doc. Mgr. Tomáš Vinař, PhD
Katedra aplikovanej informatiky
FMFI UK, Mlynská dolina
842 48 Bratislava
tel. 02/602 95 207, e-mail. vinar@fmph.uniba.sk

OPONENTSKÝ POSUDOK NA PRÁCU MGR. ING. KARLA SEDLÁŘE "METHODS FOR COMPARATIVE ANALYSIS OF METAGENOMIC DATA"

Predložená dizertačná práca sa venuje metódam spracovania metagenomických dát. Jadro práce tvoria príspevky z dvoch oblastí: nový prístup ku vizualizácii metagenomických dát prostredníctvom bipartitných grafov a využitie prístupov zo spracovania signálu na úlohu metagenomického zatrieďovania. Oba tieto príspevky zodpovedajú odboru dizertácie ako aj obsahovému zameraniu individuálneho študijného plánu doktoranda. Jedná sa o problémy, ktoré sú v bioinformatike aktuálne a u ktorých je potrebné neustále prispôsobovať metódy analýzy vývoju sekvenačných technológií, čo je aj súčasťou autorovej práce.

Metódy prezentované v časti o vizualizácii sú výsledkom autorovej spolupráce s tímom biológov na niekoľkých metagenomických štúdiách. Oceňujem, že autor kreatívne pristúpil k úlohe analýzy týchto dát a preukázal, že dokáže reagovať a prispôsobiť použité metódy tak, aby čo najlepšie ilustrovali biologické závery odvodené z dát získaných v rámci projektu. V časti o metagenomickom zatrieďovaní sa zas autor pokúsil adaptovať metódy používané na spracovanie signálu na túto úlohu. Tento prístup je obzvlášť zaujímavý v kontexte nových metód nanopórového sekvenovania, kde prvotným zdrojom informácie je práve surový signál produkovaný zariadením a DNA sekvencie sú len sekundárnym zdrojom odvodeným z tohto surového signálu. V tomto kontexte je možné považovať prístup za novátorský.

Jadro práce bolo publikované v rámci niekoľkých časopiseckých a konferenčných publikácií, ktoré boli v poskytnutých materiáloch doložené. Autorova publikačná činnosť je bohatá a vysoko nadpriemerná na študenta doktorandského štúdia a preukazuje, že autor je schopný ako vedeckej práce vo väčších tímoch (multiautorské publikácie), tak vedenia vlastných projektov (viaceré prvoautorské publikácie).

Ku samotnej dizertačnej práci mám viacero pripomienok. Práca je písaná skôr ako predhovor ku článkom, na ktorých sa autor podieľal. Na tom by nebolo nič zlé, keby práca neobsahovala časti a dáta, ktoré zatiaľ neboli publikované a o ktorých autor píše rovnakým štýlom. Vo viacerých prípadoch sa tak nedozvedáme dôležité detaily, ktoré sú potrebné na to, aby bolo možné vyhodnotiť adekvátnosť aplikovanej metodológie. Takýchto nedostatkov je v práci veľké množstvo, napríklad sa jedná o spôsob získania a základné charakteristiky dát v kapitole 4.2 (u vlastného metagenómu sa nedozvedáme ani len to, ktorou technológiou boli dáta sekvenované, ani žiadne detaily analýzy; u umelých dát nie je jasné, akým spôsobom a z akých dát boli zostavené), či o dokumentáciu postupov použitých

pri aplikácii metód strojového učenia v časti 5.2 (napr. akým spôsobom boli dáta delené na trénovaciu a testovaciu množinu, na akých dátach bolo vykonávané ladenie hyperparametrov a ktoré hyperparametre to boli). Ďalším systémovým problémom dizertačnej práce je, že prezentované metódy boli zvyčajne vyvíjané účelovo pre konkrétny príklad a nie je jasné ani preukázané, či by ich aplikácia na iné dáta viedla k zmysluplným výsledkom. **Z hľadiska repoducibility výsledkov a preukázania generalizácie vyvíjaných metód preto prácu považujem za nevyhovujúcu.**

Ako podnet do diskusie, by som rád položil nasledujúce otázky.

1. V rámci sekcie 4.2 sa venujete zväčša technickým parametrom vašej metódy vizualizácie, t.j. popisujete, akým spôsobom zobraziť konkrétne črty spracúvaných dát. Je však možné na základe vašej vizualizácie ilustrovať konkrétne biologické závery? Vedeli by ste na reálnych dátach ukázať vašu vizualizáciu a vizualizáciu tých istých dát pomocou iných bežne používaných metód (napr. metódy PCoA) a ilustrovať, že vaša vizualizácia demonštruje konkrétny biologický záver lepšie ako druhá metóda?

2. V článku [238] popisovanom v časti 5.2.4 ste dáta získali umelým zmiešaním čítaní z dvoch nezávislých sekvenačných behov. Viete povedať, že výsledky v tomto prípade nie sú ovplyvnené tzv. batch efektami – t.j. že to čo ste sa naučili identifikovať v skutočnosti nie sú jednoducho rozdiely, ktoré sú dôsledkom napr. odlišného postupu pri príprave knižníc? Keby ste robili tieto experimenty, vedeli by ste ich navrhnúť tak, aby ste minimalizovali vplyv batch efektov?

Predložená dizertačná práca predstavuje množstvo práce autora, ktorá je súčasťou mnohých recenzovaných publikácií, na ktorých sa autor podieľal. Samotná práca je však napísaná neporiadne a pre nedostatok detailov nie je možné posúdiť reproducibilitu prezentovaných výsledkov, či možnosť generalizácie predkladaných metód na nové dáta. Ako čitateľ by som radšej uvítal užšie zameranú prácu, ktorá je spracovaná s podstatne väčšou precíznosťou. Je samozrejme na posúdenie komisie, či poukazované nedostatky sú natoľko závažné, aby bolo prácu potrebné prepracovať. Na druhej strane oceňujem autorovu vysokú publikačnú aktivitu, ktorá podľa môjho názoru preukazuje autorove schopnosti pôsobiť ako vedecký pracovník, **preto považujem predloženú prácu za vyhovujúcu a zodpovedajúcu všeobecným požiadavkám pre udelenie akademického titulu PhD.**

V Bratislave, 06.11.2018 doc. Mgr. Tomáš Vinař, PhD.