Dr hab. inż. Paweł Piotr Łabaj
Małopolskie Centrum Biotechnologii
Uniwersytet Jagielloński
Gronostajowa 7a, 30-387 Kraków
pawel.labaj@uj.edu.pl

Kraków, 06.10.2021

Recenzja rozprawy doktorskiej

Tytuł rozprawy:	Analiza różnicowej ekspresji genów w populacjach bakterii
Autorka rozprawy:	mgr Julia Herman Iżycka
Promotor rozprawy:	Dr hab. Bartosz Wilczyński
Dziedzina:	nauki ścisłe i przyrodnicze
Dyscyplina:	Informatyka
Zakres dyscypliny podstawowej:	Metody analizy danych bioinformatycznych ze szczególnym uwzględnieniem danych z sekwencjonowania DNA

Ostatnie lata w naukach biologicznych i medycznych cechują się szybkim rozwojem wysokoprzepustowych technik pomiarowych. Wzrost ilości mierzonych, przechowywanych i przetwarzanych danych dotyczy nie tylko objętości, ale też rozdzielczości i złożoności. Jednym z obszarów, gdzie te techniki pomiarowe znalazły swoje zastosowanie są badania nad składem mikrobiomów w różnych niszach ekologicznych. Dla zdrowia i dobrostanu ludzi interesujące są zwłaszcza te wchodzące w skład naszego ekspozomu, czy to w postaci mikrobiomu otaczającego środowiska, mikrobiomu skóry, ust czy jelit. Techniki wysokoprzepustowego sekwencjonowania umożliwiły nie tylko jakościowe, ale też ilościowe badanie składu mikrobiomów (metagenomika) co wcześniej było możliwe jedynie w ograniczonym zakresie. Dodatkowo możliwe stało się badanie nie tylko tego jakie organizmy występują, ale również analiza które ich geny są aktywne (metatranskryptomika) aby zrozumieć stan funkcjonalny całej społeczności mikrobów, ich zależności oraz interakcji z otoczeniem. Należy pamiętać jednak, że te nowe możliwości poznawcze są powiązane ze wzrostem złożoności badanego systemu (pojedynczy organizm -> metagenom -> metatranskryptom) a w związku z tym z nowymi wyzwaniami analitycznymi.

W rozprawie szeroko omówione zostały zagadnienia badania wyników sekwencjonowania NGS danych metatranskryptomicznych. We współpracy z dr Iloną Grabowicz zaproponowana została wieloetapowa procedura używająca dostępnych narzędzi obejmująca asemblacje, mapowanie, wybór genów i ich anotację funkcjonalną, oraz analizę różnicowej ekspresji. W

dalszej części sformułowano również problem asemblacji różnicowych kontigów a następnie opisano i zanalizowano możliwość użycia do jego rozwiązania dwóch rodzajów grafów używanych do asemblacji – grafu de Bruijna i grafu nałożeń. Wykazano, że graf de Bruijna jest raczej niemożliwy do wykorzystania w przypadku asemblacji różnicowych sekwencji, natomiast w przypadku grafu nałożeń potrzebne są podejścia heurystyczne. Zaproponowany w pracy sposób wykorzystania SGA do budowy grafu nałożeń tak, aby możliwe było śledzenie liczby odczytów tworzących wierzchołki, a także pokazane heurystyki wyszukiwania różnicowych kontigów w tak przygotowanym grafie nałozen, pozwoliły zaproponować zbiory sekwencji różnicowych, które są różne od sekwencji uzyskanych innymi metodami. Niestety nie jest możliwe jednoznaczna ocena jakości wyników, ponieważ nie istnieją ogólnie przyjęte metryki do takiej oceny. Metatranskryptomika bazująca na wynikach sekwencjonowania wysokoprzepustowego jest jeszcze zbyt młodą dziedziną i dalsze badania, zwłaszcza takie, gdzie jest zdefiniowana jakaś forma „prawdy podstawowej" są konieczne.

Z tego też względu należą się słowa uznania dla Doktorantki, która podjęła bardzo odważny temat o wysokim stopniu interdyscyplinarności. Rzadko się zdarza tak solidne przygotowanie zarówno od strony matematyczno-informatycznej jak i biologicznej. Doktorantka musiała posiąść bardzo szeroką wiedzę z wielu zakamarków bioinformatyki aby być w stanie podejść do tematu w tak kompleksowy sposób. Dlatego należy bardzo wysoko ocenić wkład Autorki we wszechstronne rozpoznanie problemu.

Układ rozprawy jest typowy dla tego typu opracowań. Wprowadzenie i wstęp szeroko omawiają obecny stan wiedzy odnośnie: i) podstawowych pojęć z genetyki, ii) mikroorganizmów i ich znaczenia, oraz iii) charakterystyki technologii do badań metagenomicznych i metatranskryptomicznych. A w następnej części Doktorantka dokładnie omawia obliczeniowe metody analizy wyników eksperymentów NGS. Następnie przedstawiona jest analiza z wykorzystaniem obecnie istniejących narzędzi. Autorka poprawnie zauważa, że narzędzia te w większości nie są zaprojektowane do analizy metatranskrytpomicznej a raczej metagenomicznej, ale przy odpowiednim użyciu spełniają stawiane wymagania. Ta część podsumowana jest dwoma bardzo istotnymi obserwacjami:

- *„Odejście od patrzenia na mikrobiom jako na zbiór obecnych gatunków w stronę spojrzenia na ogół funkcjonalnych możliwości mikrobiomu może pozwolić lepiej zrozumieć sposób, w jaki całość mikrobiomu oraz organizm gospodarza oddziałują na siebie wzajemnie."*
- Użycie istniejących narzędzi powoduje, że z jednej strony niektóre kroki są niepotrzebnie powtarzane, a z drugiej akumulują się ich ograniczenia, np. wykorzystanie niekompletnych baz referencyjnych

Pierwsza obserwacja jest zgodna z obecnym trendem analizy danych mikrobiomowych, gdzie należy zauważyć, że skład mikrobiomu jest zależny od środowiska bytowania, tak więc to dana nisza ekologiczna definiuje poniekąd zestaw wymaganych funkcji, aby społeczność mikrobów mogła w niej bytować. Kwestia tego, który organizm dostarczy, którą funkcję jest do pewnego stopnia drugorzędna. Należy więc traktować taką społeczność jako jeden złożony organizm, zwłaszcza przy analizie funkcjonalnej.

Natomiast druga obserwacja pchnęła Doktorantkę do próby stworzenia nowego narzędzia, które bezpośrednio będzie dokonywało asemblacji z identyfikacją różnicowych sekwencji. Opis

tego problemu, jego rozwiązanie i użycie na przykładowych danych są przedstawione w kolejnej części. Niestety moja ekspertyza nie pozwala na dogłębną ocenę poprawności tego rozwiązania. Jednakże przedstawione rozumowanie jak i wyniki wydają się być przekonujące.

Jest jednak jedna dość istotna sprawa, która została niejako pominięta zarówno w podejściu z wykorzystaniem istniejących narzędzi jak i zbudowania nowego podejścia. W przypadku szukania fragmentów/genów różnicujących bada się czy zmiana poziomu ekspresji jest istotna statystycznie lub „znacząca" w inny sposób. Aby móc to poprawnie ocenić należy wykonać normalizację częstości występowania pomiędzy próbkami. W przypadku transkryptomiki koncepcyjnie nie jest to zadanie trudne, ponieważ zakłada się, że większość genów nie zmienia swojego poziomu ekspresji nawet przy mocnych zaburzeniach systemu (np. choroba nowotworowa). W przypadku metatranskryptomiki nie mamy podstaw, aby twierdzić, że to założenie jest spełnione:

- już na poziomie metagenomicznym normalizacja w klasyczny sposób nie ma racji bytu i często wybiera się jako referencyjny organizm, który nie jest związany z badanym zjawiskiem
- tutaj dochodzi nam dodatkowy poziom złożoności, bo chcemy badać aktywność genów nie w jednym, ale jednocześnie w wielu organizmach, dla których kompozycja pomiędzy próbkami nie jest znormalizowana a jednocześnie te same funkcjonalności (np. metabolity) mogą być dostarczone przez różne organizmy

Doktorantka w części, gdzie wykorzystuje istniejące narzędzia używa DESeq2, który dokonuje normalizacji między próbkami, ale bazuje on na założeniach prawdziwych dla transkryptomiki. Może się zdarzyć, że w analizie danych metatranskryptomicznych te założenia też będą spełnione. Ale tak naprawdę tego nie wiemy, a sprawa ta nie została omówiona w pracy. W przypadku wykorzystania grafu nałożeń nie jestem w stanie ocenić, czy to zagadnienie jest tam rozwiązane, wg mojej wiedzy jest to raczej trudne do rozwiązania, ale jeśli sytuacja tutaj jest inna to nie jest to zaznaczone i objaśnione w tekście rozprawy, a jest to sprawa naprawdę kluczowa. Decyduje ona bowiem o fakcie czy narzędzie ma praktyczne zastosowanie. Sugerowałbym zapoznanie się z następującymi publikacjami:

- https://doi.org/10.1093/bib/bbx104
- https://doi.org/10.1101/2020.10.01.322164

Całościowo należy zauważyć, że rozprawa jest napisana bardzo starannie i na tyle zrozumiałym językiem na ile złożoność zagadnienia pozwala. Tezy rozprawy są sformułowane jasno i przystępnie oraz są w pełni poparte danymi zawartymi w poszczególnych rozdziałach rozprawy. Podsumowanie rozprawy jest syntetycznym wykazaniem, że założone w pracy cele zostały osiągnięte z uwzględnieniem kluczowych wyników.

Rozprawa zawiera 26 szczegółowych rycin oraz 18 tabel, które co do zasady są jasne i czytelne. Zawarte w stopkach i nagłówkach opisy są sporządzone czytelnie i umożliwiają czytelnikowi szybką orientację w treści rycin i tabel. Piśmiennictwo obejmuje 100 dobrze dobranych i aktualnych pozycji, choć w tym szybko rozwijającym się temacie niemożliwe jest nadążenie i zawsze coś może umknąć.

Dotychczasowy dorobek Autorki nie jest związany z zagadnieniem opisywanym w rozprawie i ponieważ nie został w niej wykazany nie podlega ocenie. Zgodnie z deklaracją Autorki w przygotowaniu są publikacje dotyczące zagadnień opisanych w rozprawie:

- Metatranscriptomic changes upon high-fat diet in a Down syndrome mouse model autorstwa: Ilona E. Grabowicz, Julia Herman-Izycka, Marta Fructuoso, Mara Dierssen, Bartek Wilczynski
- *application note* opisujący heurystyki asemblacji różnicowych kontigów w grafie nałożeń. https://github.com/juliahi/diffcog

Ponadto pod adresem https://github.com/juliahi/kallisto dostępna jest zmodyfikowana wersja *kallisto* pozwalająca uzyskać informacje o mapowaniu bądź powodach niemapowania każdego analizowanego odczytu.

Podsumowując, pomimo niedostatecznego uwzględnienia kwestii normalizacji pomiędzy próbkami, która niestety tutaj jest kluczowa, przedstawiona do oceny praca doktorska stanowi bardzo ciekawe i wartościowe rozwiązanie ważnego zagadnienia naukowego. Jednakże może wymagać dalszej pracy, aby uwzględnić wykazane braki. Rozprawa jest ważnym przyczynkiem w zakresie wiedzy na temat analizy danych metatranskryptomicznych. Praca ta w pełni odpowiada warunkom stawianym rozprawom doktorskim oraz rozpoczyna dyskusję w kwestii wypełnienia istotnej luki, a mianowicie w zakresie dostępności narzędzi do analizy danych metatranskryptomicznych, a także podnosi kwestię braku powszechnie przyjętego sposobu oceny jakości tychże narzędzi.

Na podstawie powyższej oceny wnioskuję zatem do Rady Dyscyplin Naukowych Matematyki i Informatyki Uniwersytetu Warszawskiego o dopuszczenie Doktorantki do dalszych etapów przewodu doktorskiego. Nie mam wątpliwości, że doświadczenia zgromadzone przez Autorkę przy analizie danych oraz rozwijaniu nowego narzędzia stawia cały zespół badawczy w doskonałej pozycji wśród międzynarodowych grup zajmujących się tą tematyką.

Elektronicznie
podpisany przez
Paweł Łabaj
Data: 2021.10.06
14:01:57 +02'00'

Dr hab. inż. Paweł Piotr Łabaj