POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań, tel. +48 61 665 2997, fax +48 61 877 1525
e-mail: office_cs@put.poznan.pl, www.cs.put.poznan.pl

Instytut Informatyki

Poznań, 07.11.2025 r.

Dr hab. inż. Aleksandra Świercz
Instytut Informatyki
Politechnika Poznańska

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY NAUKOWEJ INSTYTUTU GENETYKI ROŚLIN POLSKIEJ AKADEMII NAUK

Tytuł rozprawy: Metodologia statystyczna badania zmian jakościowych transkryptomu roślin w aspekcie interakcji genotypu i środowiska

Autor rozprawy: mgr Maria Nuc

Promotor: prof. dr hab. Paweł Krajewski

Promotor pomocniczy: dr inż. Hanna Ćwiek – Kupczyńska

Tematyka badawcza

Przedmiotem recenzowanej rozprawy doktorskiej jest opracowanie oraz zastosowanie nowej metody analizy danych transkryptomicznych, pozwalającej na uwzględnienie aspektu jakościowego w badaniach ekspresji genów. Autorka podjęła się problemu istotnego z punktu widzenia współczesnej biologii molekularnej i bioinformatyki, w której dominują metody ilościowe oparte na porównywaniu poziomów ekspresji genów. W pracy zaproponowano alternatywne podejście – analizę jakościową, umożliwiającą badanie różnic w składzie sekwencji RNA pomiędzy próbkami, a nie jedynie ich liczebności. Do tego celu wykorzystano odległość Mash, będącą miarą podobieństwa między zbiorami sekwencji nukleotydowych. Zaproponowana metoda została zastosowana do analizy danych z eksperymentu trójczynnikowego przeprowadzonego na jęczmieniu, w którym badano wpływ genotypu, dnia i temperatury na transkryptom roślin. Opracowany pipeline analityczny obejmował obliczanie odległości Mash, analizę głównych składowych (PCoA) oraz testowanie istotności czynników eksperymentalnych. Dodatkowo autorka wykorzystała ontologię GOSlim GOA w celu pogłębienia interpretacji biologicznej otrzymanych wyników. Przeprowadzone badania wpisują się w aktualne trendy w

bioinformatyce, łącząc metody ilościowe i jakościowe w analizie danych transkryptomicznych oraz podkreślając znaczenie wieloczynnikowych podejść eksperymentalnych w biologii roślin.

Zakres pracy i wkład autora

Wkład autorki ma charakter zarówno metodyczny, jak i aplikacyjny. W warstwie koncepcyjnej zaproponowała ona nową metodę jakościowej analizy danych RNA-seq, bazującą na odległości Mash, a w warstwie praktycznej opracowała kompletny schemat analityczny pozwalający na jej zastosowanie w doświadczeniach wieloczynnikowych. Autorka sformułowała trzy cele badawcze, stanowiące zarazem hipotezy pracy.

1. Opracowanie nowej metody analizy wyników doświadczeń wieloczynnikowych, w których uzyskano zbiory sekwencji transkryptomicznych.
2. Znalezienie sposobu na włączenie do tej metody interpretacji biologicznej, umożliwiającej powiązanie wyników z funkcjami genów przy użyciu ontologii GOSlim GOA.
3. Sprawdzenie, czy uwzględnienie odległości Mash w analizie ilościowej ekspresji może poprawić trafność i kontekst wyników biologicznych.

Uzyskane rezultaty potwierdziły zasadność postawionych hipotez – autorka wykazała, że odległość Mash może być skutecznym narzędziem do oceny jakościowej zmienności transkryptomów, a jej wykorzystanie prowadzi do wzbogacenia klasycznej analizy różnicowej o nowy wymiar interpretacyjny. Praca stanowi cenny wkład w rozwój metod bioinformatycznych wykorzystywanych w analizie danych RNA-seq i ma potencjał do dalszego zastosowania w badaniach genomiki funkcjonalnej.

Ocena strony merytorycznej

Praca doktorska liczy 98 stron i składa się z 7 rozdziałów oraz suplementu z tabelami. Zacytowane zostały 43 pozycje literaturowe, wliczając cytowane strony internetowe.

W rozdziale pierwszym Autorka wprowadza w problematykę współczesnych badań transkryptomicznych, podkreślając rosnące znaczenie analiz wieloczynnikowych w biologii roślin. We wstępie zaakcentowano potrzebę łączenia danych o różnym charakterze — ilościowym i jakościowym — w celu pełniejszego zrozumienia procesów ekspresji genów w kontekście interakcji genotypu i środowiska. Przedstawiono także motywację badawczą: ograniczenia istniejących metod analizy RNA-seq, które koncentrują się głównie na porównaniach ilościowych, pomijając aspekt jakościowy sekwencji, w tym także w ujęciu wieloczynnikowym. Autorka w sposób przejrzysty wskazuje miejsce swojej pracy w szerszym kontekście bioinformatyki i statystyki stosowanej. Cennym elementem jest ujęcie pracy w perspektywie interdyscyplinarnej — łączy aspekty biologiczne, matematyczne i informatyczne.

Rozdział drugi stanowi obszerne omówienie literatury przedmiotu dotyczącej sekwencjonowania DNA i RNA oraz metod badania zmienności transkryptomu. Autorka opisuje rozwój technologii sekwencjonowania od pierwszej generacji po współczesne techniki drugiej i trzeciej generacji. W dalszej części wprowadza pojęcia metryki i odległości, poczynając od klasycznych definicji (Euklidesowa, Hamminga, Damerau-Levenshteina) po nowoczesne metody przybliżonego porównywania sekwencji – w szczególności odległość Mash bazującą na indeksie Jaccarda. Mgr Nuc w sposób przystępny, ale naukowo precyzyjny, omawia metody porównywania sekwencji, a następnie przechodzi do uzasadnienia wyboru metody Mash jako centralnej koncepcji badawczej. Cenne jest przedstawienie podstaw matematycznych i dowodów spełniania przez Mash warunków metryki. Rozdział miejscami jest bardzo obszerny — część historyczna dotycząca rozwoju sekwencjonowania mogłaby być skrócona na rzecz szerszego omówienia współczesnych zastosowań analizy odległościowej w biologii molekularnej. Przy opisie, szczególnie w pierwszej części (2.1-2.2) jest dużo wtrąceń w nawiasach () co znacznie utrudnia czytanie tego rozdziału.

W rozdziale trzecim przedstawione zostały trzy główne cele pracy (przytoczone powyżej). Cele są jasno określone, hierarchicznie uporządkowane i dobrze korespondują z treścią dalszych rozdziałów. Widać, że autorka dąży do opracowania metody nie tylko teoretycznie poprawnej, ale także praktycznej i interpretowalnej biologicznie.

Rozdział czwarty zawiera szczegółowy opis procedur bioinformatycznych i statystycznych zastosowanych w pracy. Opisane zostały etapy przetwarzania danych RNA-seq: kontrola jakości, składanie odczytów, obliczanie odległości Mash, analiza współrzędnych głównych (PCoA), analiza wariancji (ANOVA, MANOVA) oraz analiza ekspresji genów. Uzupełnieniem jest wykorzystanie ontologii genów (GOSlim GOA) dla interpretacji biologicznej wyników. Rozdział cechuje się wysokim poziomem technicznym i precyzją opisu. Autorka umiejętnie łączy procedury bioinformatyczne z metodami statystycznymi, co świadczy o interdyscyplinarnej kompetencji. Dobrze uzasadniono wybór narzędzi (FastQC, SOAPdenovo2, Mash, DESeq2) oraz ich konfigurację. Szczególnie cenne jest uwzględnienie analizy kowariancji jako sposobu łączenia informacji jakościowej (PCoA) z ilościową (ekspresja genów).

W kolejnym rozdziale przedstawiono dane eksperymentalne, które zostały wykorzystane w omawianej pracy. Dane pochodzą eksperymentu trójczynnikowego na jęczmieniu (*Hordeum vulgare*), w którym analizowano wpływ genotypu, dnia i temperatury.

Rozdział szósty to najobszerniejszy rozdział pracy, w który przedstawione zostały wszystkie wyniki przeprowadzonych analiz. Autorka pokazała iż miary odległości Mash dla różnych etapów złożenia transkryptomów: kontigów oraz skafoldów są bardzo zbliżone, dlatego też w dalszej części analiz wyniki zostały przedstawione tylko dla jednej z nich (dla skafoldów), oraz dla danych surowych NGS. Przeprowadzono analizę PCoA dla macierzy odległości Mash pomiędzy każdą parą zbiorów (różny genotyp, powtórzenie biologiczne, temperatura oraz dzień), a następnie poprzez analizę wariancji (ANOVA, MANOVA) zbadano związek współrzędnych głównych z czynnikami biologicznymi (np. genotypem). Mgr Nuc wskazała, że największy wpływ

na zmienność jakościową mają czynniki temperatura i dzień. Następnie przeprowadziła analizę funkcjonalną GOSlim, pokazując różnice w kategoriach biologicznych genów, oraz analizę różnicowej ekspresji genów z wykorzystaniem modelu rozszerzonego o współrzędne główne.

Rozdział jest dobrze napisany i bogaty w interpretacje. Wyniki są spójne z założeniami metodologicznymi, a doktorantka wykazuje umiejętność interpretacji biologicznej danych statystycznych. Szczególnie wartościowe jest połączenie analiz jakościowych i ilościowych oraz porównanie modeli liniowego i kowariancyjnego. Wyniki są przedstawione klarownie, z poprawnym użyciem figur i tabel. Niektóre wykresy i tabele można by uzupełnić o jednostki miar i opisy osi (np. rys 6.9). Dodatkowo, interpretacja biologiczna mogłaby być pogłębiona przez odniesienie do konkretnych szlaków metabolicznych lub genów różnicujących.

Pojawiło się na tym etapie kilka uwag lub nieścisłości:

Przedstawione przykłady nie są dobrze opisane. Np. 'secondary metbolic process'. Policzone zostały odległości Mash dla każdej pary podzbiorów sekwencji związanych z tym procesem. Nie wiadomo ile było takich sekwencji? Można przypuszczać że chodzi tu o wszystkie geny, które są związane z tym procesem, a każdy gen mógł być reprezentowany przez więcej niż jeden skafold. Powinno to zostać bardziej doprecyzowane.

Tabela 6.15 oraz analiza różnicowej ekspresji przy pomocy DESeq2. Nie podano jaki jest prób odcięcia istotności statystycznej, p-value lub adjusted p-value? Jaki zastosowano próg odcięcia przy stwierdzeniu że jest zmiana ekspresji w dół lub w górę? Czy próg zmiany logFC może być np. 0.5?

Kolejny, szósty rozdział obejmuje dyskusję na temat syntezy wyników oraz ich odniesienie do literatury. Autorka pokazuje, że opracowana metoda jest efektywna obliczeniowo, a zarazem biologicznie interpretowalna. Wskazuje, że Mash może stanowić użyteczne narzędzie do analizy transkryptomów w kontekście wieloczynnikowym. Zwraca uwagę na potencjał zastosowania metody w innych gatunkach oraz możliwość integracji z klasycznymi metodami analizy ekspresji genów. Dyskusja jest spójna i logicznie powiązana z celami pracy. Autorka potrafi krytycznie ocenić ograniczenia swojej metody (m.in. uproszczenia w analizie k-merów, losowość haszowania). Wskazuje kierunki dalszych badań, w tym potrzebę walidacji międzygatunkowej i optymalizacji parametrów k-merów. Dyskusja mogłaby w większym stopniu akcentować oryginalność metody w porównaniu z innymi narzędziami (np. Mashmap, MinHash w innych kontekstach). Warto również uzupełnić część o refleksję nad możliwym zastosowaniem metody w analizach epigenetycznych lub metatranskryptomicznych.

Ostatni rozdział stanowi podsumowanie wyników pracy doktorskiej.

W pracy doktorskiej zabrakło informacji o dostępności skryptów lub narzędzia, które mogłoby zostać wykorzystane przez innych użytkowników. Choć część użytych narzędzi, jak np. AdapterRemoval, DESeq2, TopHat jest powszechnie znana i szeroko stosowana, to zaproponowana przez doktorantkę metoda obliczania

odległości Mash oraz wykorzystania macierzy odległości w analizie PCoA powinna być dostępna publicznie. Brakuje również szczegółów dotyczących jej implementacji.

W trakcie lektury, oprócz pytań zawartych w powyższym podsumowaniu pracy, pojawiły się również następujące kwestie:

Jakie czynniki zdecydowały o wyborze jęczmienia (Hordeum vulgare) jako organizmu modelowego? Czy metoda mogłaby być bezpośrednio zastosowana również do innych gatunków (np. roślin dwuliściennych lub organizmów modelowych typu Arabidopsis thaliana)? Czy większa różnorodność pomiędzy genotypami wpłynęłaby na jakość i interpretowalność osiągniętych wyników?

W analizie odległości Mash zastosowano wartość k = 21. Czy autorka przeprowadziła testy wrażliwości wyniku na zmianę długości k-merów? W jakim stopniu wartość tego parametru wpływa na dokładność i interpretowalność wyników?

Jakie były kryteria wyboru metody PCoA do analizy współrzędnych głównych? Czy rozważano alternatywne podejścia redukcji wymiarów, np. MDS, t-SNE, UMAP lub NMDS?

Ile prób biologicznych i technicznych uwzględniono dla każdego wariantu czynnika? Czy autorka przeprowadziła analizę mocy statystycznej, aby potwierdzić, że liczba powtórzeń była wystarczająca do wykrycia efektów czynników?

W jakim stopniu opracowana metoda może zastąpić klasyczne podejście do analizy RNA-seq, a w jakim powinna je uzupełniać?

Czy możliwe byłoby zastosowanie tej metody do danych pochodzących z sekwencjonowania długich odczytów (np. Oxford Nanopore)? Czy większy odsetek błędów w danych trzeciej generacji wpłynie istotnie na jakość wyników? Czy konieczny byłby dodatkowy etap oczyszczania odczytów?

Ocena od strony redakcyjnej

Praca od strony redakcyjnej wymaga dopracowania. Zabrakło informacji o autorze pracy doktorskiej, promotorze oraz jednostce przeprowadzającej jest przewód doktorski.

Do wniosku powinna być dostarczona lista publikacji autorki oraz wykaz jej udział w konferencjach naukowych. Na dorobek naukowy doktorantki składają się publikacje w Nature Plants, BMC Genomics, Scientific Data, Scientific Reports, Frontiers in Plant Science oraz International Journal of Molecular Sciences. Żadna z tych prac nie została jednak wymieniona ani zacytowana w rozprawie.

Praca miejscami napisana jest w formie raportu, na przykład wnioski przedstawiono w pracy w formie wypunktowanej. Zdarzają się drobne błędy stylistyczne oraz interpunkcyjne, a także nieścisłości w zapisie nazw metod i odległości, np.

- miara odległości Damerau-Levenshteina - w pracy jest ‘Damareau-Levenshteina’
- metoda DESeq2, w pracy jest ‘Deseq2’

Powyższe usterki nie mają znacznego wpływu na jakość i czytelność pracy i nie umniejszają jej wartości. Nie zmieniają również mojej pozytywnej oceny recenzowanego doktoratu.

Wnioski końcowe

W mojej ocenie autorka recenzowanej pracy doktorskiej przeprowadziła szereg złożonych i wymagających analiz. Zaproponowała nową metodę jakościowej analizy danych RNA-seq opartą na odległości Mash, wykazując, że może ona stanowić skuteczne narzędzie do oceny zmienności transkryptomów. Miara odległości Mash może również zostać wykorzystana jako modyfikacja klasycznej analizy różnicowej ekspresji, pozwalająca uwzględnić podobieństwo sekwencji transkryptomicznych. Zastosowana aparatura badawcza oraz metody analityczne są adekwatne do założonych celów i przeprowadzonych badań.

Stwierdzam, że praca doktorska pani mgr Marii Nuc pt. „Metodologia statystyczna badania zmian jakościowych transkryptomu roślin w aspekcie interakcji genotypu i środowiska" spełnia **warunki określone w art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz.U. 2018 poz. 1668 z późn. zm.)** ***oraz stanowi oryginalne rozwiązanie problemu naukowego.***

Wnoszę o dopuszczenie mgr Marii Nuc do dalszych etapów przewodu doktorskiego.

Dr hab. inż. Aleksandra Świercz

