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

Kraków, 09.01.2024

Recenzja rozprawy doktorskiej

Tytuł rozprawy:	Modeling count data in molecular biology and sociophysics: selected applications
Autor rozprawy:	mgr Agnieszka Geras
Promotor rozprawy:	Dr hab. Ewa Szczurek, prof. UW
Dziedzina:	nauki inżynieryjno-techniczne
Dyscyplina:	Informatyka Techniczna i Telekomunikacja

Rozwój wysokoprzepustowych biomedycznych technik pomiarowych pozwolił w ostatnich latach na znaczny postęp w naukach biologicznych i medycznych. Szczególne przyspieszenie nastąpiło w związku z rozwojem Next Generation Sequencing (NGS) i technik pochodnych. Jednym z przykładów jest zastosowanie NGS do ilościowej oceny ekspresji genów i ich alternatywnych transkryptów. Domyślną strategią jest charakteryzacja transkryptomu w populacjach komórek pochodzących z określonej tkanki (ang. bulk RNAseq). To podejście jest motywowane założeniem, że komórki z tego samego typu tkanki są jednorodne i jednocześnie takie podejście jest historycznie odziedziczone po technikach nisko-przepustowych oraz mikromacierzowych. W rezultacie uzyskane wyniki są swego rodzaju średnią ważoną z różnych profili ekspresji z badanej populacji komórek. Podejście to bardzo dobrze sprawdza się w ogólnej charakteryzacji tkanki, jednak w szczególnych przypadkach, gdzie interesujący nas sygnał nie pochodzi od dominującej grupy komórek, jest niewystarczające. Alternatywnym podejściem jest sekwencjonowania RNA pojedynczych komórek (scRNA-Seq), które pod względem technologicznym dynamicznie rozwija się w ostatnich latach. W tym podejściu co do zasady uzyskujemy profile ekspresji dla każdej komórki z osobna, co pozwala na rozróżnienie nie tylko pod względem typu komórki ale także ich fazy „cyklu życia".

Jednak analiza danych pochodzących z eksperymentów scRNA-Seq jest trudnym zadaniem, znacznie trudniejszym niż z „klasycznego" RNA-Seq (bulk RNA-Seq). Z jednej strony musimy mierzyć się z wysoką wymiarowością danych, a z drugiej z szumem o złożonej etiologii. W typowym eksperymencie scRNA-Seq dla każdej próbki dla dziesiątek a nawet setek tysięcy komórek mierzymy poziomy ekspresji dla setek tysięcy alternatywnych transkryptów genów (często dla uproszczenia sumowanych do poziomu genów, co wiąże się z utratą części informacji). Jednocześnie cały proces generacji danych jest złożony z wielu etapów, gdzie na każdym sygnał może ulec zaszumieniu. Do tego oczywiście dochodzą ograniczenia i charakterystyka samej technologii

sekwencjonowania. Na całość nakłada się także charakterystyka biologiczna: z naturalną, stochastyczną różnicą w sygnałach pomiędzy komórkami tego samego typu, różnicą między typami komórek, jak i różnicą wynikająca z fazy „cyklu życia" danej komórki. Większość z tych składowych jest też obecna w danych z eksperymentów typu „bulk RNA-Seq", jednak większa złożoność przygotowania próbki dla scRNA-Seq jak i dodatkowy wymiar złożoności w postaci dziesiątek/setek tysięcy komórek powoduje, że analiza tego typów danych jest trudniejsza zarówno ze względu na bardziej złożony szum jak i większą wymiarowość danych w stosunku do klasycznego RNA-Seq.

Mówiąc o analizie danych (sc)RNA-Seq mamy na myśli szereg etapów, gdzie w pewnym momencie otrzymujemy macierz zliczeń, gdzie typowo kolumny reprezentują analizowane próbki a rzędy geny lub ich alternatywne transkrypty, natomiast w komórkach znajdują się wartości estymacji częstości/zliczeń. Dane zliczeniowe są wszechobecne nie tylko w biologii molekularnej ale też w wielu dziedzinach jak ekonomia, epidemiologia, ekologia czy socjofizyka. Składają się one z nieujemnych liczb całkowitych reprezentujących częstotliwość zdarzeń, takich jak liczba transakcji wykonywanych na rynkach finansowych lub liczba elementów lub osobników w określonej kategorii, np. liczba gatunków w siedlisku.

Przedstawiona do oceny rozprawa doktorska koncentruje się na danych zliczeniowych pojawiających się w biologii molekularnej, a konkretnie transkryptomice oraz socjofizyce, gdzie w analizie mediów społecznościowych analizowane są dane zliczeniowe generowanych z interakcji międzyludzkich na internetowej platformie StackOverflow. W tym drugim obszarze zbudowany model jest następnie przenoszony na dane bibliometryczne a mianowicie index Hirscha.

W projektach transkryptomicznych (scRNA-Seq, ST RNA-Seq) autorka przedstawia dwa innowacyjne modele probabilistyczne, które umożliwiają fenotypowanie komórek na podstawie danych z sekwencjonowania o wysokiej przepustowości. W celu estymacji parametrów modelu, zaproponowano algorytmy statystyczne oparte na próbkowaniu Monte Carlo łańcuchami Markowa. Pierwszą wprowadzoną metodą jest hierarchiczny model Bayesowski o nazwie *Celloscope*. Model ten wykorzystuje wiedzę na temat genów markerowych do dekompozycji mieszanek typów komórek w danych transkryptomiki przestrzennej. Autorka wykazuje, że *Celloscope* skutecznie zidentyfikował znane struktury mózgu myszy, a w szczególności potrafił różnicować pomiędzy neuronami hamującymi a pobudzającymi. Drugi opisany i wyprowadzony w pracy model, *ST-Assign*, służy do jednoczesnego fenotypowania pojedynczych komórek na podstawie danych sekwencjonowania RNA oraz dekompozycji mieszanek typów komórek w danych transkryptomiki przestrzennej. Autorka wykazuje jego skuteczność na danych symulowanych, a także integrując dane z przedniej części mózgu myszy powstałe w efekcie dwóch różnych eksperymentów.

W ostatnim przytaczanym projekcie, tym razem z dziedziniy socjofizyki, przeprowadzono analizę problemu "negatywnych cytowań" na podstawie danych z serwisu *Stack Exchange*. Opracowany model agentowy pozwala na przybliżanie wskaźników cytowalności oraz na wysnucie uniwersalnych wniosków na temat zagadnień, takich jak ocena osiągnięć naukowych. Uzyskane przez autorkę wyniki pokazują minimalny wpływ wprowadzenia "negatywnych cytowań" na najbardziej popularny wskaźnik cytowalności, czyli indeks Hirscha.

Układ rozprawy jest typowy dla tego typu opracowań. W kolejnych rozdziałach Autorka:

- przybliża statystyczne i probabilistyczne koncepcje niezbędne do wyjaśnienia, w jaki sposób szacowane są parametry modeli *Celloscope* i *ST-Assign* (Rozdział 2)
- przedstawia narzędzie *Celloscope* oraz wyniki badań na symulowanych danych oraz wyniki identyfikacji typów komórek w mózgu myszy (Rozdział 3)
- opisuje rozszerzenie *Celloscope*, aby dodatkowo wykonać zadanie przypisywania typów komórek w danych sekwencjonowania scRNA-Seq (Rozdział 4)
- wprowadza podstawowe koncepcje wywodzące się z nauki o sieciach (Rozdział 5),
- opisuje nowy model oparty na agentach wykorzystywany do ilościowego określenia siły wpływu wprowadzenia "negatywnych cytowań" na ocenę badań naukowych (Rozdział 6).

Autorka bardzo dokładnie omawia obecny stan wiedzy i szeroko opisuje wymagane podstawy co świadczy o dużym oczytaniu i dobrym przygotowaniu do podjęcia zagadnień poruszanych w dalszych częściach pracy. Praca jest dobrze napisana i widać, że Autorka posiada rzadką umiejętność płynnego przeprowadzenia czytelnika po szerokim spektrum zagadnień bez gubienia z oczu celu ostatecznego. W tym aspekcie przydarzyły się jednak dwa potknięcia: i) we wprowadzeniu brakuje jednoznacznego wytłumaczenia, że w projekcie trzecim dokonywana była eksploracja i budowa modelu na danych z *Q&A StackOverflow* i tak zbudowany model został następnie przeniesiony na dane cytowalności publikacji naukowych w celu estymacji ilości „negatywnych cytowań“ oraz określenia ich wpływu na indeks Hirscha (ja momentami byłem zagubiony); ii) na stronie 102 pojawia się zdanie: „*Let us note that information about the approximated positive scores of every StackOverflow user can be represented by a triple (n, P, κU).*“ gdzie *κu* pojawia się jedynie wcześniej w równaniu 6.4 na stronie 100 z wyjaśnieniem „*κu, κd are parameters minimising RMSLE*“. Wydaje mi się, że jest to trochę za mało zwłaszcza, że *κu* jest określany jako jeden z trzech podstawowych parametrów używanych w modelu.

O ile praca mi się bardzo podoba jest jedna kwestia, która wymaga głębszego zastanowienia przez Autorkę i być może wprowadzenia poprawek w kolejnych wersjach narzędzi. Autorka we wprowadzeniu słusznie zauważa, że:
The processes that generate the count data are often complex and elusive, and we can only observe them through the counts they produce. Empirical observations are critical to uncover the hidden mechanisms that govern the observed processes. Moreover, proposing adequate models is often crucial to draw meaningful conclusions about the system's present state or assess the impact of hypothetical changes.
Dlatego krytyczne jest zrozumienie, że wynikiem eksperymentu (sc)RNA-Seq nie są dane zliczeniowe ale sekwencje, z których dopiero po analizie i różnego rodzaju korekcjach estymujemy częstości występowania czyli dane zliczeniowe gotowe do dalszej analizy. W przypadku (sc)RNA-Seq na etapie przygotowania bibliotek do sekwencjonowania w co najmniej kilku krokach protokół postępowania mówi, że do następnego kroku należy wziąć konkretną (pod względem objętości) ilość materiału. Ilość materiału po ekstrakcji będzie zależna z jednej strony od ilości komórek, a z drugiej od chwilowej aktywności danej komórki, bo im więcej „się działo” w danym momencie tym więcej mRNA zostało wyprodukowane. Wpływ też będą miały (dla ST i bulk RNA-Seq) proporcje między różnymi

typami komórek czy komórek o różnym poziomie aktywności. W rezultacie stosowanie ustandaryzowanych procedur laboratoryjnych prowadzi do zaburzenia estymacji częstości występowania danego mRNA (na poziomie genów czy alternatywnych transkryptów). Oznacza to, że te same wartości zliczeń dla różnych komórek (scRNA-Seq) czy spotów (ST RNA-Seq) nie oznaczają tego samego poziomu ekspresji. Dlatego też wymienione w tabelach 3.1 i 4.1 oraz oznaczone w graficznych reprezentacjach modeli stałe:

λ0 - base gene expression level, shared across all genes and cell types,

Λgt - over-expression level for a marker gene g in its specific cell type t

będą prawdziwe jedynie w szczególnych, a przez to rzadkich przypadkach. Dla danych transkryptomicznych istnieją sprawdzone metody normalizacji/korekcji międzypróbkowej jak TMM (https://doi.org/10.1186/gb-2010-11-3-r25). W dostępnej dokumentacji narzędzi 10X Genomics zarówno dla scRNA-Seq jak i ST RNA-Seq nie znalazłem wzmianki aby normalizacja TMM lub równoważna była stosowana. Nie widzę jej również w omawianej pracy. Brak odpowiedniej normalizacja międzypróbkowej może prowadzić do fałszywych wyników i dlatego zalecałbym wbudowanie jej do kolejnych wersji narzędzi.

Z rzeczy o mniejszym kalibrze podoba mi się dyskusja czy dane ze StackOverflow mogą być referencją dla zbudowania modelu dla danych o cytowaniu publikacji naukowych, ale w mojej ocenie jest ona niepełna. Jednak w StackOverflow krytyka nawet ostra zaproponowanego rozwiązania nie jest niespotykana w wersji słownej, a co dopiero w postaci kliknięcia „łapki w dół", co znaczy, że dana odpowiedź nie jest użyteczna. W artykułach naukowych gzie nie wiemy, kto będzie recenzentem, a nie zawsze jest możliwość wykluczenia konkretnych osób, raczej nie ma otwartej krytyki. Standardem jest określenie cech metod do których się porównujemy a następnie pokazania co ponad to ma nasze rozwiązanie. Otwarta krytyka występuje raczej jedynie w listach do edytorów. Uważam, że tą dyskusję można było w pracy pociągnąć dalej.

W kwestii „drobiazgów" w podpisie do Ryc 6.8 interpretacja wyników dla scenariusza I i II są zamienione. W tekście jest już poprawnie.

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

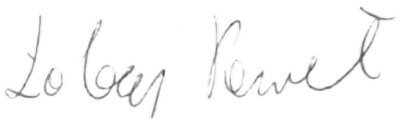
Rozprawa zawiera 26 opisanych rycin oraz kilka nienumerowanych szkiców oraz 9 tabel, które co do zasady są jasne i czytelne. W pracy ze względu na charakter zagadnienia znajduje się też znaczna ilość wzorów, uwag, definicji, przykładów czy opisów algorytmów. Piśmiennictwo obejmuje 145 co do zasady dobrze dobranych i aktualnych.

Podsumowując, przedstawiona do oceny praca doktorska stanowi bardzo wartościowe uzupełnienie obecnego stanu wiedzy odnośnie analizy danych zliczeniowych zarówno w obszarze biologii molekularnej jak i socjofizyki. W dziedzinie transkryptomiki przedstawiono dwa innowacyjne modele probabilistyczne: *Celloscope* i *ST-Assign*. Pierwszy z nich ma na celu identyfikację typów komórek w przestrzennych danych transkryptomicznych na podstawie wcześniejszej wiedzy o markerach typów komórek. Badanie na symulowanych danych wykazało wyjątkową wydajność *Celloscope* w

porównaniu do innych modeli, natomiast analiza danych rzeczywistych z powodzeniem ujawniła wzorce strukturalne w mózgu myszy. W przypadku drugiego narzędzia, *ST-Assign*, Autorka wykazała, że radzi sobie z trudnym zadaniem jednoczesnego przypisywania typów do poszczególnych komórek na podstawie scRNA-Seq i dekonwolucji na podstawie ST RNA-Seq. W dziedzinie socjofizyki Autorka opracowała nowy, oparty na agentach, model wyników cytowań użytkowników na popularnej platformie *StackOverflow*. Badanie empiryczne wykazało zdolność modelu do przewidywania wektoru cytowani dla indeksu Hirscha. Co więcej, analizy przeprowadzone z wykorzystaniem modelu wykazały, że wpływ negatywnych cytowań na h-index jest co do zasady nieistotny zarówno dla danych ze StackOverflow jak i cytowń publikacji naukowych.

Na podstawie powyższej oceny stwierdzam, że wymieniona rozprawa doktorska w pełni odpowiada warunkom stawianym w ustawie Prawo o szkolnictwie wyższym i nauce / Dz. U. z 2022 r. poz. 574, w zakresie nadawania stopni naukowych i na tej podstawie wnoszę do Wysokiej Rady Dyscypliny Informatyki Technicznej i Telekomunikacji Politechniki Warszawskiej o dopuszczenie mgr Agnieszki Geras do dalszych etapów przewodu doktorskiego.

Nie mam wątpliwości, że doświadczenie zgromadzone przez Autorkę stawia cały zespół badawczy w doskonałej pozycji wśród międzynarodowych grup zajmujących się tą tematyką.



Dr hab. inż. Paweł Piotr Łabaj