Recenzja pracy doktorskiej mgr Agnieszki Geras

dr hab. Witold Rudnicki, prof. UwB
Wydział Informatyki, Uniwersytet w Białymstoku,
ul. Konstantego Ciołkowskiego 1M, 15-245 Białystok,
tel. +48 85 738 8276
e-mail: W.Rudnicki@uwb.edu.pl

Ocena pracy doktorskiej pt. „*Modeling count data in molecular biology and sociophysics: selected applications*" przygotowanej przez mgr Agnieszkę Geras, pod opieką dr hab. Ewy Szczurek, prof. UW, w dziedzinie nauk inżynieryjno-technicznych, w dyscyplinie informatyki technicznej i telekomunikacji.

I Omówienie rozprawy

Rozprawa doktorska jest napisana w języku angielskim. Została ona przygotowana na podstawie materiału z trzech artykułów:

- A. Geras, et. al. "Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics data," Genome Biology, vol. 24, no. 1, 2023.
- A. Geras and E. Szczurek, "ST-Assign: a probabilistic model for joint cell type identification in spatial transcriptomics and single-cell RNA sequencing data," bioRxiv, 2023.
- A. Geras, G. Siudem, and M. Gagolewski, "Should we introduce a dislike button for academic articles?" Journal of the Association for Information Science and Technology, vol. 71, no. 2, 2020.

We wszystkich doktorantka była pierwszą autorką. Należy podkreślić, dwa artykuły zostały opublikowane w renomowanych czasopismach, wysoko cenionych w swoich dziedzinach, trzeci jest dostępny na platformie biorxiv.org. W rozprawie nie jest wspomniane czy artykuł został wysłany do recenzowanego czasopisma, ale można założyć, że wersja recenzowana pojawi się w równie dobrym czasopiśmie jak pozostałe.

Rozprawa jest poświęcona dwóm bardzo różnym zagadnieniom, które są połączone na dość podstawowym poziomie charakterem formalnym danych, co znajduje odzwierciedlenie w tytule. Pierwsze z tych zagadnień to rozwój metod analizy ekspresji genów w transkryptomice przestrzennej (spatial transcriptomics). Drugie to analiza hipotetycznego zmodyfikowanego indeksu Hirsha, który uwzględniałby ujemne oceny przy cytowaniach.

Rozprawa składa się z siedmiu rozdziałów.

Rozdział pierwszy to wstęp, wprowadzający przedmiot badań w pierwszej części rozprawy. **Sekcja 1.1** wprowadza podstawowe pojęcia w zakresie transkryptomiki, w szczególności opisuje dwa nowoczesne rodzaje pozwalające na badanie ekspresji genów z dużo większą rozdzielczością niż standardowa – czyli transkryptomikę jednokomórkową i transkryptomikę przestrzenną. W szczególności, autorka przedstawiła podstawowe wyzwania przy analizie danych transkryptomiki przestrzennej. W tej odmianie transkryptomiki materiał biologiczny pobierany jest w postaci bardzo cienkiego wycinka badanej tkanki. Wycinek jest dzielony przestrzennie na rozłączne pola, po izolacji mRNA z każdego pola oddzielnie jest ono znakowane, tak, że można zlokalizować jego pochodzenie w próbce. W najczęściej stosowanych protokołach, jedna lokalizacja zawiera 10 do 100 różnych komórek, które mogą należeć do różnych typów, które należy zidentyfikować dla poprawnej analizy. Doktorantka następnie wymienia i opisuje pokrótca podstawowe metody stosowane do identyfikacji typów komórek obecnych w próbce. **Sekcja 1.2** z kolei stanowi wprowadzenie do problematyki analizy sieci powiązań w dwóch sieciach społecznych: sieci cytowań w artykułach naukowych oraz ocen odpowiedzi na pytania zadawane na platformie Stack Exchange. W **sekcji 1.3** doktorantka przedstawia pokrótce projekty, w których brała udział i których wyniki są źródłem metod i wyników opisanych w rozprawie. Następnie doktorantka przedstawia przedmiot pozostałych rozdziałów rozprawy.

Trzy następne rozdziały są poświecone opisowi projektu rozwoju narzędzi do analizy ekspresji genów w transkryptomice przestrzennej i jednokomórkowej.

Rozdział drugi, wprowadza podstawy teoretyczne dla dwóch rozdziałów trzeciego i czwartego, poświęconych opisowi narzędzi do opracowanych przez doktorantkę narzędzi do analizy danych transkryptomicznych. Rozdział składa się z trzech sekcji. Krótka **sekcja 2.1** wprowadza pojęcie niezależności warunkowej. **Sekcja 2.2** zawiera krótkie wprowadzenie do sieci Bayesowskich. W szczególności wprowadzone zostały pojęcia acyklicznego grafu skierowanego i przedstawiona definicja sieci Bayesowskich, a także koca Markowa. W szczególności pokazana została możliwość wykorzystania sieci Bayesowskich do generowania rozkładów zmiennych losowych. W **sekcji 2.3** trzeciej autorka przedstawiła łańcuchy Markowa i metodę generowania rozkładu prawdopodobieństwa metodą próbkowania Monte Carlo łańcuchów Markowa. W szczególności przedstawiła dwa powszechnie stosowane algorytmy, czyli algorytm Metropolisa-Hastingsa, oraz próbkowanie Gibbsa.

W Rozdziale trzecim, doktorantka przedstawia opracowane przez siebie narzędzie do analizy danych z transkryptomki przestrzennej – Celloscope. Rozdział jest oparty o wspomnianą wcześniej pracę A. Geras, et. al., Genome Biology, vol. 24, no. 1, 2023.

Projekt opisany w tym rozdziale miał na celu zaproponowanie ulepszonego rozwiązania opisanego w sekcji 1.1 istotnego problemu w transkryptomice przestrzennej, a mianowicie zidentyfikowania typów komórek znajdujących się w każdej z badanych lokalizacji. Rozdział trzeci zaczyna się od dokładniejszego przedstawienia rozwiązywanego problemu. W

szczególności przedstawia wyzwania związane ze stosowaną powszechnie metodą przypisania typów komórek, a mianowicie zastosowania jednocześnie metod sekwencjonowania jednokomórkowego i przestrzennego dla próbki. Niestety, w wielu wypadkach nie tej metody zastosować, a tam gdzie to się udaje, błędy wynikające z integracji wyników dwóch metod obciążonych dużymi błędami, mogą być bardzo duże.

Założeniem projektu było wymaganie, że identyfikacja typów komórek znajdujących się w lokalizacji powinna odbywać się z wykorzystaniem wyłącznie danych uzyskanych z eksperymentu transkryptomiki przestrzennej. Jedyne wcześniejsze narzędzie o podobnym charakterze pozwalało na identyfikację różnych typów komórek i ich liczebności w próbce na podstawie istniejących korelacji między ekspresją genów.

Punktem wyjścia dla metodologii opracowanej przez doktorantkę było założenie, że dla typów komórek mogących pojawić się w próbce znane są geny markerowe o podwyższonej ekspresji w danym typie, w stosunku do innych typów.

Sekcja 3.1 zawiera ogólną prezentację modelu Celloscope, pokazującą sposób działania metody.

Sekcja 3.2 Poświęcona jest bardzo szczegółowej prezentacji modelu, opisu parametrów modelu ich roli w modelu i powiązania z analizowanymi danymi. Prezentacja zawiera zarówno opis słowny jak i precyzyjne sformułowania matematyczne.

Sekcja 3.3 przedstawia z kolei algorytm wykorzystujący próbkowanie Monte Carlo łańcuchów Markowa prowadzący do określenia wartości parametrów modelu. Opis jest szczegółowy i precyzyjny, poza opisem słownym algorytm jest opisany w sposób formalny z wykorzystaniem precyzyjnej notacji matematycznej.

Sekcja 3.4 przedstawia wyniki zastosowania modelu Celloscope dla danych symulowanych i porównania z algorytmami CellAssign, Stereoscope, RCTD, SpatialDWLS i BayesPrism. Badania przeprowadzono dla kilkunastu różnych scenariuszy, z różnymi założeniami o liczbie komórek i ich liczebności. We wszystkich przypadkach algorytm Celloscope uzyskało bardzo dobre wyniki, nawet w wypadku bardzo ograniczonej informacji czy wysokiego szumu. W szczególności wyniki te były lepsze niż uzyskane w analizach przeprowadzonych przy użyciu konkurencyjnych agorytmów.

Sekcja 3.5 przedstawie z kolei wyniki analiz uzyskanych w badanich dwóch wycinków tkanki mózgowej myszy, uzyskanych z dwóch różnych lokalizacji przy pomocy dwóch różnych technologii eksperymentalnych. Użycie Celloscope pozwala na znacznie dokładniejsze przedstawienie przestrzennego rozkładu różnych typów komórek w porównaniu z narzędziem CellAssign. W szczególności widoczne jest, że różne obszary funkcjonalne nie są homogeniczne, ale oprócz dominującego w danym obszarze typu komórek, powiązanych z bezpośrednio z funkcją, zawierają często duże i różnorodne populacje innych typów komórek o innych funkcjonalnościach. Doktorantka przedstawiła również analizę wrażliwości wyników na dobór parametrów wejściowych. Analiza pokazuje, że dobór właściwych parametrów jest bardzo ważny dla uzyskania dobrych wyników.

Sekcja 3.6 zawiera podsumowanie i dyskusję opisanych wyników.

Rozdział czwarty jest poświęcony opisowi narzędzia ST Assign przeznaczonego do jednoczesnej analizy danych z transkryptomiki przestrzenej i jednokomórkowej. ST Assign jest rozwinięciem metodologii opisanej w poprzednim rozdziale. W szczególności punktem wspólnym jest oparcie analiz na znajomości genów markerowych dla typów komórek. Należy podkreślić, że znane są jedynie same geny, a nie ich profile ekspresji. Te ostatnie należą do parametrów modelu i są określane w procesie dopasowania do wyników eksperymentalnych. ST Assign jest hierarchicznym modelem Bayesowskim, w którym obserwowane wartości ekspresji w obu rodzajach eksperymentów są zależne od nieznanych wartości ekspresji genów w różnych typach komórek w transkryptomice jednokomórkowej oraz od liczby komórek określonego typu w transkryptomice przestrzennej. Należy podkreślić, że wyniki analizy

Struktura rozdziału jest podobna do struktury rozdziału poprzedniego.

Sekcja 4.1 zawiera generalny opis modelu, w sekcji 4.2 model przedstawiony jest bardzo szczegółowo, z wyczerpującym opisem parametrów i algorytmu dopasowywania nieznanych parametrów modelu. Sekcja 4.3 przedstawia matematyczny formalizm i szczegóły implementacji modelu. W sekcji 4.4 przedstawione są wyniki analiz przeprowadzonych dla sześciu różnych scenariuszy opisujących różne rodzaje danych. Sekcja 5 zawiera opis badań przeprowadzonych na danych rzeczywistych – jednego z zestawów użytych wcześniej w rozdziale 3. Podobnie jak w rozdziale 3, sekcja 4.6 zawiera podsumowanie i dyskusję wyników.

Rozdział piąty jest poświęcony na wprowadzenie do problemu, oraz krótki opis pojęć i narzędzi matematycznych wykorzystanych do analizy sieci cytowań. W szczególności opisane są dwa rozkłady prawdopodobieństwa stosowane do opisu sieci oddziaływań społecznych, a mianowicie rozkład potęgowy oraz rozkład Tsallisa-Pareto.

Rozdział szósty jest poświęcony analizie wpływu uwzględnienia „negatywnych cytowań" na ocenę wpływu w wybranych sieciach społecznościowych, w szczególności wpływu na indeksy bibliograficzne.

W **sekcji 6.1** opisane są dwa zbiory danych używane w rozprawie, jeden to cytowania w literaturze naukowej pobrane z bazy stworzonej w projekcie „The Initiative for Open Citations", druga to pozytywne i negatywne oceny odpowiedzi zgromadzonych na platformie StackOverflow, służącej do dzielenia się wiedzą w dziedzinie informatyki.

W sekcji 6.2 przeprowadzona jest globalna analiza obu zbiorów danych, a w szczególności pokazano, że zarówno liczba cytowań, jak i liczba głosów negatywnych i pozytywnych są zgodne z rozkładem Tsallisa-Pareto.

Sekcja 6.3 poświęcona jest sprawdzeniu hipotezy, że rozkład głosów pozytywnych i negatywnych na platformie StackOverflow jest uzyskany w wyniku mechanizmu wiązania preferencyjnego.

W sekcji 6.4 przeprowadzona została analiza ocen pozytywnych i negatywnych uzyskanych przez pojedynczych autorów. Również w tym wypadku rozkład prawdopodobieństwa jest zgodny z rozkładem Tsalisa-Pareto.

W sekcji 6.5 została przeprowadzona dyskusja podobieństwa procesów wystawiania ocen w serwisie StackOverflow do cytowań naukowych.

W sekcji 6.6 został przeanalizowany wpływ uwzględnienia negatywnych ocen na ogólną ocenę prestiżu autora odpowiedzi na platformie StackOverflow. W tym celu posłużono się indeksem Hirscha (indeksem H), który jest często stosowaną miarą oceny wpływu badaczy w środowisku naukowym. Indeks H badacza wynosi k, jeśli co najmniej k publikacji uzyskało k cytowań. Analogicznie można zdefiniować indeks H dla autora odpowiedzi na platformie StackOverflow. W sekcji przeprowadzono wpływ uwzględnienia negatywnych ocen w na wartość liczbową indeksu H. Zostały przebadane dwa warianty tego wpływu - neutralny, oraz pesymistyczny. W obu wariantach liczba ocen pozytywnych odpowiedzi została pomniejszona przez liczbę ocen negatywnych, przy czym w wariancie pesymistycznym całkowita liczba ocen negatywnych została rozdystrybuowana między wszystkie odpowiedzi, tak aby w stopniu maksymalnym obniżyć wartość indeksu H. Okazuje się, że niezależnie od przyjętego sposobu generowania indeksu H uwzględniającego oceny negatywne, wpływ ocen negatywnych nie był duży. W większości przypadków wartość indeksu H nie była zmieniona, w pozostałych przypadkach obniżenie indeksu wynosiło 1, a w bardzo rzadkich przypadkach 2.

Sekcja 6.7 zawiera dyskusję wyników.

Rozdział siódmy zawiera podsumowanie całej rozprawy.

Ogólna ocena rozprawy

Rozprawa zawiera oryginalne rozwiązanie dwóch dość luźno powiązanych ze sobą problemów naukowych. Połączenie między nimi jest dość sztuczne - modelowanie danych zliczeniowych wspomniane w tytule. Zarówno charakter danych, jak i używana metodologia są na tyle odległe, że trudno znaleźć inne realne połączenie między problemami poza osobą doktorantki. Jak rozumiem połączenie tych dwóch odrębnych projektów w jednej rozprawie wynikało z zupełnie niepotrzebnej obawy, że materiał z projektu poświęconego rozwojowi narzędzi do analizy danych transkryptomiki przestrzennej mógłby okazać się niewystarczający do spełnienie wymagań dla rozprawy doktorskiej.

W efekcie bardzo nowatorski projekt wnoszący nową wiedzę i bardzo cenne narzędzia dla zrozumienia fundamentalnych procesów biologicznych został powiązany z projektem interesującym i ciekawym, ale o zupełnie innym ciężarze gatunkowym. W mojej opinii rozprawa zyskałaby na wartości i sile, gdyby skupiła się na wyłącznie na pierwszym temacie a w miejsce rozdziałów 5 i 6 znalazła się nieco rozszerzona prezentacja metodologii i wyników prezentowanych w rozdziałach 2, 3 i 4.

Badania opisane w pierwszej części pracy są klasy światowej. Dotyczą ważnego problemu - czyli analizy danych o ogromnym znaczeniu dla zrozumienia procesów biologicznych odbywających się na poziomie pojedynczych komórek z dodatkową wiedzą w postaci dokładnej znajomości położenia w badanej tkance dane komórki się znajdują. Określenie lokalizacji komórek danego typu z możliwością jednoczesnego uchwycenia stanu komórki opisanego przez poziomy ekspresji genów daje fenomenalną perspektywę dla

zrozumienia jak funkcjonuje organizm. Od strony medycznej te dane pozwalają badać procesy nowotworzenia z niemożliwą wcześniej rozdzielczością przestrzenną i komórkową.

Narzędzia opracowane przez doktorantkę oparte są na bardzo solidnym fundamencie matematycznym doskonale opisane od strony formalnej i co najważniejsze dają doskonałe wyniki, znacząco lepsze niż uzyskiwane z wykorzystaniem dotychczas dostępnych narzędzi.

Należy podkreślić, że ograniczenie się do określonego zbioru typów komórek, nakłada na model silne więzy, które z jednej strony pozwala na lepsze dopasowanie wyników do rzeczywistego składu komórek w badanej tkance, jeśli one rzeczywiście występują w badanej tkance, z drugiej strony ogranicza to możliwość wykrycia identyfikacji innych typów komórek.

Jednak, nie jest to ograniczenie absolutne. Wskazuje na to wykrycie przez algorytm ST-Assign spójnych wyników dla typu nieokreślonego (dummy) co wskazywało na możliwość pojawienia się w danych jednego typu komórek, które nie były pierwotnie włączone do analizy. Powtórzona analiza, w której, z jednej strony dodano ekspresje dla komórek z należących do typu „Peptigeneric neurons" w analizie ekspresji jednokomórkowej, a z drugiej dodane zostały geny markerowe dla tego typu komórek dała wyniki spójne z obserwacjami.

Wyniki prezentowane w drugiej części rozprawy są interesujące. Ciekawym pomysłem jest powiązanie dwóch zbiorów, które na pierwszy rzut oka nie są bardzo podobne - z jednej strony ocen odpowiedzi do pytań stawianych na platformie StackOverflow a z drugiej liczby cytowań artykułu naukowego. Jednak głębsza analiza przeprowadzona w pracy wskazuje, że podobieństwo jest istotne. Zostało to wykorzystane do zbadania czy wprowadzenie negatywnych ocen do prac naukowych miałoby znaczenie przy obliczaniu indeksu Hirscha. Indeks Hirscha używany jest często jako uproszczona miara wpływu danego badacza na piśmiennictwo naukowe. Okazuje się, że wpływ taki jest niewielki w wypadku analogicznego indeksu na platformie StackOverflow, a zatem należy wnioskować, że podobny efekt byłby uzyskany, gdyby można było w jakiś sposób wprowadzić negatywne oceny prac naukowych. W przeciwieństwie do wyników prezentowanych w pierwszej części pracy jest to raczej ciekawostka bez znaczenia praktycznego - w praktyce prace słabo oceniane po prostu nie są cytowane a przykłady wysoko cytowanych prac błędnych i niepotrzebnych są bardzo rzadkie - osobiście nie jestem w stanie podać takiego przykładu.

Wspomniana wcześniej niespójność pracy jest jedynym poważniejszym zarzutem dla pracy, nie umniejszającym wartości prezentowanej pracy.

II Poprawność

Ocena ogólna

Rozprawa jest bardzo dobrze napisana, z dobrą kompozycją, podziałem na rozdziały i sekcje. Bardzo podoba mi się przyjęta konwencja nadawania tytułów na poziomie sekcji - tytuły sekcji są zdaniami twierdzącymi podsumowującymi zawartość sekcji.

W rozprawach doktorskich często umieszczane są spisy tabel i ilustracji, tutaj tego brakuje, z drugiej strony wszystkie trzy rozdziały opisujące wyniki uzyskane przez doktorantkę są oparte na samodzielnych opublikowanych artykułach naukowych, więc znalezienie właściwych tabel czy ilustracji nie jest specjalnie trudne.

Od strony merytorycznej nie mam żadnych zastrzeżeń, opisy modeli i algorytmów są poprawne i wyczerpujące, wyniki są dobrze opisane.

Uchybienia

Najpoważniejszym zarzutem jest niepełna spójność i brak lepszego porównania i dyskusji wyników pomiędzy rozdziałami 3 i 4. W szczególności brakuje porównania ilościowego wyników rozdziałów 3 i 4 dla dokładnie tych samych danych i jakościowej i ilościowej odpowiedzi na to jaka jest wartość dodana modelu ST Assign w stosunku do Celloscope. Na stronach 81/82 jest zamieszczone zdanie „The obtained cell type decomposition (Figure 4.3) is in agreement with previous results acquired with Celloscope (Figure 3.4). „ ale niestety nie jest ono poparte dokładniejszymi danymi. Obszar analizowany w rozdziale 4 jest jedynie wycinkiem obszaru z rozdziału 3, co bardzo utrudnia analizę wizualną, nie ma podanych miar liczbowych tego podobieństwa na wspólnym obszarze.

Drobniejsze uchybienia.

Doktorantka dość losowo posługuje się zamiennie pojęciami StackOverflow i Stack Exchange. Relacja między nimi nie jest wyjaśniona. Być może nie jest to wielki problem w środowisku programistów, gdzie wiedza o StackOverflow jest powszecha, ale należałoby jednak zadbać o więcej precyzji.

Ilustracja 5.1 - oś Y w panelu C podana jest w bardzo dziwnych jednostkach. W szczególności niemożliwe jest by w Polsce procent miast o populacji w okolicach 2 mln wynosił 10^{-9} a procent miast z populacją około 10 000 wynosił 10^{-4}.

III Wiedza doktorantki

Doktorantka wykazała w pracy wiedzę zarówno szeroką, jak i głęboką. W obydwu badanych problemach przedstawiła dobrą znajomość stanu wiedzy w dziedzinie transkryptomiki i w dziedzinie sieci społecznych, a także w dziedzinie modelowania probablilistycznego.

Jedyne uchybienie jakie udało mi się znaleźć w pracy dotyczy przypisania opracowania metody Monte Carlo „ *The Metropolis-Hastings algorithm [64] is historically the first and still the most important MCMC algorithm.*" Istotnie algorytm Metropolisa jest prawdopodobnie najszerzej stosowanym algorytmem Monte Carlo. Jednakże palma pierwszeństwa należy do Stanisława Ulama i Johna von Neumana (Oczywiście można również dyskutować, czy algorytm Metropolisa jako najczęściej stosowany jest ważniejszy niż pierwszy algorytm Ulama).

1. S. Ulam, R. D. Richtmyer, and J. von Neumann. 1947. Statistical methods in neutron diffusion. Alamos Scientific Laboratory report LAMS-55 1.
2. N. Metropolis and S. Ulam. 1949. The Monte Carlo method. Journal of the American Statistical Association 44:335-341.
3. S. Ulam. 1950. Random processes and transformations. Proceedings of the International Congress of Mathematicians 2~264-275.
4. Metropolis, Nicholas, et al. "Equation of state calculations by fast computing machines." *The journal of chemical physics* 21.6 (1953): 1087-1092.

IV Podsumowanie

Po przeczytaniu i przeanalizowaniu przedstawionej rozprawy oceniam, że:

1. Rozprawa przedstawia oryginalne rozwiązanie problemu naukowego (zdecydowanie TAK)
2. Kandydatka posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka techniczna i komunikacja (zdecydowanie TAK)
3. Kandydatka posiada umiejętność samodzielnego prowadzenia pracy naukowej (zdecydowanie TAK).

Stwierdzam, że rozprawa doktorska pani magister Agnieszki Geras spełnia warunki określone w art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce, tekst jednolity: Dz.U. z 2021 r. poz. 478. W związku z powyższym wnioskuję o dopuszczenie pani Agnieszki Geras do dalszych etapów przewodu doktorskiego.

Jednocześnie wnioskuję o wyróżnienie rozprawy.

dr hab. Witold Rudnicki, prof. UwB

Białystok, 5.12.2023