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

Kraków, 16.05.2024

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

Małopolskie Centrum
Biotechnologii

Recenzja rozprawy doktorskiej

Tytuł rozprawy:	Analiza metylacji sekwencji CpG w oparciu o uczenie maszynowe i sieci neuronowe
Autor rozprawy:	lek. Tomasz Falgowski
Promotor rozprawy:	Prof. dr hab. n. med. Tadeusz Sulikowski
Dziedzina:	nauki medyczne i nauki o zdrowiu
Dyscyplina:	nauki medyczne

Rozwój wysokoprzepustowych biomedycznych technik pomiarowych pozwolił w ostatnich latach na znaczny postęp w naukach biologicznych i medycznych. Szczególne przyspieszenie nastąpiło w związku z rozwojem technik mikromacierzowych oraz sekwencjonowania następnych generacji (Next Generation Sequencing - NGS). Pozwoliły one nie tylko znacznie poszerzyć naszą wiedzę w obszarach genomiki czy transkryptomiki ale też wejść w nowe rejony takie jak epigenetyka. Jak pisze autor rozprawy „złożoność ludzkiego genomu, polega nie tylko na określonej kompozycji miliardów par zasad, ale także chemicznej modyfikacji, która może być odczytywana i interpretowana przez enzymy i inne czynniki molekularne". Powszechnymi podejściami do analizy metylacji są platforma Illumina Infinium BeadChips, sekwencjonowanie z użyciem wodorosiarczynu (bisulphite sequencing), a ostatnio bezpośrednie sekwencjonowanie nanoporowe. Za metodę referencyjną uważa się metodę mikromacierzową i w przedstawionej do oceny pracy autor również prezentuje ten pogląd. Niezależnie od użytej techniki pomiarowej stopień metylacji danej sekwencji CpG określa się parametrem β, który wyraża proporcję pomiędzy badanymi komórkami metylowanymi i niemetylowanymi, oraz osiąga wartości z przedziału [0, 1].

Mogłoby się wydawać, że analiza danych metylacyjnych, a zwłaszcza analiza różnicowa, nie powinna przysparzać problemów. Jednak charakterystyka rozkładu danych metylacyjnych jest mocno nietypowa w porównaniu do innych danych biomedycznych i w związku z tym klasyczne podejścia nie zawsze sobie radzą. Ten problem został zauważony przez doktoranta, który zaproponował, że wykorzystanie technik uczenia maszynowego może być alternatywą, gdyż rozwiązania te są na tyle elastyczne, że są w stanie nauczyć się wzorców nawet w bardzo złożonych zagadnieniach. W szczególności autor zdefiniował, że celem rozprawy jest

Adres:
Gronostajowa 7a,
30-387 Kraków

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

Małopolskie Centrum
Biotechnologii

ocena czy architektura sieci neuronowej oparta o kombinację sieci neuronowej konwolucyjnej i typu transformers, jest porównywalnym lub lepszym narzędziem do analizy metylacji DNA.

Autor w swojej analizie wykorzystuje trzy zbiory danych, które z punktu widzenia analizy metylacji powinny być odpowiednio: łatwy (metylacja DNA limfocytów B oraz limfocytów T CD4+), trudniejszy (metylacja DNA zdrowych limfocytów B oraz limfocytów B pacjentów chorujących na przewlekła białaczkę limfatyczną) i trudny (metylacja DNA pacjentów z przewlekłą białaczką limfocytową z IGHV 100% lub mniej).

Autor wykazuje, że zaproponowana metoda była porównywalna lub lepsza niż metody standardowe. Dla prostszych zbiorów danych uzyskana selektywność była porównywalna do metod standardowych. Dla trudniejszych zbiorów danych zaproponowana metoda z jednej strony zidentyfikowała większą ilość metylacji różnicujących, a z drugiej wykazała się wyższą wydajnością w ich doborze z punktu widzenia zdolności do rozróżnienia pomiędzy grupą kontrolną i badaną. Ponadto autor wykazuje, że zaproponowana metoda dostarcza wyników o większym znaczeniu biologicznym. Autor konkluduje, że zaproponowana metoda może być alternatywną i skuteczną metodą do analizy danych metylacji DNA.

Układ rozprawy jest dla mnie atypowy co w mojej opinii ma duży wpływ na trudność w zrozumieniu rozprawy. W kolejnych rozdziałach Autor:

- Przybliża różne obszary medycyny, biologii, biotechnologii i (bio)informatyki (Wstęp)
- Definiuje cel pracy (Cel Pracy)
- Opisuje zbiory danych oraz metodykę badania wraz ze zdefiniowaniem metryk do oceny skuteczności metod w analizie metylacji DNA (Materiały i Metodyka)
- Przedstawia uzyskane wyniki (Wyniki),
- Dokonuje interpretacji wyników oraz formułuje wnioski (Wnioski)
- Przedstawia dalsze plany (Dodatkowe moduły i dalszy plan rozwoju)

Po Bibliografii oraz spisach znajdują się jeszcze skróty/streszczenia pracy w języku angielskim i polskim.

Tematyka pracy jest bardzo ciekawa i zdecydowanie jest to obszar rozwojowy, jednak sposób prezentacji w znacznym stopniu utrudnia właściwy odbiór, a w wielu wypadkach pewnie i zrozumienie pracy i jej wyników. Ogólny układ pracy powoduje, że czytelnik przez pierwsze 64 strony (czyli prawię połowę efektywnej części) nie wie o co chodzi. Opisywane są różne zagadnienia z zakresu medycyny, biologii, biotechnologii czy (bio)informatyki, które nie stanowią logicznej całości. Ponadto cały wstęp jest bardzo nierówny, są fragmenty bardzo dobrze napisane, o dużym stopniu szczegółowości gdzie widać, że autor świetnie wie co pisze. Ale też są takie, gdzie ważne kwestie są omówione pobieżnie, a następujące po sobie zdania nie tworzą całości. Odnosi się wrażenie, że autor coś wie (bo w końcu używa tych narzędzi), ale wiedza jest powierzchowna i często brakuje głębszego zrozumienia problemu. Dodatkowym utrudnieniem jest kwestia, że wielokrotnie po pobieżnym

Adres:
Gronostajowa 7a,
30-387 Kraków

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

Małopolskie Centrum
Biotechnologii

omówieniu jakiegoś tematu jest uwaga, że dokładniejszy opis nastąpi w dalszej części pracy. Taki zabieg co do zasady nie jest zły, ale wymaga aby czytelnik już na tym etapie miał jakiś pogląd odnośnie tego co jest robione, dlaczego i w jaki sposób. W przeciwnym wypadku, czytelnik po dojściu do fragmentu gdzie jest ten dokładniejszy opis nie jest w stanie tego właściwie odnieść. O ile w mojej ocenie układ powinien być kompletnie inny, to tak naprawdę już samo przesunięcie streszczeń na sam początek znacznie poprawiło by czytelność i zrozumienie rozprawy.

Mniejszego kalibru sprawami ale też irytującymi z punktu widzenia przejrzystości i czytelności są: i) brak numeracji rozdziałów, podrozdziałów itd.; ii) stosunkowo dużo „błędów" w postaci nadmiarowych spacji lub ich braku przed/po przecinkach, kropkach, nawiasach, itd.; iii) brak rozpoczynania rozdziałów czy podrozdziałów od nowej strony, a w innych miejscach urywanie zdania w połowie i kontynuowanie na kolejnej stronie. Występują też błędy typu: „0,1 % procenta" (strona 21), „brak odwołania" (strona 60), jest „Bart" a powinno być „BERT" (strona 62), słowo „Wyniki" w streszczeniu w języku angielskim itp. Ogólnie w mojej ocenie praca została przygotowana i złożona dość niechlujnie.

Przechodząc do kwestii merytorycznych widać, że o ile na pewno wiedza autora odnośnie technik uczenia maszynowego ma tendencję wzrostową to brakuje solidnych podstaw co objawia się wieloma lukami. Algorytmy AI to jest szczególny przypadek uczenia maszynowego a nie na odwrót (strona 25), kwestia technik Deep Learning (głębokiego uczenia, strona 33) to nie tylko zwiększenie ilości warstw ale cała koncepcja jak taką głęboką strukturę nauczyć, tutaj adekwatną pozycją do zacytowania jest praca magisterska prof. Hochreitera o LSTM. LSTM pojawia się wprawdzie później w pracy ale w przypadku tak przełomowej techniki przyzwoitość nakazuje zacytowanie oryginału, który jest z roku 1997. Brakuje mi też rozpoznania co w kwestiach wykorzystania technik eksploracji danych i uczenia maszynowego w analizie danych biomedycznych, jak choćby danych obrazowych, dzieje się na naszym Polskim podwórku. Brakuje choćby odniesienia się do olbrzymich sukcesów na tym polu zespołu prof. Polańskiej z Politechniki Śląskiej w Gliwicach.

Ważnym tematem, który autor „poruszył" na stronie 40 (i dalszych) jest kwestia tego, że sieci neuronowe, a zwłaszcza rozwiązania głębokiego uczenia wymagają dużej ilości danych do uczenia z jednej strony, a z drugiej powstaje model o dużej złożoności. Naturalnym kierunkiem rozważań powinno być odniesienie się do Brzytwy Ockhama i kwestii metryk złożoności modelu jak Kryterium informacyjne Akaikego (AIC) czy Bayesowskie kryterium informacyjne Schwartza (BIC). Autor sam dobrał zestawy danych tak, żeby były zarówno proste jak i złożone. W wynikach i wnioskach autor konkluduje, że dla prostszych zbiorów wyniki są porównywalne. W związku z tym nasuwa się pytanie „Czy warto?" i właśnie metryki takie jak AIC czy BIC mogłyby na takie pytanie odpowiedzieć. Niestety temat ten nie został podjęty w rozprawie, a szkoda. Powiązana z tym jest też kwestia złożoności obliczeniowej, która gdzieś przemyka we wstępie ale nie jest później rozwijana w odniesieniu do zaproponowanego rozwiązania. Złożoność ma

Adres:
Gronostajowa 7a,
30-387 Kraków

wpływ na wymagane zasoby (moc obliczeniową, ilość pamięci RAM, konieczność użycia GPU itd.) oraz na czas obliczeń. I znowu w przypadku prostszych zbiorów danych nasuwa się pytanie „Czy warto?”. Niestety autor nie przedstawia danych, które by mogły dać odpowiedź. Powoduje to, że choć zaproponowana metoda wydaje się być właściwym rozwiązaniem w złożonych przypadkach, to nie jesteśmy w stanie określić przy jakim poziomie złożoności koszty zaczną przeważać, a w związku z tym dla których problemów, zgodnie z Brzytwą Ockhama, prostsze rozwiązania będą „wystarczająco dobre”.

Autor powinien też pogłębić swoją widzę odnośnie korekcji przy testowaniu wielokrotnym. W przypadku eksperymentów gdzie mierzone jest wiele cech zupełnie bezcelowe jest patrzenie na *p-value* przed korektą. Ponadto należy raczej powiedzieć, że stosuje się poprawkę „z rodziny FDR”, czyli taką która stabilizuje FDR na zadanym poziomie. Przykładem jest poprawka „Benjamini–Hochberg”, która właśnie często określana jest jako FDR, z tym, że nie jest ona jedyna w tej rodzinie, a ponadto zakłada ona niezależności testowanych cech, co w przypadku prawie wszystkich danych biomedycznych nie jest warunkiem spełnionym. W takim wypadku lepszym rozwiązaniem jest poprawka „Benjamini–Yekutieli”, która wprawdzie jest bardziej restrykcyjna ale nie ma warunku o niezależności badanych cech.

Ostatnią kwestią, jest sprawa użycia „map cieplnych, autor napisał: „Analiza map cieplnych z hierarchicznym grupowaniem jest metodą obserwacyjną i eksploracyjną, która w prosty i intuicyjny sposób pozwala na interpretację zależności i wzorców obecnych w wielkoskalowych wynikach takich jak analiza macierzy metylacji”. Niestety, ale fundamentalnie nie zgadzam się z takim stwierdzeniem. Mapy cieplne mogą być użyte do prezentacji wyników, które zostały już przeanalizowane innymi obiektywnymi metodami, niezależnymi od zdolności percepcji osoby analizującej. Mapy cieplne nie nadają się do analizy eksploracyjnej bo bardzo łatwo jest je „zmanipulować”, żeby było widać „jakiś” wzorzec.

W rezultacie pomimo wykonania olbrzymiej pracy i zaprezentowania znacznej ilości wyników należy stwierdzić, że poza przytoczonymi wnioskami w wynikach drzemie znaczny potencjał, który mógłby być uwolniony przy zastosowaniu bardziej zaawansowanych, ale przy tym obiektywnych metod eksploracji i analizy.

Pomimo tych wszystkich zarzutów, należy podkreślić dużą wartość pracy, która łączy dość trudne do połączenia obszary wiedzy. Tezy rozprawy są sformułowane jasno (choć kilkadziesiąt stron za późno) i przystępnie oraz są w pełni poparte wynikami i wnioskami zawartymi w rozprawy. Podsumowanie rozprawy jest syntetycznym wykazaniem, że założone w pracy cele zostały osiągnięte

Rozprawa zawiera 68 opisanych rycin z tym, że w spisie następuje przesunięcie numeracji z 26 na 28 i w związku z tym numeracja kończy się na 69. Ponadto jako ryciny uwzględniono też wzory, co jest nietypowym zabiegiem. Praca zawiera też 31 tabel. Piśmiennictwo mimo, że jest bardzo obszerne (373 pozycje) ma jednak swoje braki, co zostało wspomniane wcześniej.

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

Małopolskie Centrum
Biotechnologii

Adres:
Gronostajowa 7a,
30-387 Kraków

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

Małopolskie Centrum
Biotechnologii

Podsumowując, przedstawiona do oceny praca doktorska stanowi przyczynek do eksploracji bardzo ciekawego ale też trudnego obszaru wiedzy z pogranicza medycyny, biologii, biotechnologii i (bio)informatyki. Autor wykazał się znaczną determinacją w zagłębieniu się w temacie, jednak jest to chyba przedsięwzięcie przerastające możliwości pojedynczej osoby, co widać w pracy. Autor opracował nową metodę analizy danych metylacyjnych wykorzystującą techniki uczenia maszynowego (kombinacja sieci neuronowych konwolucyjnych i typo transformers). Autor wykazał, że zaproponowana metoda była porównywalna lub lepsza niż metody standardowe. Oraz, że jej przewaga rośnie wraz z wzrostem złożoności problemu.

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 Nauki Medyczne Pomorskiego Uniwersytetu Medycznego o dopuszczenie lek. Tomasza Falgowskiego do dalszych etapów przewodu doktorskiego.

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

Elektronicznie
podpisany przez
Paweł Piotr Łabaj
Data: 2024.05.16
10:01:43 +02'00'

Dr hab. inż. Paweł Piotr Łabaj, Prof. UJ

Adres:
Gronostajowa 7a,
30-387 Kraków