Review of Master's Thesis

Student: Valešová Nikola, Bc.
Title: Bioinformatic Tool for Classification of Bacteria into Taxonomic Categories Based on the Sequence of 16S rRNA Gene (id 21517)
Reviewer: Hon Jiří, Ing., UIFS FIT VUT

1. Assignment complexity — **more demanding assignment**

Cílem práce bylo vytvořit klasifikátor sekvencí 16S rRNA do taxonomických tříd na všech úrovních taxonomického stromu. Zadání považuji za obtížnější díky nutnosti pochopit množství biologických souvislostí.

2. Completeness of assignment requirements — **assignment fulfilled**

3. Length of technical report — **in usual extent**

4. Presentation level of technical report — **90 p. (A)**

Předložená práce je celkově napsána přehledně a kapitoly jsou uspořádány v logickém sledu. Rozsah teoretické a praktické části je vyvážený. Důležité pojmy jsou vhodně vysvětleny již při první zmínce a práce je tak dobře čitelná. Za mírně matoucí považuji vlastní redefinici metriky AUC (Area Under Curve).

5. Formal aspects of technical report — **85 p. (B)**

Práce je psána anglicky na průměrné úrovni. Překlepy se objevují minimálně, častější jsou stylisticky neobratná souvětí. Typografická stránka je kvalitní. Text je doplněn názornými obrázky, schémata a grafy. Na druhou stranu, u grafů systematicky chybí popis os. Dále bych u grafů doporučil vyhnout se duplicitním hodnotám na ose x, nahradit ji jednou hodnotou a skupinu souvisejících prvků oddělit např. odlišnými rozestupy.

6. Literature usage — **85 p. (B)**

Práce s literaturou je na dobré úrovni. Čerpáno bylo z kvalitních časopiseckých publikací z oblasti analýzy metagenomických dat. Převzaté části textu a obrázky jsou řádně označeny a odděleny od vlastního přínosu. Citační styl je mírně matoucí. Běžně je v práci citace uvedena na konci odstavce, což vede k pochybnosti, která z tvrzení v odstavci jsou vlastně citována.

7. Implementation results — **95 p. (A)**

Hlavním výstupem práce jsou zdrojové kódy v jazyce Python, které zajišťují zpracování datové sady, prohledávání prostoru možných modelů a jejich parametrů a implementují klasifikátor KTC (K-mer Tree Classifier). Dalším výstupem je množství experimentů s různými různými variantami modelů a jejich zhodnocení. Práce byla úspěšně prezentována na přehlídce Excel@FIT 2019, kde získala ocenění odborným panelem, partnery i odbornou veřejností.

8. Utilizability of results

Práce staví na publikovaných metagenomických datech a snaží se je využít ke klasifikaci do taxonomických tříd. Výsledný prediktor vykazuje nižší přesnost než některé publikované nástroje. Publikace nástroje KTC v odborném časopise by byla možná po dalším vylepšení a překonání existujících přístupů.

9. Questions for defence

- Proč je nástroj TOP přesnější než váš model? Mohlo by to souviset se ztrátou informace při použití k-merového spektra?
- U algoritmu NMDK vybíráte N prvků k-merového spektra s největšími rozdíly. Co kdyby se použili všechny prvky k-merového spektra, které by měly rozdíl větší než pevně stanovený práh? Mohlo by to vést ke zlepšení klasifikace?

10. Total assessment — **95 p. excellent (A)**

Studentka prokázala, že je schopna se orientovat v obtížnější problematice analýzy metagenomických dat. Navrhla a implementovala stromový klasifikátor sekvencí 16S rRNA do taxonomických tříd na všech úrovních taxonomie a provedla mnoho experimentů s různými variantami klasifikátoru. Přestože výsledný nástroj KTC vykazuje nižší přesnost než některé existující přístupy, četné experimenty a jejich kvalitní interpretace svědčí o výborné úrovni práce. S přihlédnutím na úspěch na přehlídce Excel@FIT navrhuji hodnocení **výborně (A)** a navrhuji komisi další ocenění práce.

In Brno 5. June 2019

Hon Jiří, Ing.
reviewer