Warszawa, 16 sierpnia 2021 r.

dr hab. inż. Robert Nowak, prof. uczelni
Instytut Informatyki
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska
ul. Nowowiejska 15/19
00-665 Warszawa

Recenzja rozprawy doktorskiej mgr Julii Herman-Iżyckiej zatytułowanej „Analiza różnicowej ekspresji genów w populacjach bakterii"

Recenzja powstała na prośbę Rady Naukowej Dyscyplin Matematyka i Informatyka Uniwersytetu Warszawskiego na podstawie: (1) rozprawy doktorskiej liczącej 108 stron z maja 2021 r, (2) autoreferatu, (3) dorobku naukowego Kandydatki uwzględnionego w bazach Scoupus i Google Scholar, (4) kodów źródłowych prezentowanych rozwiązań dostępnych w repozytorium GitHub.

1 Tematyka badań

Rozprawa dotyczy metod analizy danych metatranskryptomicznych (sekwencji RNA) uzyskiwanych przez sekwencjonowanie nowej generacji metodą Illumina. Tytuł rozprawy odpowiada jej treści. Mgr Julia Herman-Iżycka zaproponowała nowe metody analizy tego typu danych pozwalających badać aktywność mikroorganizmów (głównie bakterii) zamieszkujących daną niszę środowiskową. Metody te wykorzystano dla nowych, rzeczywistych danych. Rozprawa definiuje problem asemblacji różnicowych kontigów, pokazuje nowe algorytmy oraz ich realizacje w języku Python.

Cel badawczy jest postawiony właściwie, jest on interesujący i istotny. Przedstawione rozwiązania są poprawne i potwierdzone eksperymentem.

2 Główne wyniki rozprawy

W rozprawie w sposób właściwy przeprowadzono analizę źródeł z literatury światowej, co świadczy o dostatecznej wiedzy Autorki. Wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonujący, pozwoliły one postawić opisane w pracy problemy badawcze.

Mgr Julia Herman-Iżycka użyła właściwej metody do rozwiązania postawionych problemów.

W rozprawie pokazano rozwiązanie dwóch spójnych tematycznie problemów. Pierwszym z nich, opisanym w rozdziale 3, jest uzyskanie ilościowej oceny zmian w mikrobiomie jelitowym myszy na różnych dietach. Istniejące metody pozwalały wykorzystać

tylko niewielką część sekwencji. Zespół, którego członkiem jest Autorka rozprawy zaproponował badanie różnic ekspresji genów, wykorzystanie bazy genomów referencyjnych i nowe algorytmy wykorzystujące m.in. asemblację. Pozwoliło to w ciekawy sposób opisać w jaki sposób mikroorganizmy żyjące w przewodzie pokarmowym myszy oraz organizm gospodarza oddziałują na siebie.

Drugim problemem, opisanym w rozdziale 4, jest uzyskanie metody, która wykorzystując odczyty z sekwenatorów jednocześnie tworzy sekwencję wynikową (problem asemblacji) i identyfikuje sekwencje, które różnicują dwie grupy próbek (problem identyfikacji). Zagadnienie takie występuje m.in. przy analizach odczytów z mikrobiomów. Kandydatka przeprowadziła rozważania teoretyczne dla struktur danych wykorzystywanych w asemblacji: pod-grafów grafów de Bruijna oraz dla grafów nałożeń, zaproponowała nowy algorytm i go zrealizowała jako moduł programu String Graph Assembler (SGA). Porównała wyniki z innymi rozwiązaniami i wyprowadziła właściwe wnioski.

Uzyskane wyniki są pokazane poprawnie. Wywód jest zwięzły i jasny, drobne nieścisłości przedstawiłem w sekcji 4. Nie stwierdzam znaczących błędów redakcyjnych.

Rozprawę doktorską stanowi samodzielna i wyodrębniona część pracy zbiorowej, opisanej w rozdziale 3 i w rozdziale 4 rozprawy. Recenzja musi zawierać ocenę indywidualnego wkładu kandydata w powstanie tej pracy (paragraf 5 pkt. 5 rozporządzenia MNiSW z dn. 26 września 2016 r), dlatego proszę o uzupełnienie dokumentacji, aby jednoznacznie wskazać wkład mgr Julii Herman-Iżyckiej. Na podstawie rozprawy nie mogę stwierdzić, co stanowi samodzielny i oryginalny dorobek Autorki.

3 Propozycja uzupełnienia rozprawy

Proponowane uzupełnienia rozprawy są następujące:

1. szerszy opis własnych dokonań,
2. wyraźne określenie indywidualnego wkładu.

3.1 Szerszy opis własnych rozwiązań

Rozprawa liczy 108 stron, z których większość to wprowadzenie lub opis istniejących rozwiązań. Wprowadzenie i opis metod NGS zajmuje pierwsze 61 stron (rozdziały 1 i 2). Rozdział 3, zajmujący 19 stron, to opis eksperymentu i metod w nim użytych, gdzie tylko niektóre są nowe (własne). Rozdział 4 (17 stron) w całości opisuje nowe (własne) ozwiązanie. Wydaje mi się, że opis własnych prac stanowi zbyt małą część rozprawy.

Nie znalazłem publikacji omawiających przedstawione w rozprawie rozwiązania. Autorka w autoreferacie na str. 10 wspomina, że publikacje takie są w przygotowaniu. Proponuję opublikować wyniki badań w czasopismach naukowych lub w materiałach konferencyjnych i cytować je w rozprawie.

Proponuję więcej uwagi poświęcić na opis własnych rozwiązań. Na przykład, jeżeli opracowano własny algorytm w ramach prac opisanych w rozdziale 3 i w rozdziale 4, to

R

myślę, że można przedstawić użycie tych algorytmów na wielu różnych zbiorach danych, na danych rzeczywistych i danych symulowanych. Myślę, że nie zaszkodzi analiza złożoności obliczeniowej poparta eksperymentami numerycznymi, czy analiza skalowalności.

Przy dostarczaniu kodów źródłowych warto zadbać o czytelność kodu. Przydałaby się chociaż minimalna dokumentacja, w postaci komentarzy w kodzie, gdzie opisane będą odpowiedzialności pakietów, modułów czy funkcji. Warto dodać testy jednostkowe.

3.2 Określenie indywidualnego wkładu w badania

Badania prezentowane w rozprawie były prowadzone w zespołach naukowych, co jest moim zdaniem wielką ich zaletą i pokazują umiejętność Kandydatki do pracy w interdyscyplinarnych zespołach. Aby poprawnie ocenić rozprawę chciałbym znać indywidualny wkład mgr Herman-Iżyckiej w badania. W rozdziale 3, przy opisie metody barania różnicowej ekspresji, pojawiła się jedynie informacja, że procedurę analizy zaproponowała dr Ilona Grabowicz (str 74 rozprawy). Niektóre analizy i eksperymenty numeryczne również wyknała dr Grabowicz, patrz przypisy na str. 74, 76 i 77. Proszę więc szczegółowo opisać, na czym polegał wkład Kandydatki. Opis metody asemblacji z jednoczesną identyfikacją nie zawiera sugestii, że autorem koncepcji jest ktoś inny niż autor tekstu, ale niestety nie jest napisane wprost, co zrobiła Autorka. Być może pomogłoby przekształcenie niektórych zdań ze strony biernej na czynną (opracowałam, zdefiniowałam, itd.). Pliki źródłowe utworzonego przez mgr Herman-Iżycką modułu dla potoku SGA, dostępne w repozytorium GitHub, nie mają autora w komentarzach, proponuję to uzupełnić. Kandydatka jest autorem wszystkich zmian w repozytorium, więc tutaj nie mam wątpliwości.

Konieczność określenia indywidualnego wkładu wynika z paragrafu 5, punkt 2 rozporządzenia MNiSW, definiującego proces uzyskania stopnia naukowego doktora. Mowa jest tam o precyzyjnym określeniu wkładu i o dostarczeniu oświadczeń współautorów.

4 Drobne poprawki i uzupełnienia

Zagadnienie badawcze uważam za ciekawe, przedstawione rozwiązania uważam za bardzo dobre. Poniżej jest kilka uwag, które proponuję uwzględnić przy okazji wprowadzania poprawek:

- algorytm Smitha-Watermana różni się od algorytmu Needlemana-Wunscha nie tylko warunkami brzegowymi (tak jak napisano w rozprawie na str 52), ale także regułą wypełniania poszczególnych komórek, algorytm ten nie dopuszcza wartości ujemnych w macierzy;

- Przy opisie alg. Needlemana-Wunscha pojawia się zdanie '... możliwe jest rozwiązanie korzystające z programowania dynamicznego, takie jak alg. Needlemana-Wunscha ... korzystając z kwadratowej pamięci ...' (str 52 rozprawy). Istnieje al-

R

gorytm programowania dynamicznego dostarczający sekwencję uliniowienia w liniowej pamięci (Myers and Miller, 1988).

- wydaje mi się, że zgrabniej napisać 'kryteria są wzajemnie przeciwstawne' niż 'kryteria wzajemnie się wykluczają' (str. 82 rozprawy), bo to drugie oznacza brak rozwiązania;
- str 11 linia 14, usunąć przyimek 'w';
- proszę usunąć powtórzenia: informacja o tym, że mikroorganizmy zamieszkują wiele miejsc jest na str. 3, 9 i 16;
- j.w., informacja o tym, że potok SGA to zbiór programów jest na str 86 i 88;
- w akapicie rozpoczynającym się od 'Łącząc te dwa podejścia...' (str 52 rozprawy) powinno być 'z punktu widzenia odczytu globalne' zamiast 'z punktu widzenia genomu globalne';
- proszę dać jedną nazwę do programu będącego częścią SGA, bo raz jest 'assemble', a innym razem 'assembly' (str 87);
- moim zdaniem wygodniej czytać tekst, w którym są 2-liczbowe indeksy tablic i rysunków (tzn. numer rozdziału i numer porządkowy tablicy/rysunku wewnątrz rozdziału), zamiast 3-liczbowych, użytych w rozprawie (numer rozdziału, numer podrozdziału i numer porządkowy), przy 2-liczbowym indeksie łatwiej odnaleźć tabelę/rysunek w tekście;
- spotkałem się z nazwami kontig fizyczny i kontig sekwencyjny zamiast kontig (ang. contig) i skafold (ang. scaffold), str 44, proszę rozważyć taką terminologię.

5 Podsumowanie

Rozprawa wymaga wprowadzenia poprawek i ponownego recenzowania. Proponowane poprawki zamieściłem w sekcji 3, dotyczą one jedynie prośby o jednoznaczne określenie własnego wkładu w rozwiązania oraz szerszego opisania własnych rozwiązań. Temat badawczy uważam za bardzo ciekawy i istotny.

Osiągnięcia naukowe Kandydatki są wysokie, mgr Julia Herman-Iżycka posiada w swoim dorobku kilka prac współautorskich opublikowanych w renomowanych czasopismach naukowych i brała udział w grantach NCN i NCBiR.

Robert Nowak