Mpt. RPITT – 22.06.2023
M. Słomy

instytut biologii doświadczalnej
im. Marcelego Nenckiego PAN

Dr hab. Bartosz Wojtaś, Profesor Instytutu
Pracownia Sekwencjonowania,
Instytut Biologii Doświadczalnej
im. M. Nenckiego, PAN

Recenzja pracy doktorskiej mgr inż. Aleksandry Suwalskiej
"Developing a system of automatic identification of cellular subpopulations in data from single-cell mass cytometry with the use of algorithms for grouping of high-dimensional data"

Recenzję opracowano na podstawie uchwały 31/2023 Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej z dnia 28 marca 2023 r. Przedstawiona mi do oceny praca miała na celu stworzenie automatycznej analizy danych pochodzących z eksperymentów cytometrii masowej (CyTOF) z komórek pochodzących z bronchoskopii od pacjentów z gruźlicą lub z innymi chorobami płucnymi w porównaniu do osób zdrowych. Badano panel białek błonowych i wewnątrzkomórkowych w celu identyfikacji różnych populacji komórkowych i zidentyfikowaniu różnic pomiędzy komórkami płuc w stanie patofizjologicznym w porównaiu do płuc zdrowych oraz stymulowanych lub nie stymulowanych pochodną tuberkuliny (PPD) lub fitohemoaglutyniną (PHA). Podstawowym celem badawczym, było przetestowanie istniejących metod analiz w zakresie ustawiania bramek (gating) komórek, czyli filtrowaniu danych wstępnych w eksperymencie, jak i normalizacji, usuwania efektu paczki, jak i wizualizacji, klastrowania, poszukiwania subpopulacji komórkowych i wreszcie samej analizy różnicowej pomiędzy grupami pacjentów oraz różnego typu traktowania komórek. Etap testowania prowadzony był głównie na danych ze zbioru Samusik et al., gdzie etapy bramkowania oraz identyfikacji populacji komórkowych były ustalane przez ekspertów po czym parametry i ustawienia były testowane na zbiorze od współpracowników doktorantki, gdzie takie adnotacje nie były znane. Przedstawiona praca w dużej części jest więc pracą typu meta-analizy oraz porównania istniejących narzędzi, dodatkowo jednak doktorantka stosuje własne podejście analityczne do identyfikacji subpopulacji komórek oparty o Modele mieszanin Gaussowskich. Niezwykle ciekawym podejściem było dla mnie poszerzenie zbioru zmiennych (expanded feature space), co przy wynikach z metody CyTOF, gdzie mamy zazwyczaj jedynie kilkadziesiąt badanych zmiennych a setki tysięcy badanych komórek, może mieć bardzo ważne znaczenie dla analizy, co zresztą doktorantka pokazała w swoich wynikach.

Przedstawiona do oceny rozprawa doktorska ma oryginalną formułę, zawierającą 4 główne części z bardzo intuicyjnym układem, w którym pierwsza jest klasycznym wstępem do 1) metody CyTOF, 2) problemu badawczego od strony biologicznej (gruźlica) oraz 3) zarysem głównych problemów analitycznych. Trzy następne części pracy doktorskiej opisują osobne części analizy, takie jak: Pre-gating danych z eksperymentów CyTOF, usuwanie efektu serii oraz właściwa identyfikacja subpopulacji komórek. Każda z trzech części dotyczących analizy zawiera własną część Materiałów i Metod, Wyników oraz Dyskusji. Piątym rozdziałem pracy jest podsumowanie, a szóstym suplement, po którym jest spis literatury, który zawiera 59 pozycji. Spis literatury zawiera wszystkie pozycje, związane z metodami analizy, z których doktorantka korzystała w swojej pracy. Na tyle, na ile jestem w stanie to ocenić, uważam że zakres wybranej literatury jest wyczerpujący.

Na uwagę zasługuje fakt, że przedstawione wyniki mogą służyć za przydatny przewodnik dla ludzi wchodzących w analizy CyTOF, gdzie opisane są metody, które zostały też przez doktorantkę przetestowane w praktyce. Wysoko oceniam wartość merytoryczną i praktyczną wykonanych badań. Wysoko oceniam sprawność doboru i posługiwaniem się zastosowanych przez doktorantkę metod badawczych. Przykładowo, w części dotyczącej analizy subpopulacji komórkowych doktorantka porównuje 9 istniejących metod.

Mam do pracy kilka pytań ogólnych, kilka szczegółowych oraz kilka komentarzy, które pozwolą mi lepiej zrozumieć pracę i zorientować się co kierowało doktorantką w podejmowaniu pewnych decyzji analitycznych:

Pytania ogólne:

1. W danych scRNAseq stan cyklu komórkowego można oszacować w pojedynczej komórce, co samo w sobie jest interesującą obserwacją, ale w wielu analizach cykl komórkowy jest uważany bardziej za zmienną zakłócającą, ponieważ często jest to jedno z głównych źródeł zmienności, maskującym różnice pomiędzy subpopulacjami komórkowymi. Jak to jest w danych CyTOF, czy poziomy badanych białek są stabilne na różnych etapach cyklu komórkowego?
2. Mam wrażenie że pełna automatyzacja analiz przedstawionych przez doktorantkę może być bardzo trudna, ponieważ wydzielenie niektórych klastrów komórek bez wiedzy eksperckiej może nie być możliwe. Na przykład w procesie różnicowania komórek pojawienie się jednego markera może wyraźnie określić los komórki, która może nadal wyglądać jak komórka typu A, ale jest już na nieodwracalnej ścieżce, aby być komórką typu B. W świetle tej wiedzy czy

byłoby możliwe dodanie wagi do niektórych znaczników, aby wydzielać bardziej dyskretne populacje komórek ?

3. Dyskusja w rozprawie nie zawiera prawie odniesień do innych publikacji, co jest o tyle zaskakujące, że zazwyczaj jest to część rozprawy, w której uzyskane wyniki są omawiane w kontekście istniejącej literatury. Jaki jest tego powód?
4. Czy doktorantka może w trakcie obrony przejść przez wyniki str 24, Fig 2.2. ponieważ dla mnie są one niejasne. W szczególności nie wiem skąd bierze się wielkość i kształt okręgów w pierwszej kolumnie. Myślę, że warto by tu omówić czym są obiekty w lewym dolnym rogu oraz górnym prawym rogu ryciny A, które zostały wyfiltrowane z analizy. O ile rozumiem, że np. w prawym dolnym rogu to może być po prostu cell debris, to czym w takim razie są obiekty w górnym prawym rogu ?
5. Na końcu strony 27 doktorantka sugeruje, że strategię pre-bramkowania można jeszcze zoptymalizować, aby skrócić czas obliczeń, ale bez podawania szczegółów. Rodzi to pytanie: jakie idee kryją się za tym stwierdzeniem?
6. Na początku rozdziału 4.2.5 doktorantka pisze: „Some artifacts must be corrected to make the results reliable and use conditional probabilities as new features.”. Jakiego rodzaju artefaktów się tutaj spodziewamy i jakie może być ich źródło?

Zależałoby mi żeby podczas obrony doktorantka ustosunkowała się do pytań ogólnych, z uwagi na ograniczenia czasowe pytania szczegółowe oraz drobne komentarze będę mógł omówić indywidualnie z doktorantką.

Pytania szczegółowe:

1. Na rycinie 2.1: dlaczego ostatni wykres w rzędzie C jest tak różny od A i B? Wygląda na to, że nastąpiła transformacja danych, której nie było w A i B ? Czy był inny powód ?
2. Na stronie 27 rozprawy opisany jest czas obliczeń dla Gaussian Mixed Models z algorytmem Expectation-Maximization, czy w tym czasie obliczeń (do 48h) wykorzystano wszystkie 2TB RAM i 256 wątków ?
3. W sekcjach 4.2.4 i 4.2.5 użyto etykiety „pdf” osi y, ale nie mogę znaleźć informacji w podpisach, co ona oznacza, zakładam, że jest to akronim, ale od czego ?
4. Nie do końca rozumiem punktu drugiego na stronie 52. Co rozumiemy przez słowo “restored”? Czy odnosi się to do tego, że prawdopodobieństwa pierwszego i ostatniego składnika prawdopodobieństwa (component probability) powinny wynosić 1?

5. Wizualizacja na rycinie 4.8. nie jest dla mnie jasna, czerwone strzałki wskazują nieaktywne elementy, które należy usunąć, podczas gdy w ostatnim wierszu ryciny jeden z elementów wskazanych czerwoną strzałką jest zachowany?
6. Na stronie 63 podano, że wykluczono wszystkie komponenty zawierające mniej niż 6,4% zbioru danych. Dla mnie z opisu nie wynika jasno jak ten próg został ustawiony. Proszę mnie również poprawić, jeśli się mylę, ale słowo „zbiór danych" nie jest tutaj trafne, ponieważ nie mówimy tutaj o zbiorze danych (wszystkie dane), ale o niewielkiej części naszego zbioru danych dotyczącej jednego typu komórki?

Komentarze:

1. Podrozdział 2.1, pierwsze zdanie odnosi się do martwych komórek i uważam to za nieco mylące. Czy wszystkie nie są martwymi komórkami, które zostały utrwalone przed procedurą barwienia? Nie mogłem znaleźć tej konkretnej informacji w rozdziale 1, ale zakładam, że są one utrwalone przed eksperymentem CyTOF.
2. Nie wyjaśniono, co oznacza „E" we wzorach ARI i AMI.
3. Coś złego stało się na stronie 75, gdzie czcionka została zmieniona na mniejszą, a opis rysunku został zdublowany.
4. W większości części rozprawy język angielski jest, o ile mogę to ocenić, poprawny i zrozumiały, ale są fragmenty, w których szyk zdania jest niepoprawny i wpływa to na proces czytania i rozumienia tematu, np. zdanie na stronie 85: „ However, the significant impact (p-value<0.05) that is also confirmed with the large effect size (>0.14) has the patient group on the composition of the cell populations.".
5. Tabela 4.9: proszę uważać przy zaokrąglaniu wartości p, ponieważ w tabeli znalazły się wartości p równe 0.
6. W niektórych częściach rozprawy doktorskiej - przykład 4.1 znajduje się opis kilku porównywanych algorytmów, a ich opis jest czasami zbyt skąpy, aby dokładnie ocenić, jak one działają, np. w opisie SPADE jest fragment: „the density of outliers and targets". Bez dokładniejszego opisu trudno powiedzieć, jak definiowane są wartości odstające i docelowe w tym algorytmie. Rozumiem, że w tego rodzaju częściach benchmarkingu/metaanalizy często zdarza się, że ciężko jest dać wyczerpujący opis wszystkich zastosowanych metod, ale uważam, że ważne byłoby podkreślenie, jaka byłaby metoda do oceny wartości odstających w tej i innych użytych metodach.

Większość niedociągnięć w tekście jest drobnych i nie przesłania to ogólnego bardzo dobrego wrażenia, jakie sprawia przedstawiona mi do oceny praca doktorska.

W podsumowaniu stwierdzam, że przedstawiona rozprawa stanowi oryginalne rozwiązanie problemu naukowego. Dodatkowo stwierdzam, że doktorantka prezentuje bardzo wysoki poziom wiedzy teoretycznej w zakresie dyscypliny oraz szczegółową wiedzę i umiejętności samodzielnego prowadzenia pracy naukowej a zatem odpowiada warunkom określonym w art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (t.j. Dz. U. 2022, poz. 85 z poźn. zm.). Wnioskuję zatem do Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej o dopuszczenie mgr inż. Aleksandry Suwalskiej do dalszych etapów przewodu doktorskiego.

Dodatkowo ze względu na dużą wartość naukową i wartość praktyczną uzyskanych wyników oraz imponującą aktywność publikacyjną doktorantki, wnioskuje o wyróżnienie tej pracy przez Radę Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej.

Bartosz Wojtaś
Warszawa, 15.06.2023