



UNIWERSYTET WARSZAWSKI

Instytut Informatyki
Uniwersytet Warszawski
ul. Banacha 2
02-097 Warszawa
POLSKA

dr hab. Bartosz Wilczyński
profesor uczelni
Phone: +(48 22) 5544 577
Fax: +(48 22) 5544 400
e-mail: bartek@mimuw.edu.pl

Warszawa, 29. czerwca 2023 r.

Recenzja rozprawy doktorskiej pt. „Advanced data exploration techniques for augmented transcriptional landscape and its better quantification”
przedstawionej przez mgr inż. Agatę Muszyńską

Recenzja niniejsza została sporządzona na zlecenie Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej zgodnie z wymogami ustawy dotyczącej procedur nadawania stopnia doktora.

Opis rozprawy

Przedstawiona rozprawa napisana jest w języku angielskim, składa się z 5 rozdziałów (Introduction, Background, Methods, Results, Summary) i liczy wraz z dodatkami 149 stron. Naturalnie, większość materiału istotnego dla oceny rozprawy znajduje się w rozdziałach trzecim i czwartym opisujących odpowiednio metodologię i wyniki rozprawy.

Dwa pierwsze rozdziały rozprawy wprowadzają do tematu i zarysowują tło przedstawionych badań. W szczególności rozdział 1 zawiera kilka fragmentów, które opisują zakres i cele rozprawy doktorskiej, a także przedstawia te aspekty rozprawy, które autorka uważa za nowatorski wkład naukowy. W szczególności, podrozdział 1.1 zawiera opis celu rozprawy, który autorka widzi jako przede wszystkim przeprowadzenie „stress-testów” obecnie stosowanych narzędzi do analizy transkryptomów i przedstawienie „rozwiązania” w postaci zestawu skryptów snakemake, które pozwalają na uruchomienie zestawu standardowych narzędzi w kilku różnych scenariuszach typowych dla analiz RNA-Seq (analiza różnicowej ekspresji genów, wykrywanie alternatywnego splicingu itp.) a także (w mniejszym zakresie) analiz danych mikromacierzowych. Podrozdział 1.2 zaś poświęcony jest nowatorskiemu wkładowi rozprawy, który, w ocenie autorki, polega na stworzeniu nowego zestawu skryptów snakemake, będących pewnym ulepszeniem obecnie dostępnych zestawów narzędzi do analizy RNA-Seq.

Rozdział trzeci zawiera opis metodologii wykorzystywanych w rozprawie. Obejmuje on zarówno podstawy technologii sekwencjonowania RNA/cDNA, podstawy technologii mikromacierzowych, jak i wszelkie techniki analizy danych, zarówno od strony statystycznej jak i

ogy).

Uwagi krytyczne

Tytuł rozprawy sugeruje, że autorka planowała stworzenie nowych „zaawansowanych” metod eksploracji danych, które miały prowadzić do „lepszich” metod kwantyfikacji transkryptomów. W istocie, po przeczytaniu rozprawy widać jest, że aż tak ambitne cele nie zostały osiągnięte. Szkoda, że nie udało się zmienić tytułu, aby bardziej odpowiadał zakresowi rozprawy. Podejrzewam, że być może taka zmiana byłaby kłopotliwa formalnie, ale myślę, że przy tak istotnym okrojeniu zakresu pracy, zmiana tytułu pozwoliłaby uniknąć efektu zbyt dużych oczekiwań wobec pracy, które taki ambitny tytuł wywołuje.

Praca napisana jest w języku angielskim i w niektórych miejscach widać jest, że nie jest to język ojczysty autorki. (np. sformułowania „confounding effects” zamiast „factors”, „faster run time” zamiast „shorter...” itp. są używane niepoprawnie), ale zapewne wykorzystanie języka angielskiego podyktowane było udziałem prof. Kreila jako ko-promotora. Mimo tych, drobnych, niedociągnięć, praca daje się czytać dość dobrze i nie odbiega od poziomu typowego dla rozpraw doktorskich broniących w Polsce.

Podział treści pomiędzy rozdziały budzi niekiedy wątpliwości. Np. transformata Burrows’a Wheelera opisana jest w podrozdziale 2.3 „NGS technologies summary”. Wydaje się, że lepszym miejscem dla takiego opisu byłby rozdział 3, jako że nie jest to metoda przynależna do technik NGS, a raczej struktura danych pomocna m.in. w tej dziedzinie. Metoda ta jest potraktowana dość pobieżnie, co, być może, nie ma wpływu bezpośrednio na wyniki rozprawy, ale nie najlepiej świadczy o wiedzy kandydatki w dyscyplinie informatyka.

Podobnie, wykorzystanie rozkładu ujemnego dwumianowego (NB) w modelowaniu różnicowej ekspresji genów przedstawione jest w rozdziale wstępnym (2.4) i potraktowanie bardzo skrótowo. Autorka nie ustrzegła się przy tym sformułowań ocierających się o błąd, pisząc, że „Random sampling of RNA-Seq reads [...] can be modelled quite well by Poisson distribution” co nie jest prawdą, z czego autorka zapewne zdaje sobie sprawę pisząc dalej o wykorzystaniu rozkładu NB do tego celu.

Opisując pakiet snakemake w podrozdziale 3.1, autorka używa sformułowania „comes with a set of additional advantages”, co jest dość zaskakujące, bo nie dowiadujemy się jakie to zalety, a mowa nie o jakimś marginalnym narzędziu, a głównym, wybranym przez autorkę, rozwiązaniu.

Opisując metodę TMM autorka używa określenia „trimmed mean of M-values values”, co jest zapewne po prostu błędem pisarskim.

W podrozdziale 3.5.5, autorka wymienia metody poprawek związanych z testowaniem wielu hipotez, ale wspomina wyłącznie o metodzie Holma (zwanej zwykle Holma-Bonferroniego) i o podejściu FDR, pomijając prostszą i często stosowaną metodę Bonferroniego.

uważam, że – mimo wspomnianych słabości - **rozprawa wypełnia ustawowe wymagania stawiane wobec rozpraw doktorskich** i może zostać skierowana do kolejnych etapów postępowania w kierunku nadania stopnia doktora.

Bartosz Wilczyński