Posudek oponenta diplomové práce

Student: Pelantová Lucie, Bc.
Téma: Klasifikace bakterií pomocí markerových genů (id 22672)
Oponent: Hon Jiří, Ing., UIFS FIT VUT

1. Náročnost zadání **obtížnější zadání**

Cílem práce bylo predikovat taxonomické zařazení neznámé bakterie podle sekvencí několika markerových genů. Zadání považuji za obtížnější kvůli nutnosti nastudovat množství biologických souvislostí.

2. Splnění požadavků zadání **zadání splněno**

3. Rozsah technické zprávy **je v obvyklém rozmezí**

4. Prezentační úroveň předložené práce **70 b. (C)**

Předložená práce je celkově členěna přehledně a kapitoly jsou uspořádány v logickém sledu. Rozsah teoretické a praktické části je vyvážený. Samotný text do značné míry předpokládá znalost dané oblasti. K prezentační úrovni mám několik připomínek. Vlastní algoritmus MultiGene by si zasloužil lépe vysvětlit - některé formulace jsou matoucí a zápisy dílčích algoritmů 1 a 2 obsahují nedefinované nebo nepoužívané proměnné. Pro pochopení algoritmu je tak nutné nahlédnutí do zdrojového kódu. Dále se v práci 9x opakuje delší text (~300 znaků) popisující význam krabicového grafu, vždy v popisu obrázku. Stačilo by ho uvést jednou.

5. Formální úprava technické zprávy **70 b. (C)**

Typografická stránka je kvalitní s výjimkou zápisu algoritmů. Text je doplněn názornými obrázky, schématy a grafy. Často se ale objevují překlepy a práce obsahuje stylisticky neobratná souvětí, která místy velmi stěžují čitelnost práce. Text obsahuje neplatné reference na obrázky.

6. Práce s literaturou **90 b. (A)**

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.

7. Realizační výstup **75 b. (C)**

Hlavním výstupem práce jsou zdrojové kódy v jazyce Python pro trénování a spuštění algoritmu MultiGene a vytvořená datová sada. Skripty pro tvorbu datové sady nejsou součástí odevzdaných zdrojových kódů. Dalším výstupem jsou experimenty s různými variantami nastavení a jejich zhodnocení. Rozsah zdrojových kódu není velký, což je pochopitelné vzhledem k experimentálnímu zaměření práce. Schází však více experimentů s různými variantami predikce. Nabízí se například natrénování jednoduchého konsenzuálního prediktoru, který by kombinoval/průměroval výstupy predikcí podle jednotlivých genů.

8. Využitelnost výsledků

Práce řeší velice aktuální téma analýzy metagenomických vzorků, které má velký potenciál pro publikaci v odborném časopise. Prezentovaný algoritmus MultiGene se novým způsobem snaží predikovat taxonomické zařazení bakterií. Aby mohl být úspěšně publikován a využit v praxi, musel by být nejdříve významně dopracován a pečlivěji ohodnocen.

9. Otázky k obhajobě

1. Ve fázi trénování zjišťujete, s jakou přesností který gen rozlišuje danou taxonomickou skupinu. Při predikci tuto znalost ale využíváte jen nepřímo, tj. zařadíte nejlepší gen do množiny genů. Dal by se váš algoritmus upravit tak, aby zohlednil i přesnost jednotlivých genů? Např. použitím vhodného váhování sekvenční identity k různým genům?
2. Velikost vaší datové sady (631 organismů) je omezena nutností znát sekvence všech osmi markerových genů. Jak velkou datovou sadu byste mohla získat, pokud byste pro predikci potřebovala pouze gen 16S rRNA?

10. Souhrnné hodnocení **75 b. dobře (C)**

Studentka prokázala, že je schopna se orientovat v obtížnější problematice analýzy metagenomických dat. Oceňuji, že vytvořila vlastní datovou sadu a ověřila vlastní algoritmus MultiGene pro predikci taxonomického zařazení pomocí více markerových genů. Celkově však mohla vyzkoušet a ohodnotit více jednoduchých způsobů predikce, které se nabízí. Vzhledem průměrnému rozsahu realizačního výstupu a průměrné prezentační úrovni práce navrhuji hodnocení stupněm **dobře (C)**.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 18. srpna 2020

Hon Jiří, Ing.
oponent