačných intervalov.
steps	pole celých čísel	Pre každú hodnotu poľa je spustené prehľadávanie stavového priestoru s daným počtom krokov. Zvyšné parametre sú pri každom spustení rovnaké.

Tabuľka 7.1: Parametre potrebné na spustenie klasifikácie na základe variánt 16S rRNA.

7.1 Základné dátové sady

Prvé testovanie a ladenie klasifikátoru prebehlo na dvoch ručne pripravených dátových sadách, ktoré sú určené na overenie funkčnosti a nájdenie problematických častí implementovaného klasifikátoru. Prvá dátová sada predstavuje jednoznačné mapovanie a sekvencie tejto dátovej sady boli získané priamo z databázy známych baktérií, ktorá sa používa aj pri samotnej klasifikácii. Boli zvolené dve baktérie: *Bifidobacterium bifidu*, kmeň *PRL2010* a *Mesorhizobium ciceri*, kmeň *WSM1271*. Baktéria *Bifidobacterium bifidu* obsahuje 3 kópie génu 16S rRNA, ktoré sú identické a baktéria *Mesorhizobium ciceri* obsahuje 2 kópie, ktoré sa líšia len pridaním jednej bázy. Podobnosť sekvencií medzi týmito dvoma baktériami je približne 76 %. Dátová sada obsahuje sekvencie zodpovedajúce baktérii *Bifidobacterium bifidu* s abundanciou 3 a baktérii *Mesorhizobium ciceri* s abundanciou 4.

Druhá dátová sada pozostáva zo sekvencií patriacich trom baktériám: *Staphylococcus lugdunensis* kmeň *HKU09-01*, *Staphylococcus lugdunensis* kmeň *N920143* a *Hyperthermus butylicus* kmeň *DSM 5456*. Kmeň *HKU09-01* obsahuje 6 kópií génu 16S rRNA, pričom unikátnych z nich je 5, a teda jedna kópia má abundanciu 2. Kmeň *N920143* obsahuje 4 kópie génu 16S rRNA. Všetky kópie sú v rámci kmeňa *N920143* unikátne, ale dve kópie sú zdieľané s kmeňom *HKU09-01*. Baktéria *Hyperthermus butylicus* obsahuje jednu kópiu génu, ktorá je unikátna medzi kópiami tejto dátovej sady. Dátová sada obsahje sekvencie, ktoré zodpovedajú uvedeným baktériám s abundanciami 2, 4 a 7. Zhrnutie týchto dátových sád je zobrazené v tabuľke 7.2.

Dátová sada	**Baktéria**	**Abundancia**	**Počet variant**	**Pomery abundancií variant**	**Popis**
A	Bifidobacterium bifidu	3	1	3	Tri zhodné kópie génu 16S rRNA.
	Mesorhizobium ciceri	4	2	1:1	Dve kópie génu 16S rRNA, ktoré sa líšia o jednu bázu.
B	Staphylococcus lugdunensis, HKU09-01	2	5	1:1:1:1:2	Šesť kópií génu 16S rRNA, z ktorých je unikátnych päť.
	Staphylococcus lugdunensis, N920143	4	4	1:1:1:1	Dve kópie génu 16S rRNA sú zdieľané s kmeňom HKU09-01.
	Hyperthermus butylicus, DSM_5456	7	1	1	Kópia je unikátna v rámci celej dátovej sady.

Tabuľka 7.2: Dátové sady vytvorené na úvodné vyhodnotenie implementovaného klasifikátoru na základe variánt 16S rRNA.

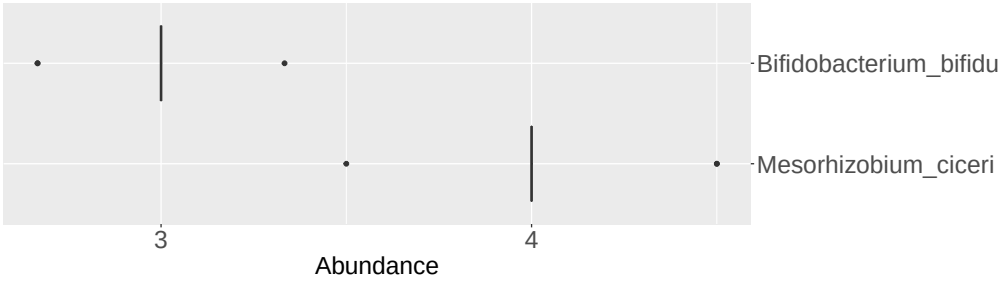
Klasifikátor bol spustený s parametrami uvedenými v tabuľke 7.3 s dátovou sadou A a následne s dátovou sadou B. Pre každú dátovú sadu bola klasifikácia spustená dvakrát:

predfiltrovanie databázy známych baktérií pomocou metódy `k-mer` a následne pomocou metódy `blast`. Hodnota parametru `threshold` bola zvolená ako 1 z dôvodu overenia schopnosti metódy predfiltrovania databázy nájsť baktérie s presnou zhodou sekvencií. Ďalej boli zvolené dva evaluátory, pričom evaluátor `water` ohodnocuje sekvenčnú podobnosť priradených sekvencií a evaluátor `binomial` ohodnocuje pomery priradených sekvencií k variantám.

Parameter	Hodnota
whole_DB	data/knownDB.csv
threshold	1
evaluation_methods	[water, binomial]
normalization	False
normalization_steps	0
steps	[10000]

Tabuľka 7.3: Parametre klasifikácie použité na overenie schopnosti metód predfiltrovania databázy nájsť baktérie s presnou zhodou sekvencií. Normalizácia skóre je vypnutá a použitá je jedna kombinácia evaluátorov – `[water, binomial]`.

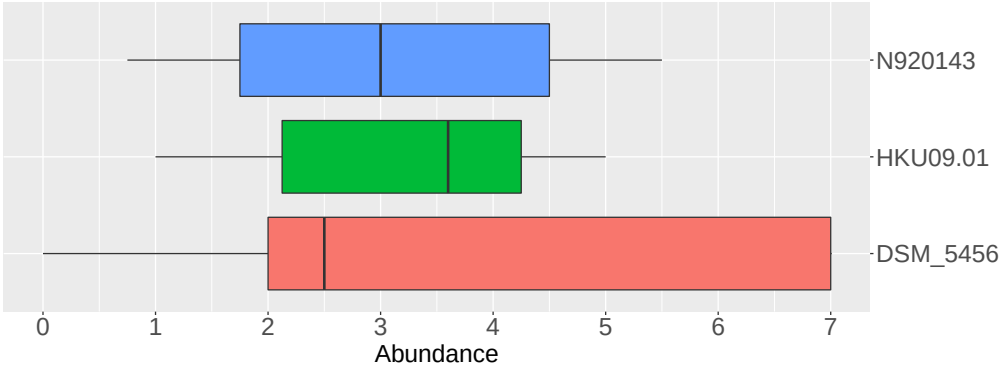
Klasifikátor vo všetkých prípadoch zvolil korektné baktérie v kroku predfiltrovania databázy baktérií. Počas následného prehľadávania stavového priestoru boli pre dátovú sadu A určené aj správne abundancie jednotlivých baktérií. Pri dátovej sade B sa klasifikátoru nepodarilo skonvergovať k správnym abundanciám. Rozloženie abundancií v prijatých stavoch systému pri predfiltrovaní databázy pomocou metódy `blast` pre dátovú sadu A je zobrazné na obrázku 7.1 a pre dátovú sadu B na obrázku 7.2.



Obr. 7.1: Rozloženie abundancií baktérií v prijatých stavoch systému pre dátovú sadu A. Na predfiltrovanie databázy bola použitá metóda `blast`. Zvyšné parametre klasifikácie sú uvedené v tabuľke 7.3. Očakávané abundancie boli 3 pre baktériu *Bifidobacterium bifidu* a 4 pre baktériu *Mesorhizobium ciceri*.

7.2 Porovnanie kombinácií evaluátorov

Nasledujúci experiment spočíva v porovnaní jednotlivých evaluátorov a ich schopnosti nájsť požadovaný stav systému. Klasifikátor bol opakovane spustený s parametrami uvedenými v tabuľke 7.4. Ako metóda na predfiltrovanie databázy baktérií bola zvolená metóda `blast`, aby bolo neskôr možné použiť zodpovedajúci evaluátor. Počet krokov prehľadávania bol znížený na 5000, pretože v predchádzajúcom experimente bolo zistené, že klasifikátor dokáže pre dátovú sadu A skonvergovať aj s takýmto počtom krokov. Evaluačné metódy klasifiká-



Obr. 7.2: Rozloženie abundancií baktérií v prijatých stavoch systému pre dátovú sadu B. Na predfiltrovanie databázy bola použitá metóda `blast`. Zvyšné parametre klasifikácie sú uvedené v tabuľke 7.3. Očakávané abundancie boli 4 pre kmeň *N920143*, 2 pre kmeň *HKU09-01* a 7 pre kmeň *DSM_5456*.

toru boli postupne nastavené na všetky kombinácie zo zoznamu dostupných evaluátorov: `[kl, js, binomial, blast, water, structure]`.

Parameter	**Hodnota**
whole_DB	data/knownDB.csv
DB_method	blast
threshold	1
input_samples	data/datasetA.fasta
evaluation_methods	všetky kombinácie
normalization	False
normalization_steps	0
steps	[5000]

Tabuľka 7.4: Parametre klasifikácie použité s dátovou sadou A popísanou v tabuľke 7.2 pri experimente na porovnanie všetkých kombinácií evaluátorov. Normalizácia skóre je vypnutá.

Klasifikácia bola spustená 10-krát pre všetky kombinácie evaluátorov a boli sledované nasledovné parametre:

- výpočetný čas prehľadávania stavového priestoru pri 5000 krokoch,
- krok, v ktorom boli prvý krát objavené korektné abundancie baktérií,
- abundancie baktérií, ku ktorým systém skonvergoval – nachádzajú sa vo výstupe najčastejšie.

Na obrázku A.1 v prílohe A sú zobrazené časy výpočtu prehľadávania stavového priestoru pri 5000 krokoch a pri použití rôznych kombinácií evaluátorov. Výrazné spomalenie prehľadávania stavového prietoru nastáva pri použití evaluátoru `binomial`.

Ďalším sledovaným parametrom bol krok, v ktorom boli prvýkrát objavené korektné abundancie baktérií. Získané výsledky sú zobrazené na obrázku A.2 v prílohe A. Niektoré

kombinácie evaluátorov neboli schopné objaviť správne abundancie (na obrázku A.2 v prílohe A zobrazené bielou farbou) v niektorých behoch klasifikácie ani po 5000 krokoch. Jedná sa o evaluátor `binomial` a jeho kombinácie s evaluátormi `js`, `kl` a `structure`.

Pri evaluátoroch `kl`, `binomial` a `js` je to z dôvodu, že tieto evaluátory ohodnocujú len pomery sekvencií priradených k variantám baktérií. Pri dátovej sade, kde je na vstupe 17 sekvencií a pri zodpovedajúcich baktériách, kde prvá baktéria obsahuje len jednu variantu a druhá baktéria obsahuje dve varianty v pomere 1:1, existuje niekoľko možností namapovania sekvencií na varianty, ktoré tieto pomery spĺňajú. Uvedené evaluátory všetky tieto možnosti ohodnotia ako korektný stav, a preto je potrebné použiť aj ohodnocovanie na základe podobnosti sekvencí.

Ďalšie kombinácie evaluátorov, ktorým sa nepodarilo nájsť správne riešenie, obsahovali evaluátor `structure`. Ukázalo sa, že pridanie evaluátora, ktorý ohodnocuje podobnosť štruktúr, nemusí vždy stačiť na nájdenie správneho výsledku. Zaujímavé však je, že použitie evaluátoru `structure` osamote, alebo v kombinácií len s evaluátorom `js`, viedlo k správnemu výsledku. Možným vysvetlením je, že evaluátory, ktorých skóre je v rôznych rádoch, nie sú schopné spolu úspešne ohodnocovať stavy systému.

Schopnosť klasifikátoru objaviť korektný stav ešte nie je vypovedajúca o jeho úspešnosti. Dôležité je, aby klasifikátor počas prehľadávania stavového priestoru do tohto stavu skonvergoval, aby daný stav prevažoval medzi prijatými stavmi. Rozloženia abundancií baktérií, ku ktorým klasifikátor skonvergoval pri použití rôznych kombinácií evaluátorov, sú zobrazené na obrázku A.3 v prílohe A. Podmnožina týchto výsledkov pre kombinácie evaluátorov, ktorým sa nepodarilo skonvergovať k správnym abundanciám, je na obrázku 7.3. Problém skonvergovať k správnemu riešeniu majú kombinácie evaluátorov `kl`, `js`, `binomial` a `structure`. V nasledujúcich experimentoch sú preto použité len kombinácie evaluátorov, kde jeden hodnotí podobnosť sekvencií a druhý dodržanie pomerov abundancií variant.

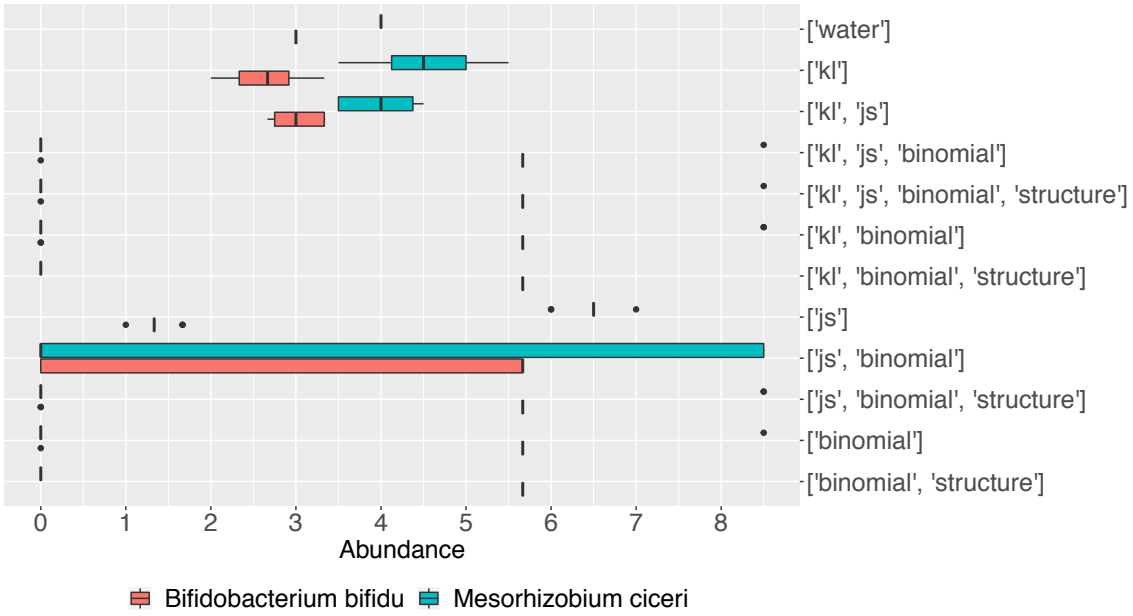
7.3 Vplyv normalizácie skóre

Následne bol vyhodnotený vplyv normalizácie skóre na presnosť získaných výsledných abundancií. Klasifikácia bola spustená s parametrami uvedenými v tabuľke 7.5. Evaluačné metódy boli tvorené dvojicami, kde jeden evaluátor hodnotí podobnosť sekvencií a druhý evaluátor hodnotí dodržanie pomerov abundancií variant.

Parameter	**Hodnota**
whole_DB	data/knownDB.csv
DB_method	blast
threshold	1
input_samples	data/datasetB.fasta
evaluation_methods	kombinácie dvojíc (evaluátor na ohodnotenie dodržania pomerov abundancií variant, evaluátor na ohodnotenia podobnosti sekvencií)
steps	[50000]

Tabuľka 7.5: Parametre klasifikácie použité s dátovou sadou B popísanou v tabuľke 7.2 pri experimente na vyhodnotenie vplyvu normalizácie skóre na presnosť získaných výsledkov.

Experiment pozostával z dvoch častí, kde v prvej časti bola normalizácia skóre vypnutá a v druhej časti bola zapnutá s 50000 krokmi na nájdene pôvodných intervalov skóre. Každá



Obr. 7.3: Výsledné odhadnuté abundancie baktérií, ku ktorým klasifikátor skonvergoval – majú najvyššie zastúpenie vo výsledných prijatých stavoch. Zobrazené abundancie sú z 10 iterácií pri použití parametrov uvedených v tabuľke 7.4 pre kombinácie evaluátorov, ktorým sa nepodarilo skonvergovať k správnym abundanciám. Evaluátor `[water]` dosiahol rovnaký (korektný) výsledok ako všetky nezobrazené kombinácie evaluátorov. Skutočné hodnoty abundancií sú abundancia 3 pre baktériu *Bifidobacterium bifidu* a abundancia 4 pre baktériu *Mesorhizobium ciceri*. Kompletné výsledky sú na obrázku A.3 v prílohe A.

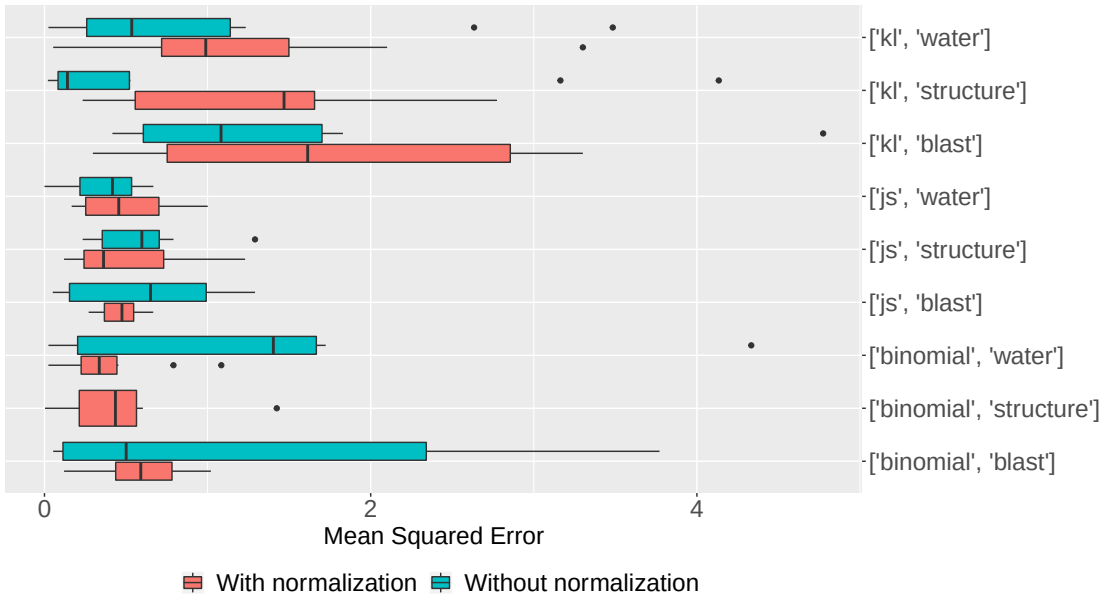
časť bola spustená 10-krát pre každú kombináciu evaluátorov. Pre každý výsledok bola vypočítaná stredná kvadratická chyba[1] podľa vzťahu:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

kde Y_i sú očakávané abundancie baktérií a $\hat{Y}_i$ sú abundancie baktérií, ku ktorým klasifikátor skonvergoval. Porovnanie výslednej strednej kvadratickej chyby pre klasifikácie bez použitia normalizácie skóre a s použitím normalizácie sú zobrazené na obrázku 7.4.

Bez použitia normalizácie sú najlepšie výsledky dosahnuté dvojicou evaluátorov `[kl, structure]`, ale nie v každej iterácii. Konzistentne dobré výsledky dosahujú dvojice evaluátorov `[js, water]`, `[js, structure]` a `[js, blast]`. Najväčšej chyby sa dopúšťa dvojica evaluátorov `[binomial, structure]`, ktorá nie je zobrazená, pretože stredná priemerná kvadratická chyba má medián väčší než 24. V prípade použitia normalizácie skóre dosahujú najlepšie výsledky dvojice evaluátorov `[binomial, water]` a `[binomial, structure]`. Dvojice evaluátorov, ktoré obsahujú evaluátor `kl`, sa dopúšťajú väčšej chyby v prípade použitia normalizácie.

[1] anglicky: Mean Squared Error



Obr. 7.4: Stredné kvadratické chyby výsledných abundancií od očakávaných abundancií baktérií. Klasifikácia bola spustená s parametrami uvedenými v tabuľke 7.5 10-krát pre každú dvojicu evaluátorov, bez normalizácie a s normalizáciou. Výsledok dvojice evaluátorov `[binomial, structure]` bez normalizácie nie je zobrazený, pretože stredná kvadratická chyba mala medián väčší než 24.

7.4 Pokročilejšie dátové sady

Ďalšie testovanie implementovaného nástroja bolo vykonané na dvoch náhodne vygenerovaných dátových sadách. Obe dátové sady obsahujú sekvencie zodpovedajúce variantám dvadsiatich náhodne vybraných baktérií z databázy známych baktérií. V prvej dátovej sade boli baktérie vybraté podľa uniformného rozloženia a každá baktéria má abundanciu jedna. Druhá dátová sada bola vygenerovaná podľa logaritmického rozloženia s parametrom $p = 0.9$. Logaritmické rozloženie sa používa na aproximáciu pomerov abundancií baktérií v bakteriálnych spoločenstvách [47]. Druhá dátová sada teda obsahuje sedem unikátnych baktérií s abundanciami zodpovedajúcimi logaritmickému rozloženiu. Zhrnutie dátový sád je uvedené v tabuľke 7.6.

Dátová sada	**Počet baktérií**	**Počet unikátnych baktérií**	**Počet sekvencií**	**Pomery abundancií baktérií**
C	20	20	85	všetky baktérie majú abundanciu 1
D	20	7	87	10:5:1:1:1:1:1

Tabuľka 7.6: Charakteristika pokročilejších dátových sad vygenerovaných na vyhodnotenie implementovaného klasifikátoru na základe variánt 16S rRNA.

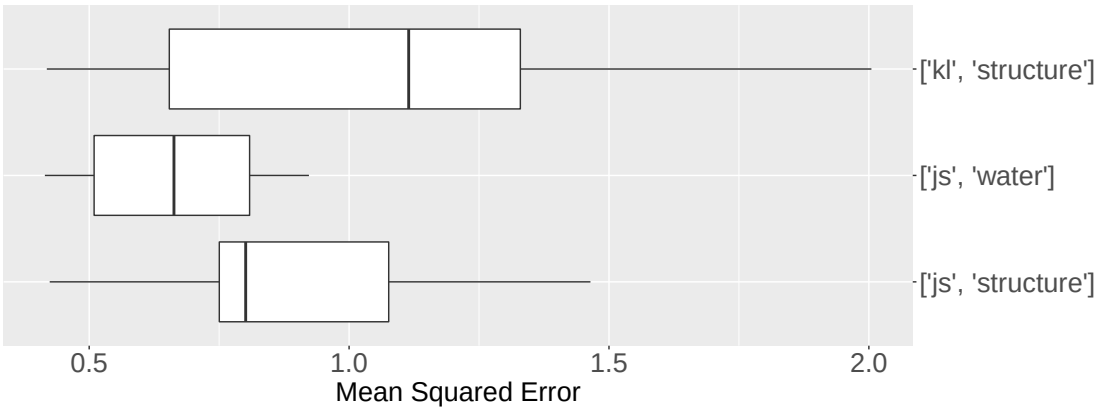
Na klasifikáciu pokročilejších dátových sád boli použité parametre uvedené v tabuľke 7.7. Použité dvojice evaluátorov vychádzajú z výsledkov experimentu zobrazených na obrázku 7.4, kde dvojice `[js, water]`, `[js, structure]` a `[kl, structure]` vykazovali naj-

lepšie výsledky bez použitia normalizácie. Kvôli komplexnosti dátových sád bol zvýšený počet krokov prehľadávania stavového priestoru na 500000.

Parameter	Hodnota
whole_DB	data/knownDB.csv
DB_method	k-mer
threshold	1
input_samples	[data/datasetC.fasta, data/datasetD.fasta]
evaluation_methods	[[js, water], [js, structure], [kl, structure]]
normalization	False
normalization_steps	0
steps	[500000]

Tabuľka 7.7: Parametre klasifikácie použité s pokročilejšími dátovými sadami popísanými v tabuľke 7.6. Normalizácia skóre je vypnutá a počet krokov prehľadávanai stavového priestoru je oproti predchádzajúcim experimentov zvýšený na 500000.

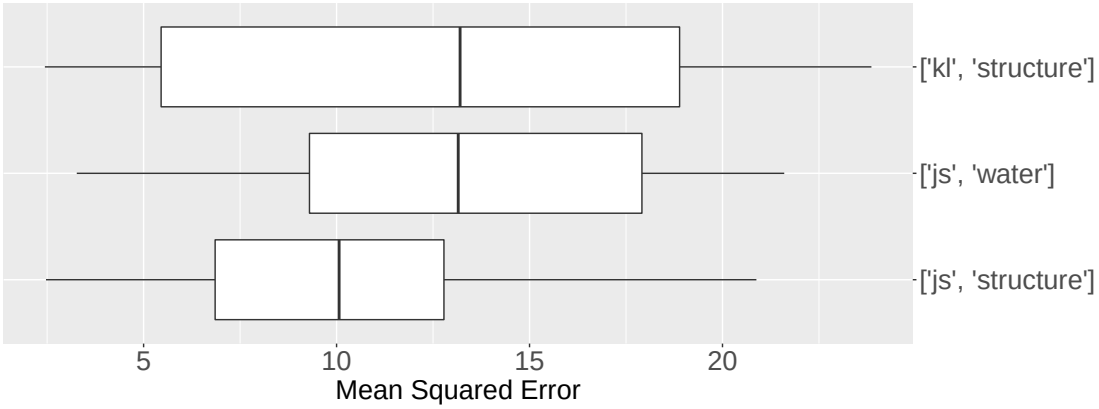
Experiment bol spustený 10-krát pre obe dátové sady a pre všetky použité dvojice evaluátorov. Porovnanej výslednej strednej kvadratickej chyby pre dátovú sadu C je na obrázku 7.5 a pre dátovú sadu D na obrázku 7.6.



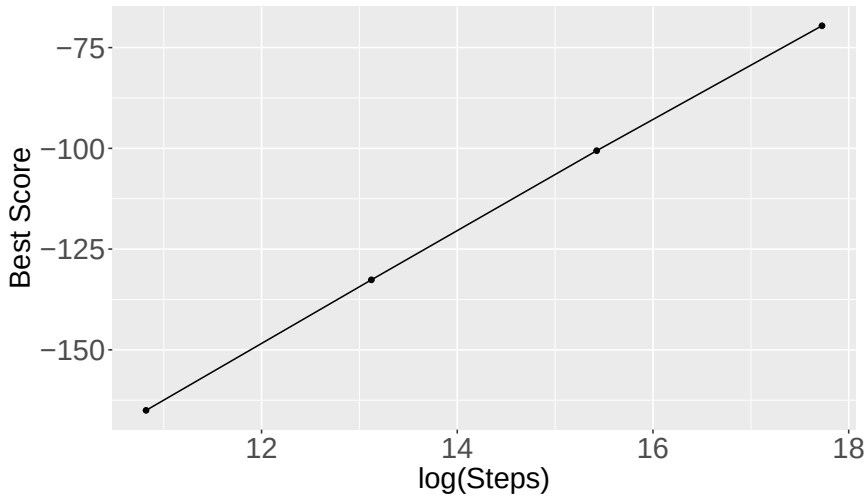
Obr. 7.5: Stredné kvadratické chyby výsledných abundancií od očakávaných abundancií baktérií. Klasifikácia bola spustená s parametrami uvedenými v tabuľke 7.7 10-krát pre každú dvojicu evaluátorov s dátovou sadou C popísanou v tabuľke 7.6.

Žiadnej dvojici evaluátorov sa nepodarilo skonvergovať k správnemu riešeniu v poskytnutom počte krokov prehľadávania stavového priestoru. Analýza získaných výsledkov a hodnôt skóre priradených výsledným stavom ukázala, že samotný nástroj považoval nájdené stavy za pomerne nepravdepodobné z pohľadu zvolených modelov.

Pri ďalšom experimente bola použitá dátová sada D, dvojica evaluátorov `[js, water]` a normalizácia skóre bola vypnutá. Experiment bol spustený s počtami krokov $5 \cdot 10^4$, $5 \cdot 10^5$, $5 \cdot 10^6$ a $5 \cdot 10^7$. Najlepšie dosiahnuté skóre ohodnotenia stavu pri jednotlivých počtoch krokov prehľadávania stavového priesoru je zobrazená na obrázku 7.7. Systém nebol schopný skonvergovať ku korektnému stavu, ale pri zvyšovaní počtu krokov prehľadávania sa zlepšuje skóre výsledných stavov.



Obr. 7.6: Stredné kvadratické chyby výsledných abundancií od očakávaných abundancií baktérií. Klasifikácia bola spustená s parametrami uvedenými v tabuľke 7.7 10-krát pre každú dvojicu evaluátorov s dátovou sadou D popísanou v tabuľke 7.6.



Obr. 7.7: Vzťah medzi počtom krokov prehľadávania stavového priestoru a najlepším dosiahnutým skóre ohodnotenia stavu. Experiment bol spustený s počtami krokov $5 \cdot 10^4$, $5 \cdot 10^5$, $5 \cdot 10^6$ a $5 \cdot 10^7$. Pre názornosť je na zobrazenie krokov použitá logaritmická stupnica. Hodnoty skóre predstavujú logaritmus pravdepodobnosti, že získaný stav odpovedá zvolenému modelu. Pri nájdení korektného stavu by sa hodnota skóre rovnala 0.

Kapitola 8

Záver

Hlavným cieľom tejto práce bolo navrhnúť a implementovať nástroj na klasifikáciu sekvencií bakteriálneho génu 16S rRNA do taxonomických kategórií s využitím vlastností 16S rRNA. Za týmto účelom bolo najskôr nutné získať rozsiahle vedomosti z oblasti molekulárnej biológie, bioinformatiky a metagenomiky. Získané znalosti sú prezentované v kapitolách 2, 3 a 4.

Získané informácie boli následne použité na špecifikáciu nástroja, ktorá je popísaná v kapitole 5. Princíp tohto nástroja je založený na vlastnosti génu 16S rRNA, ktorá hovorí, že bakteriálny genóm obsahuje niekoľko kópií génu 16S rRNA, ktoré sa môžu líšiť svojou sekvenciou. Z tohto dôvodu bol zvolený prístup, kedy sa vstupné sekvencie neklasifikujú jednotlivo, ale analyzujú sa ako celok.

Navrhovaný nástroj využíva algoritmus Metropolis Hastings na prehľadávanie priestoru možných riešení. Tento priestor je reprezentovaný dvojrozmernou maticou, kde stĺpce predstavujú vstupné sekvencie, riadky predstavujú sekvencie génu 16S rRNA známych baktérií a hodnoty v bunkách matice určujú, koľko sekvencií z daného stlpca je priradených k sekvencii v danom riadku.

Pri prehľadávaní stavového priestoru je potrebné vedieť ohodnotiť, ako pravdepodobný je aktuálny stav z pohľadu zvoleného modelu. Na ohodnotenie stavu systému bolo navrhnutých niekoľko prístupov, ktoré sa delia do dvoch skupín. Jedna skupina ohodnocuje stav z pohľadu podobnosti priradených sekvencií a druhá skupina z pohľadu dodržania pomerov abundacií variánt 16S rRNA pri jednotlivých baktériách. Jedna z navrhovaných metód ohodnotenia stavu z pohľadu podobnosti priradených sekvencií využíva podobnosť sekundárnej štruktúry 16S rRNA, pretože štruktúra je evolučne konzervovanejšia ako sekvencia a môže sa stať, že dve veľmi podobné štruktúry vznikli z pomerne odlišných sekvencií.

Navrhnutý nástroj bol úspešne implementovaný s využitím programovacích jazykov Python 3, R a Bash. Vyhodnotenie výkonnosti a analýza schopností implementovaného nástroja bola vykonaná na osobnom počítači. Základná funkcionalita bola testovaná na jednoduchých dátových sadách, kde bol porovnávaný vplyv rôznych kombinácií ohodnocovania na schopnosť skonvergovať k stavu, ktorý reprezentuje korektné baktérie a pomery ich abundancií. Ukázalo sa, že ohodnotenie stavu len z pohľadu dodržania pomerov abundacií variánt 16S rRNA pri jednotlivých baktériách nie je dostatočné na konvergenciu k správnemu výsledku.

Pokročilejšie testovanie implementovaného nástroja bolo vykonané na dvoch náhodne vygenerovaných dátových sadách. V prvej dátovej sade boli baktérie vybraté podľa uniformného rozloženia a druhá dátová sada bola vygenerovaná podľa logaritmického rozloženia, ktoré sa používa na aproximáciu pomerov abundancií baktérií v bakteriálnych spoločen-

stvách. Nástroj nebol schopný skonvergovať ku korektnému stavu, ale pri zvyšovaní počtu krokov prehľadávania stavového priestoru sa zlepšovala presnosť výsledných stavov. Počet krokov, ktorý by viedol ku korektnému riešeniu, sa však na osobnom počítači nepodarilo dosiahnuť.

V budúcom rozšírení systému by bolo najvhodnejšie umožniť paralelné prehľadávanie stavového priestoru, čo by dovolilo efektívnejšie použitie nástroja aj na klasifikáciu rozsiahlejších vstupných dátových sád. Čas výpočtu a pamäťová náročnosť by mohli byť zlepšené aj profilovaním, ktoré by odhalilo náročné časti výpočtu. V náväznosti na to by bolo potrebné nástroj vyhodnotiť na rozsiahlych dátových sadách získaných z reálnych experimentov.

Počas implementácie nástroja bolo zistených ešte niekoľko možných budúcich vylepšení. Pri hodnotení stavu systému s využitím sekundárnej štruktúry RNA sa na vyhodnotenie, či môže párovanie nastať, používa len Watson-Crickovo párovanie. Ako však bolo uvedené v kapitole 2.2, až 40 % párov báz v štruktúre RNA môže byť nekánonických, a preto je vhodné zapojiť tieto znalosti do procesu ohodnocovania podobnosti. Jednou z možností je vytvoriť štatistiku párovania báz v šablónach známych baktérií a použiť ju na ohodnotenie schopnosti vytvoriť párovanie.

Ďalšia úprava hodnotenia stavu systému s využitím sekundárnej štruktúry RNA je nutná pre testovanie na dátach z reálnych experimentov. Toto hodnotenie aktuálne očakáva na vstupe sekvencie celého génu 16S rRNA, a preto používa globálne zarovnanie na zarovnanie sekvencií so šablónami, ktoré tiež pokrývajú celý gén 16S rRNA. Sekvencie z reálnych experimentov však obvykle pokrývajú len určitý región. Z toho dôvodu by bolo vhodné použiť lokálne zarovnanie a šablóny dynamicky orezať na zarovnanú časť 16S rRNA.

Literatúra

[1] *16S Illumina Amplicon Protocol* [online]. [cit. 2020-01-13]. Dostupné z: http://www.earthmicrobiome.org/protocols-and-standards/16s/.

[2] *16S rRNA Amplicon Sequencing Offers Enhanced Metagenomic Detection* [online]. [cit. 2019-24-12]. Dostupné z: https://cdn.technologynetworks.com/TN/Resources/PDF/16S%20rRNA%20Amplicon%20Sequencing%20Offers%20Enhanced%20Metagenomic%20Detection.pdf.

[3] *Dot Bracket Notation for RNA and DNA nanostructures* [online]. [cit. 2019-12-23]. Dostupné z: https://users.cs.duke.edu/~reif/courses/molcomplectures/DNA.Modeling/DotBracketNotationForRNA&DNAnanostructures/DotBracketNotationForRNA&DNAnanostructures.pdf.

[4] *Global alignment of two sequences – Needleman-Wunsch Algorithm* [online]. [cit. 2019-31-12]. Dostupné z: https://vlab.amrita.edu/?sub=3&brch=274&sim=1431&cnt=1.

[5] *High-Speed, Multiplexed 16S Microbial Sequencing on the MiSeq® System* [online]. [cit. 2020-01-13]. Dostupné z: https://www.illumina.com/documents/products/appnotes/appnote_miseq_16S.pdf.

[6] *Metagenome Analysis* [online]. [cit. 2019-26-12]. Dostupné z: https://www.eurofinsgenomics.eu/en/eurofins-genomics/material-and-methods/metagenome-analysis/.

[7] *RNAlib-2.4.14: RNA Structure Notations* [online]. [cit. 2019-12-23]. Dostupné z: https://www.tbi.univie.ac.at/RNA/ViennaRNA/doc/html/rna_structure_notations.html.

[8] *What is FASTA format?* [online]. [cit. 2019-12-23]. Dostupné z: https://zhanglab.ccmb.med.umich.edu/FASTA/.

[9] RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic acids research*. Oxford University Press. 2018, roč. 47, D1, s. D221–D229.

[10] *16S rRNA and 16S rRNA Gene* [online]. 2019 [cit. 2019-30-12]. Dostupné z: https://help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/.

[11] *K-Nearest Neighbors* [online]. 2019 [cit. 2019-27-12]. Dostupné z: https://machine-learning-course.readthedocs.io/en/latest/content/supervised/knn.html.

[12] Abrahamsson, T. R., Jakobsson, H. E., Andersson, A. F., Björkstén, B., Engstrand, L. et al. Low diversity of the gut microbiota in infants with atopic

eczema. *Journal of allergy and clinical immunology*. Elsevier. 2012, roč. 129, č. 2, s. 434–440.

[13] Achsan, B. M. *Support Vector Machine: Classification* [online]. [cit. 2020-06-02]. Dostupné z: https://medium.com/it-paragon/grid-search-f24a73a8a0ac.

[14] Acinas, S. G., Marcelino, L. A., Klepac Ceraj, V. a Polz, M. F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *Journal of bacteriology*. Am Soc Microbiol. 2004, roč. 186, č. 9, s. 2629–2635.

[15] Aggarwal, C. C. *Data mining: the textbook*. Springer, 2015.

[16] Alberts, B., Wilson, J., Johnson, A., Hunt, T., Lewis, J. et al. *Molecular Biology of the Cell*. Garland Science. ISBN 9780815341116.

[17] Alberts, B., Bray, D., Hopkin, K., Johnson, A. D., Lewis, J. et al. *Essential cell biology*. Garland Science, 2013.

[18] Allard, G., Ryan, F. J., Jeffery, I. B. a Claesson, M. J. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC bioinformatics*. BioMed Central. 2015, roč. 16, č. 1, s. 324.

[19] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. a Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology*. Elsevier. 1990, roč. 215, č. 3, s. 403–410.

[20] Andrieu, C., De Freitas, N., Doucet, A. a Jordan, M. I. An introduction to MCMC for machine learning. *Machine learning*. Springer. 2003, roč. 50, 1-2, s. 5–43.

[21] Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. a Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science*. American Association for the Advancement of Science. 2005, roč. 307, č. 5717, s. 1915–1920.

[22] Bisen, P. *Microbes in Practice*. I.K. International Publishing House Pvt. Limited, 2014. ISBN 9789382332961.

[23] Cain, A. Taxonomy. Encyclopædia Britannica, inc. [online]. 2018, [cit. 2019-26-12]. Dostupné z: https://www.britannica.com/science/taxonomy.

[24] Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M. et al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics*. BioMed Central. 2002, roč. 3, č. 1, s. 2.

[25] Carroll, I. M., Ringel Kulka, T., Keku, T. O., Chang, Y.-H., Packey, C. D. et al. Molecular analysis of the luminal and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. *American Journal of Physiology-Heart and Circulatory Physiology*. 2011.

[26] Chang, J. Y., Antonopoulos, D. A., Kalra, A., Tonelli, A., Khalife, W. T. et al. Decreased diversity of the fecal microbiome in recurrent Clostridium difficile—associated diarrhea. *The Journal of infectious diseases*. The University of Chicago Press. 2008, roč. 197, č. 3, s. 435–438.

[27] Clarridge, J. E. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews*. Am Soc Microbiol. 2004, roč. 17, č. 4, s. 840–862.

[28] Cleaves, H. J. J. Wobble Pair. In: Amils, R., Gargaud, M., Cernicharo Quintanilla, J., Cleaves, H. J., Irvine, W. M. et al., ed. *Encyclopedia of Astrobiology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, s. 1–1. Dostupné z: https://doi.org/10.1007/978-3-642-27833-4_5248-1. ISBN 978-3-642-27833-4.

[29] Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. a Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*. Oxford University Press. 2009, roč. 38, č. 6, s. 1767–1771.

[30] Coenye, T. a Vandamme, P. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters*. Blackwell Publishing Ltd Oxford, UK. 2003, roč. 228, č. 1, s. 45–49.

[31] Cristianini, N., Shawe Taylor, J. et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[32] Dagan, I., Lee, L., Pereira, F. a Pereira, F. Similarity-based methods for word sense disambiguation. In: Association for Computational Linguistics. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. 1997, s. 56–63.

[33] Ding, Y. a Lawrence, C. E. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic acids research*. Oxford University Press. 2003, roč. 31, č. 24, s. 7280–7301.

[34] dinksved, J., Halfvarson, J., Rosenquist, M., Järnerot, G., Tysk, C. et al. Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *The ISME journal*. Nature Publishing Group. 2008, roč. 2, č. 7, s. 716.

[35] Do, C. B., Woods, D. A. a Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*. Oxford University Press. 2006, roč. 22, č. 14, s. e90–e98.

[36] Eichorst, S. A., Breznak, J. A. a Schmidt, T. M. Isolation and characterization of soil bacteria that define Terriglobus gen. nov., in the phylum Acidobacteria. *Appl. Environ. Microbiol.* Am Soc Microbiol. 2007, roč. 73, č. 8, s. 2708–2717.

[37] Federhen, S. The NCBI taxonomy database. *Nucleic acids research*. Oxford University Press. 2011, roč. 40, D1, s. D136–D143.

[38] Fix, E. a Hodges Jr, J. L. *Discriminatory analysis-nonparametric discrimination: consistency properties*. California Univ Berkeley, 1951.

[39] Foster, J. A. a Neufeld, K.-A. M. Gut–brain axis: how the microbiome influences anxiety and depression. *Trends in neurosciences*. Elsevier. 2013, roč. 36, č. 5, s. 305–312.

[40] Fox, G. E. a Woese, C. R. 5S RNA secondary structure. *Nature*. Nature Publishing Group. 1975, roč. 256, č. 5517, s. 505.

[41] Fox, G. E. a Woese, C. R. The architecture of 5S rRNA and its relation to function. *Journal of molecular evolution*. Springer. 1975, roč. 6, č. 1, s. 61–76.

[42] Gautheret, D., Konings, D. a Gutell, R. R. *A major family of motifs involving G? A mismatches in ribosomal RNA*. Elsevier, 1994.

[43] Gautheret, D., Konings, D. a Gutell, R. R. GU base pairing motifs in ribosomal RNA. *Rna*. Cold Spring Harbor Lab. 1995, roč. 1, č. 8, s. 807–814.

[44] Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V. et al. Current understanding of the human microbiome. *Nature medicine*. Nature Publishing Group. 2018, roč. 24, č. 4, s. 392.

[45] Gill, K. a Robinton, D. *Genetics & Genotyping* [online]. 2009 [cit. 2019-30-12]. Dostupné z: https://slideplayer.com/slide/9435095/.

[46] Golden, B. L., Podell, E. R., Gooding, A. R. a Cech, T. R. Crystals by design: a strategy for crystallization of a ribozyme derived from the Tetrahymena group I intron. *Journal of molecular biology*. Elsevier. 1997, roč. 270, č. 5, s. 711–723.

[47] Gotelli, N. J. a Chao, A. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. Elsevier. 2013.

[48] Gray, M. W., Sankoff, D. a Cedergren, R. J. On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. *Nucleic Acids Research*. Oxford University Press. 1984, roč. 12, č. 14, s. 5837–5852.

[49] Gutell, R. R., Larsen, N. a Woese, C. R. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiology and Molecular Biology Reviews*. Am Soc Microbiol. 1994, roč. 58, č. 1, s. 10–26.

[50] Gutell, R. R., Lee, J. C. a Cannone, J. J. The accuracy of ribosomal RNA comparative structure models. *Current opinion in structural biology*. Elsevier. 2002, roč. 12, č. 3, s. 301–310.

[51] Halder, S. a Bhattacharyya, D. RNA structure and dynamics: a base pairing perspective. *Progress in biophysics and molecular biology*. Elsevier. 2013, roč. 113, č. 2, s. 264–283.

[52] Harrison, O. *Machine Learning Basics with the K-Nearest Neighbors Algorithm* [online]. 2018 [cit. 2019-27-12]. Dostupné z: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[53] Hartl, D. a Ruvolo, M. *Genetics: Analysis of Genes and Genomes*. Jones & Bartlett Learning, 2012. ISBN 9781449635961.

[54] Hashimoto, J. G., Stevenson, B. S. a Schmidt, T. M. Rates and consequences of recombination between rRNA operons. *Journal of bacteriology*. Am Soc Microbiol. 2003, roč. 185, č. 3, s. 966–972.

[55] Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. Oxford University Press. 1970.

[56] Higgs, P. G. Compensatory neutral mutations and the evolution of RNA. *Genetica*. Springer. 1998, roč. 102, s. 91–101.

[57] Hugenholtz, P. a Tyson, G. W. Microbiology: metagenomics. *Nature*. Nature Publishing Group. 2008, roč. 455, č. 7212, s. 481.

[58] Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H. et al. Structure, function and diversity of the healthy human microbiome. *Nature*. Nature Publishing Group. 2012, roč. 486, č. 7402, s. 207.

[59] Illumina, I. *Quality Scores for Next-Generation Sequencing* [online]. [cit. 2019-24-12]. Dostupné z: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf.

[60] Jones, N. C., Pevzner, P. A. a Pevzner, P. *An introduction to bioinformatics algorithms*. MIT press, 2004.

[61] Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S. et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in microbiology*. Frontiers. 2016, roč. 7, s. 459.

[62] Kataria, A. a Singh, M. A review of data classification using k-nearest neighbour algorithm. *International Journal of Emerging Technology and Advanced Engineering*. Citeseer. 2013, roč. 3, č. 6, s. 354–360.

[63] Katz, N. *Metropolis Hastings Review* [online]. [cit. 2020-05-25]. Dostupné z: https://towardsdatascience.com/metropolis-hastings-review-2dfeb0c3d0eb.

[64] Kim, S.-H. Three-dimensional structure of transfer RNA. In: *Progress in nucleic acid research and molecular biology*. Elsevier, 1976, s. 181–216.

[65] Klappenbach, J. A., Dunbar, J. M. a Schmidt, T. M. RRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* Am Soc Microbiol. 2000, roč. 66, č. 4, s. 1328–1333.

[66] Klappenbach, J. A., Saxman, P. R., Cole, J. R. a Schmidt, T. M. Rrndb: the ribosomal RNA operon copy number database. *Nucleic acids research*. Oxford University Press. 2001, roč. 29, č. 1, s. 181–184.

[67] Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C. et al. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*. Nature Publishing Group. 2012, roč. 13, č. 1, s. 47.

[68] Kullback, S. a Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*. JSTOR. 1951, roč. 22, č. 1, s. 79–86.

[69] Kurt, W. *Kullback-Leibler Divergence Explained* [online]. [cit. 2020-06-02]. Dostupné z: https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained.

[70] Kusunoki, S. a Ezaki, T. Proposal of Mycobacterium peregrinum sp. nov., nom. rev., and Elevation of Mycobacterium chelonae subsp. abscessus (Kubica et al.) to Species Status: Mycobacterium abscessus comb. nov. *International Journal of Systematic and Evolutionary Microbiology*. Microbiology Society. 1992, roč. 42, č. 2, s. 240–245.

[71] Lee, B. *K-mer - PLoSWiki* [online]. [cit. 2020-05-09]. Dostupné z: http://compbiolwiki.plos.org/wiki/K-mer.

[72] Leinonen, R., Sugawara, H., Shumway, M. a Collaboration, I. N. S. D. The sequence read archive. *Nucleic acids research*. Oxford University Press. 2010, roč. 39, suppl_1, s. D19–D21.

[73] Leontis, N. B. a Westhof, E. Geometric nomenclature and classification of RNA base pairs. *Rna*. Cambridge University Press. 2001, roč. 7, č. 4, s. 499–512.

[74] Ley, R. E., Turnbaugh, P. J., Klein, S. a Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature*. Nature Publishing Group. 2006, roč. 444, č. 7122, s. 1022.

[75] Li, J., Wang, X., Kong, X., Zhao, K., He, S. et al. Variation patterns of the mitochondrial 16S rRNA gene with secondary structure constraints and their application to phylogeny of cyprinine fishes (Teleostei: Cypriniformes). *Molecular Phylogenetics and Evolution*. Elsevier. 2008, roč. 47, č. 2, s. 472–487.

[76] Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H. et al. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic acids research*. Oxford University Press. 2015, roč. 43, W1, s. W580–W584.

[77] Madden, T. The BLAST sequence analysis tool. In: *The NCBI Handbook [Internet]. 2nd edition*. National Center for Biotechnology Information (US), 2013.

[78] Markov, A. A. The theory of algorithms. *Trudy Matematicheskogo Instituta Imeni VA Steklova*. Russian Academy of Sciences, Steklov Mathematical Institute of Russian .... 1954, roč. 42, s. 3–375.

[79] Mears, J. A., Cannone, J. J., Stagg, S. M., Gutell, R. R., Agrawal, R. K. et al. Modeling a minimal ribosome based on comparative sequence analysis. *Journal of molecular biology*. Elsevier. 2002, roč. 321, č. 2, s. 215–234.

[80] Metropolis, N. a Ulam, S. The monte carlo method. *Journal of the American statistical association*. Taylor & Francis. 1949, roč. 44, č. 247, s. 335–341.

[81] Morgan, X. C. a Huttenhower, C. Human microbiome analysis. *PLoS computational biology*. Public Library of Science. 2012, roč. 8, č. 12, s. e1002808.

[82] Needleman, S. B. a Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*. Elsevier. 1970, roč. 48, č. 3, s. 443–453.

[83] Nussinov, R., Pieczenik, G., Griggs, J. R. a Kleitman, D. J. Algorithms for loop matchings. *SIAM Journal on Applied mathematics*. SIAM. 1978, roč. 35, č. 1, s. 68–82.

[84] Otsuka, J., Terai, G. a Nakano, T. Phylogeny of organisms investigated by the base-pair changes in the stem regions of small and large ribosomal subunit RNAs. *Journal of molecular evolution*. Springer. 1999, roč. 48, č. 2, s. 218–235.

[85] Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science*. American Association for the Advancement of Science. 1997, roč. 276, č. 5313, s. 734–740.

[86] Patel, J. B. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Molecular Diagnosis*. Springer. 2001, roč. 6, č. 4, s. 313–321.

[87] Peer, Y. Van de, Chapelle, S. a De Wachter, R. A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic acids research*. Oxford University Press. 1996, roč. 24, č. 17, s. 3381–3391.

[88] Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P. et al. Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl. Environ. Microbiol.* Am Soc Microbiol. 2010, roč. 76, č. 12, s. 3886–3897.

[89] Pereira, F., Carneiro, J., Matthiesen, R., Asch, B. van, Pinto, N. et al. Identification of species by multiplex analysis of variable-length sequences. *Nucleic acids research*. Oxford University Press. 2010, roč. 38, č. 22, s. e203–e203.

[90] Pruitt, K. D., Tatusova, T. a Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*. Oxford University Press. 2007, roč. 35, suppl_1, s. D61–D65.

[91] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. Nature Publishing Group. 2010, roč. 464, č. 7285, s. 59.

[92] Reischl, U., Emler, S., HORAK, Z., KAUSTOVA, J., KROPPENSTEDT, R. M. et al. Mycobacterium bohemicum sp. nov., a new slow-growing scotochromogenic mycobacterium. *International Journal of Systematic and Evolutionary Microbiology*. Microbiology Society. 1998, roč. 48, č. 4, s. 1349–1355.

[93] Rettner, R. *DNA: Definition, Structure & Discovery* [online]. [cit. 2019-12-23]. Dostupné z: https://www.livescience.com/37247-dna.html.

[94] Rzhetsky, A. Estimating substitution rates in ribosomal RNA genes. *Genetics*. Genetics Soc America. 1995, roč. 141, č. 2, s. 771–783.

[95] Sanschagrin, S. a Yergeau, E. Next-generation sequencing of 16S ribosomal RNA gene amplicons. *JoVE (Journal of Visualized Experiments)*. 2014, č. 90, s. e51709.

[96] Scher, J. U., Ubeda, C., Artacho, A., Attur, M., Isaac, S. et al. Decreased bacterial diversity characterizes the altered gut microbiota in patients with psoriatic arthritis, resembling dysbiosis in inflammatory bowel disease. *Arthritis & rheumatology*. Wiley Online Library. 2015, roč. 67, č. 1, s. 128–139.

[97] Sebastian, A. *Amplicon sequencing and high-throughput genotyping – Basics* [online]. 2017 [cit. 2019-28-12]. Dostupné z: http://www.sixthresearcher.com/amplicon-sequencing-and-high-throughput-genotyping-basics/.

[98] Sender, R., Fuchs, S. a Milo, R. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. *Cell*. Elsevier. 2016, roč. 164, č. 3, s. 337–340.

[99] Smit, S., Widmann, J. a Knight, R. Evolutionary rates vary among rRNA structural elements. *Nucleic acids research*. Oxford University Press. 2007, roč. 35, č. 10, s. 3339–3354.

[100] Smith, S. D. a Bond, J. E. An analysis of the secondary structure of the mitochondrial large subunit rRNA gene (16S) in spiders and its implications for phylogenetic reconstruction. *The Journal of Arachnology*. BioOne. 2003, roč. 31, č. 1, s. 44–55.

[101] Smith, T. F., Waterman, M. S. et al. Identification of common molecular subsequences. *Journal of molecular biology*. Elsevier Science. 1981, roč. 147, č. 1, s. 195–197.

[102] Springer, B., BÖTTGER, E. C., Kirschner, P. a Wallace Jr, R. J. Phylogeny of the Mycobacterium chelonae-Like Organism Based on Partial Sequencing of the 16S rRNA Gene and Proposal of Mycobacterium mucogenicum sp. nov. *International Journal of Systematic and Evolutionary Microbiology*. Microbiology Society. 1995, roč. 45, č. 2, s. 262–267.

[103] Stackebrandt, E. a GOEBEL, B. M. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International journal of systematic and evolutionary microbiology*. Microbiology Society. 1994, roč. 44, č. 4, s. 846–849.

[104] Suddath, F., Quigley, G., McPherson, A., Sneden, D., Kim, J. et al. Three-dimensional structure of yeast phenylalanine transfer RNA at 3. 0Å resolution. *Nature*. Nature Publishing Group. 1974, roč. 248, č. 5443, s. 20.

[105] Tataru, P. *RNA 2nd structure prediction based on multiple alignments* [online]. 2015 [cit. 2019-28-12]. Dostupné z: https://www.slideshare.net/PaulaTataru/abrnaalignments2011.

[106] Tillier, E. R. a Collins, R. A. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. *Molecular Biology and Evolution*. 1995, roč. 12, č. 1, s. 7.

[107] Tillier, E. R. a Collins, R. A. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics*. Genetics Soc America. 1998, roč. 148, č. 4, s. 1993–2002.

[108] Tsukuda, M., Kitahara, K. a Miyazaki, K. Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16 S rRNAs. *Scientific reports*. Nature Publishing Group. 2017, roč. 7, č. 1, s. 9993.

[109] Uzzell, T. a Corbin, K. W. Fitting discrete probability distributions to evolutionary events. *Science*. American Association for the Advancement of Science. 1971, roč. 172, č. 3988, s. 1089–1096.

[110] Valešová, N. *Bioinformatic Tool for Classification of Bacteria into Taxonomic Categories Based on the Sequence of 16S rRNA Gene*. Brno, CZ, 2019. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Dostupné z: https://www.fit.vut.cz/study/thesis/21517/.

[111] Větrovskỳ, T. a Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PloS one*. Public Library of Science. 2013, roč. 8, č. 2, s. e57923.

[112] Wang, Y., Zhang, Z. a Ramanan, N. The actinomycete Thermobispora bispora contains two distinct types of transcriptionally active 16S rRNA genes. *Journal of bacteriology*. Am Soc Microbiol. 1997, roč. 179, č. 10, s. 3270–3276.

[113] Watson, J. D., Crick, F. H. et al. Molecular structure of nucleic acids. *Nature*. 1953, roč. 171, č. 4356, s. 737–738.

[114] Westhof, E. Isostericity and tautomerism of base pairs in nucleic acids. *FEBS letters*. Wiley Online Library. 2014, roč. 588, č. 15, s. 2464–2469.

[115] Wheeler, W. C. a Honeycutt, R. Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. *Molecular Biology and Evolution*. 1988, roč. 5, č. 1, s. 90–96.

[116] Wilkins, J. S. What is systematics and what is taxonomy? *Evolving Thoughts*. 2011.

[117] Woese, C. R. a Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*. National Acad Sciences. 1977, roč. 74, č. 11, s. 5088–5090.

[118] Woese, C., Magrum, L., Gupta, R., Siegel, R., Stahl, D. et al. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic acids research*. Oxford University Press. 1980, roč. 8, č. 10, s. 2275–2294.

[119] Xiong, J. *Essential bioinformatics*. Cambridge University Press, 2006.

[120] Yang, B., Wang, Y. a Qian, P.-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*. BioMed Central. 2016, roč. 17, č. 1, s. 135.

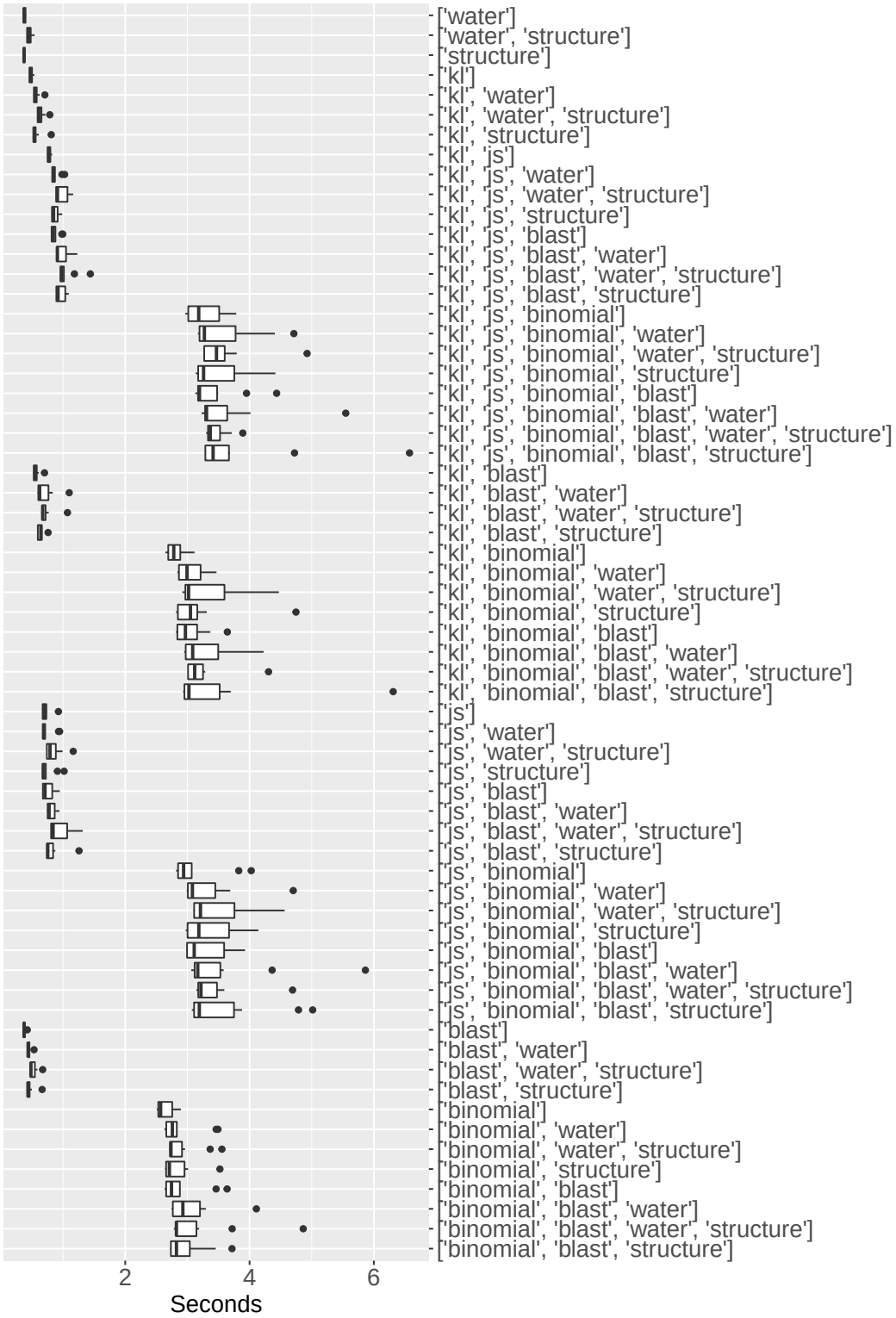
[121] Yap, W. H., Zhang, Z. a Wang, Y. Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. *Journal of bacteriology*. Am Soc Microbiol. 1999, roč. 181, č. 17, s. 5201–5209.

[122] Zuker, M. a Sankoff, D. RNA secondary structures and their prediction. *Bulletin of mathematical biology*. Springer. 1984, roč. 46, č. 4, s. 591–621.

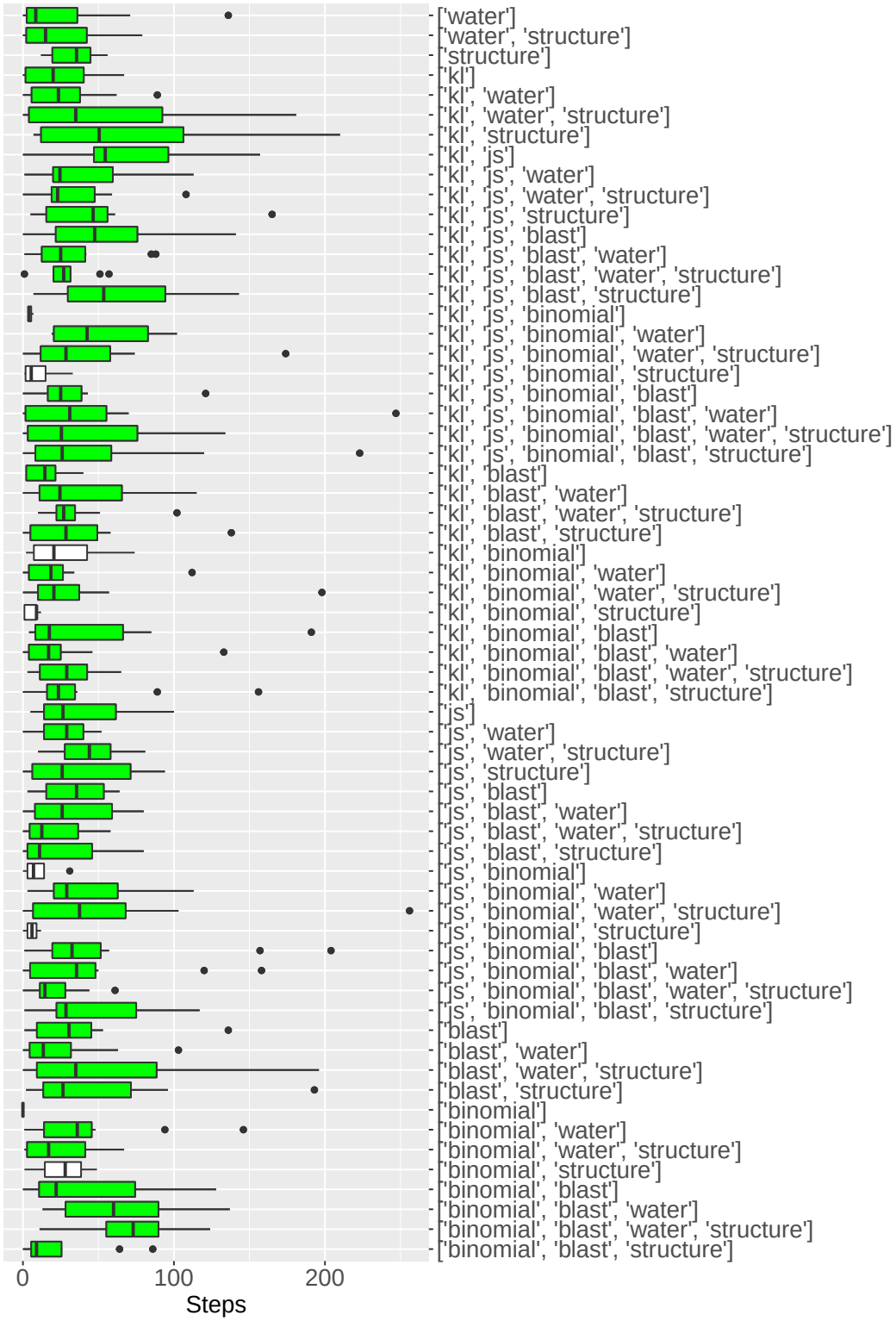
[123] Zuker, M. a Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*. Oxford University Press. 1981, roč. 9, č. 1, s. 133–148.

Príloha A

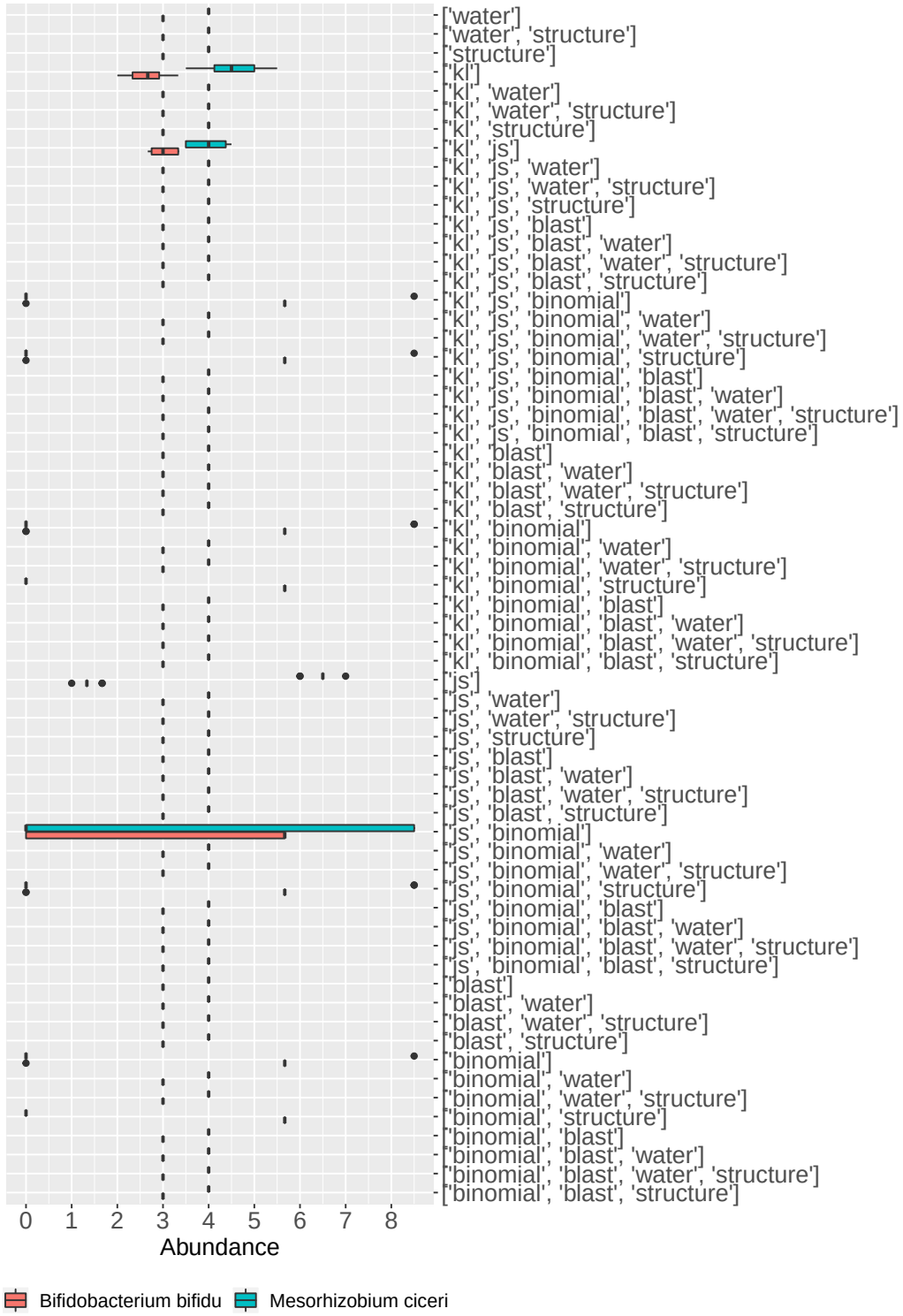
Výsledky porovnania kombinácií evaluátorov



Obr. A.1: Časy prehľadávania stavového priestoru pri 5000 krokoch a pri použití rôznych kombinácií evaluátorov. Zobrazené časy sú z 10 iterácií pri použití parametrov uvedených v tabuľke 7.4.



Obr. A.2: Krok prehľadávania stavového priestoru, v ktorom systém prvýkrát objavil a prijal stav s korektnými abundanciami baktérií. Kombinácie evaluátorov, ktoré objavili správne riešenie v každom z 10 behov, sú zobrazené zelenou farbou. Parametre klasifikácie sú uvedené v tabuľke 7.4.



Obr. A.3: Výsledné odhadnuté abundancie baktérií, ku ktorým klasifikátor skonvergoval – majú najvyššie zastúpenie vo výsledných prijatých stavoch. Zobrazené abundancie sú z 10 iterácií pri použití parametrov uvedených v tabuľke 7.4. Skutočné hodnoty abundancií sú abundancia 3 pre baktériu *Bifidobacterium bifidu* a abundancia 4 pre baktériu *Mesorhizobium ciceri*.

Príloha B

Obsah pamäťového média

- `data/` - dátové sady použité v experimentoch, vytvorené šablóny známych baktérií a podmnožina databázy známych baktérií a ich variant génu 16S rRNA
- `doc/` - zdrojové súbory tohto textu
- `src/` - zdrojové súbory implementovaného nástroja
- `definitions.py` - konfiguračné informácie pre metódy predfiltrovania databázy metódami *blast* a *k-mer*
- `DP.pdf` - elektronická verzia tohto textu
- `INSTALLATION.txt` - návod na inštaláciu potrebných závislostí
- `requirements.txt` - konfiguračný súbor so zoznamom závislostí
- `run_example.py` - skript s ukážkou použitia implementovaného nástroja