Gliwice, dn. 25.07.2023

Dr hab. Monika Pietrowska
Narodowy Instytut Onkologii im. Marii Skłodowskiej – Curie
Państwowy Instytut Badawczy
Oddział w Gliwicach
ul. Wybrzeże Armii Krajowej 15
44-102 Gliwice
tel. (32) 278 96 27
e-mail: monika.pietrowska@io.gliwice.pl

RECENZJA

rozprawy doktorskiej **mgr inż. Wojciecha Sikory** zatytułowanej „Machine learning-based workflow for the analysis of MALDI-TOF mass spectrometry cancer data" wykonanej na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej w Gliwicach pod opieką Prof. dr hab. Joanny Polańskiej.

Technika MSI jest obecnie najlepiej rozwiniętą metodą badawczą umożliwiającą skorelowanie kompletnego profilu molekularnego z obrazem morfologicznym tkanki, pozwalającą na przypisanie cech molekularnych do konkretnego dobrze zdefiniowanego obszaru tkanki czy typu komórek. Zaletą techniki MSI jest jej uniwersalność – analizowane mogą być praktycznie wszystkie typy analitów, które ulegają jonizacji i rejestrowane są za pomocą spektrometrii mas (MS). Mogą to więc być bardzo różne typy cząsteczek biologicznych zazwyczaj analizowane metodami MS: białka, peptydy, lipidy, małocząsteczkowe metabolity i inne substancje pochodzenia endogennego i egzogennego (leki, ksenobiotyki). Nawet najbardziej zaawansowane technicznie współczesne metody analityczne (np. metody hybrydowe łączące immunohistochemię i MS) umożliwiają jednoczesną analizę jedynie kilkunastu-kilkudziesięciu antygenów. W tym samym czasie metody MSI umożliwiają jednoczesną detekcję setek i tysięcy odrębnych analitów różniących się masą molekularną. Tak więc w przypadku wielowymiarowego obrazowania molekularnego białek i metabolitów technika MSI pozostaje obecnie bezalternatywna. Pierwotna idea MSI wykorzystująca spektrometrię mas (MS) z desorpcją/jonizacją laserową wspomaganą matrycą (Matrix-Assisted Laser-Desorption Ionization, MALDI) i analizatorem czasu przelotu (Time of Flight, ToF) została zaproponowana ponad dwie dekady temu w laboratorium kierowanym przez prof. Richarda M. Caprioli (Vanderbilt University, Nashville), jednak wzrost zainteresowania tą metodą i jej coraz bardziej powszechne zastosowanie w badaniach biomedycznych to kwestia ostatnich lat. Chociaż w strategii MSI wykorzystywane są obecnie różne metody MS, MALDI-MSI pozostaje najbardziej rozpowszechniona.

Przedmiotem analiz w przedstawionej do recenzji pracy doktorskiej były dane uzyskane z pomiarów wykonanych techniką MALDI-MSI dla preparatów guzów pobranych od pacjentów z rozpoznaniem nowotworu regionu głowy i szyi. W trakcie typowego eksperymentu MSI dla pojedynczego preparatu tkankowego rejestruje się zazwyczaj (w zależności od wielkości preparatu i zastosowanej rozdzielczości bocznej) od tysiąca do kilkudziesięciu tysięcy widm. W przypadku projektów, w ramach których mierzonych jest wiele preparatów, analizowanych jest więc kilkaset tysięcy widm (obecnie w skrajnych przypadkach są to już miliony widm). Natomiast każde widmo opisane jest kilkunastoma lub kilkudziesięcioma tysiącami „punktów pomiarowych" (to również zależy od badanego

zakresu m/z i wybranej rozdzielczości spektrometru). Ilustruje to rozmiar danych generowanych w eksperymentach MSI; przykładowo, wielkość surowych danych zarejestrowanych dla jednego preparatu metodą MALDI-MSI to zazwyczaj od 2 do 20 GB. Dane rejestrowane w eksperymentach MSI stanowią więc interesujący rodzaj danych wielkoskalowych, tzw. „big data", a ich przetwarzanie rodzi oczywiste problemy obliczeniowe. Projekty wykorzystujące technikę MSI stanowią więc doskonały model do rozwoju i optymalizacji metod obliczeniowych oraz narzędzi bioinformatycznych i biostatystycznych. Podstawowe problemy, które wymagają rozwiązania w takich projektach to m.in.: (i) kwestia redukcji wymiarowości danych (uzyskiwana np. poprzez zastąpienie całego widma jego skwantyfikowanymi składowymi); (ii) kwestia obliczeniowej efektywności stosowanych algorytmów; (iii) kwestia optymalizacji metod tzw. nienadzorowanej analizy danych. Rozwiązania testowane i optymalizowane na potrzeby przetwarzania obrazów MSI mają więc oczywiście zastosowanie praktyczne w naukach biomedycznych. Ze względu na powyższe, tematyka ocenianej pracy doktorskiej, której podjął się mgr inż. Wojciech Sikora jest aktualna, a uzasadnienie wyboru problemu badawczego jest wystarczające i ma niewątpliwie aspekt praktycznego wykorzystania opracowanych rozwiązań.
Podstawowym założeniem techniki MSI jest możliwość rejestracji widm MS dla zdefiniowanych punktów „ciągłej" powierzchni preparatu biologicznego zamiast dla odrębnych preparatów (np. lizatów czy homogenów tkankowych traktowanych jako całość w podejściach standardowych). Po przeprowadzeniu takiej analizy dysponujemy informacją o poziomie setek/tysięcy analitów w setkach/tysiącach punktów zlokalizowanych w ściśle zdefiniowanych miejscach (koordynaty x i y) na powierzchni preparatu. Preparat tkankowy poddany analizie MSI może być później wybarwiony dla ujawnienia jego struktury morfologicznej w standardowej analizie mikroskopowej (ew. analizowane są tzw. skrawki seryjne), a następnie „obrazy molekularne" (czyli profile MS zarejestrowane dla punktów o koordynatach x i y) „nakładane" są komputerowo na mikroskopowe obrazy morfologiczne. Przestrzenna informacja o profilu molekularnym może być ujawniona w formie obrazów intensywności (tzw. „heatmaps") wybranych analitów, jednak pełne wykorzystanie informacji wymaga zastosowania zaawansowanych metod obliczeniowych typowych dla tzw. danych multispektralnych. Pozwala to na połączenie wielowymiarowych „obrazów molekularnych" z rzeczywistą informacją o morfologii i strukturze analizowanej tkanki.

Tematem pracy doktorskiej przedstawionej do recenzji było zweryfikowanie następujących hipotez: (1) identyfikacja jonów w widmach masowych może być skuteczna za pomocą modelowania całego spektrum, podzielenia go na części i modelowania go mieszaninami normalnymi; (2) informacja o dystrybucji przestrzennej usuwa redundancje i zmniejsza wymiarowość danych, przy jednoczesnym zachowaniu jakości; (3) identyfikacja istotnych („najważniejszych") cech jest możliwa dzięki wnioskowaniu na podstawie modeli jednostkowych. W prezentowanej pracy Doktorant wyjaśnił w powody, dla których zdecydował się na zastosowanie proponowanych w czasie realizacji projektu rozwiązań. W opinii recenzenta metody analizy danych, które zostały przez Autora pracy zaimplementowane do rozwiązania postawionych przed nim hipotez świadczą o dojrzałości naukowej, wymagają bardzo wszechstronnej wiedzy. Doktorant dysponuje wiedzą z zakresu analizy danych wielkoskalowych, ale również spektrometrii mas oraz znajomością sposobów akwizycji danych technikami MSI. Bardzo szczegółowo potraktował też Autor zakres opracowania teoretycznego

będącego podstawą realizacji jego pracy doktorskiej. W pierwszych dwóch rozdziałach znajdziemy opis podstawowych założeń oraz parametrów technicznych akwizycji danych. We wstępie pracy, mającym nieco inny charakter niż przyjęta w większości prac konwencja, Doktorant wyjaśnił w sposób niepozostawiający wątpliwości powody, dla których zdecydował o kolejnych etapach swoich badań i co ważniejsze powodach oraz sposobie ich realizacji. Dobór narzędzi badawczych do rozwiązania postawionego problemu oceniam bardzo wysoko. Do potknięć utrudniających odbiór pracy zaliczam kilka rycin, których główną wadą w mojej opinii jest nieoptymalna kolorystyka. Doktorant zdecydował się na tło części rycin prezentujących dystrybucję danej cechy w tym samym kolorze którego używa do prezentacji danych. Przykłady takich rycin to: figure1.3 strona 6, figure 3.1 strona 24, figure 7.9 do figure 7.12 od strony 90 do strony 93 (tło ryciny jest niebieskie, a na tym „podkładzie" kolorem jasnoniebieskim prezentowane są analizowane dane). Niektóre ryciny są w opinii Recenzenta zbędne na przykład figure 1.1. strona 5, figure 2.5 strona 21 oraz figure 2.6 strona 22. Paradoksalnie pozostałe ryciny zawarte w pracy doktorskiej prezentują bardzo wysoki poziom, co niestety jeszcze bardziej uwypukla niedostatki tych wymienionych. Podsumowując ocenę tematyki badań i jej realizacji przez mgr inż. Wojciecha Sikorę, przedstawioną do recenzji pracę oceniam bardzo dobrze. Mimo, że nie zabrakło potknięć w zaprezentowaniu wyników badań, to badania będące przedmiotem pracy zostały bardzo dobrze zaplanowane i zrealizowane.

Pod względem redakcyjnym pracę doktorską stanowi 8 rozdziałów, jej streszczenia w języku polskim i angielskim, spis ilustracji, spis tabel oraz 78 pozycji literaturowych. Rozprawa doktorska została zredagowana w języku angielskim i obejmuje 114 stron tekstu wraz z bibliografią. Układ pracy jest nietypowy, ale zawiera standardowe (mimo niestandardowych tytułów) dla tego typu dzieł naukowych części tj. wstęp, założenia i cel pracy, materiały i metody, opracowanie wyników wraz dyskusją oraz wnioski. Wstęp pracy doktorskiej stanowi 8 stron starannie omawiających cel pracy oraz tematykę kolejnych rozdziałów. Aktualny stan wiedzy związany z prowadzonymi przez doktoranta analizami stanowi rozdział 2. Założenia i cel pracy są przejrzyste i jasno napisane. Poszczególne rozdziały są napisane z dużą znajomością tematu. Doktorant wyjaśnia powody podjęcia stworzenia narzędzia analizy danych MSI, opisuje stosowane dotychczas metody oraz te które zaimplementował w swoich analizach/rozwiązaniach. W części pracy prezentującej wyniki analiz mamy liczne ryciny, które poza kilkoma podanymi wyżej sytuacjami bardzo dobrze ilustrują proces analizy danych oraz jej wyniki. Wnioski są przejrzyste i spójne, co oceniam jako dużą zaletę prezentowanej pracy. W pracy doktorskiej mgr inż. Wojciech Sikora posługuje się terminologią jon oraz pik stosując ją naprzemiennie. W mojej dziedzinie nauki tj. spektrometrii mas termin jon (ang. ion) oznacza jon molekularny w widmie masowym, z kolei termin pik (ang. peak) jest zarezerwowany dla terminologii dotyczącej chromatografii. Czy mogłabym prosić o wyjaśnienie podczas odpowiedzi na recenzje status quo obowiązujące w metodologii analizy danych? Jest również kilka innych kwestii (poniżej) do których chciałabym, aby Doktorant odniósł się podczas publicznej obrony pracy:

1. Czy Doktorant mógłby się odnieść do ostatniej z badanych hipotez i wyjaśnić w jaki sposób definiuje „ważność" cech?
2. Jaki warunek musi być spełniony by uznać jon za pojedynczo naładowany? W jaki sposób proponowane rozwiązanie „radzi sobie" z problemem nakładających się obwiedni izotopowych?

3. Jakie ograniczenia występują podczas usuwania linii bazowej?
4. Jaką definicję szumu przyjęto dla detekcji pików/jonów?
5. Jaką minimalną intensywnością całkowitego prądu jonowego (TIC) musi charakteryzować się widmo mas, aby mogło zostać poddane detekcji pików/jonów?
6. Jaka w opinii Doktoranta jest przyczyna wyznaczenia zupełnie innych cech dla klasyfikacji z wykorzystaniem sieci neuronowych metodami LIME oraz wartości Shapley'a? Czy na podstawie dystrybucji dla najlepszych cech można ocenić wiarygodność biologiczną uzyskanych wyników?
7. Czy dokonano interpretacji biologicznej cech wyznaczonych w procesie klasyfikacji?

Mgr inż. Wojciech Sikora rozpoczął swoją drogę naukową jako student Politechniki Śląskiej na Wydziale Mechanicznym Technologicznym, rozpoczynając studia inżynierskie na kierunku Mechatronika, które ukończył w roku 2012 z wynikiem bardzo dobrym. Studia inżynierskie ukończył pracą inżynierską na temat rozszerzania funkcjonalności środowiska do analizy danych RapidMiner o komponenty umożliwiające wstępne przetwarzanie danych sejsmicznych. Podczas studiów inżynierskich zainteresował się programowaniem, a swoje umiejętności rozwijał podczas kursów i staży (staż Sieć Badawcza Łukasiewicz - Instytut Technik Innowacyjnych). Naukę kontynuował rozpoczynając studia magisterskie na wydziale Automatyki, Elektroniki i Informatyki na kierunku Informatyka, które również ukończył z wynikiem bardzo dobrym. Po roku przerwy podczas którego rozwijał swoje umiejętności programisty tworząc oraz utrzymując systemy zarządzania współpracą międzyoperatorską dla firmy Orange Polska, postanowił kontynuować rozwój naukowy w zakresie analizy danych rozpoczynając interdyscyplinarne studia doktorskie Applied Integrative Data Analysis. W ramach studiów doktoranckich mgr inż. Wojciech Sikora zajął się tematyką analizy danych pochodzących z obrazowania molekularnego tkanek z udziałem spektrometrii mas wykorzystując do tego celu metody statystycznych oraz uczenia maszynowego. Podczas studiów doktoranckich mgr inż. Wojciech Sikora był współautorem monografii oraz brał udział w czterech zagranicznych konferencjach, gdzie prezentował wyniki swojej pracy.

Podsumowując, przedstawione przeze mnie w niniejszej recenzji uwagi i wątpliwości nie umniejszają mojej dobrej oceny całokształtu rozprawy, którą uważam za wartościową. Doktorant wykazał się dużą wiedzą teoretyczną w zakresie prowadzonych badań oraz umiejętnościami prowadzenia pracy badawczej. W mojej ocenie treść i jakość przedstawionej rozprawy w pełni odpowiadają wymogom stawianym rozprawom doktorskim. Rekomenduję zatem dopuszczenie przedstawionej do mojej oceny rozprawy doktorskiej do jej publicznej obrony.

Dr hab. Monika Pietrowska, profesor NIO