INSTYTUT CHEMII BIOORGANICZNEJ
Polskiej Akademii Nauk

ul. Noskowskiego 12/14, 61-704 Poznań
tel.: centrala 61 852 85 03, sekretariat 61 852 89 19
faks: 61 852 05 32, e-mail: ibch@ibch.poznan.pl
REGON 000849327 NIP 777-00-02-062
http://www.ibch.poznan.pl

Poznań, 24.07.2023

Dr hab. inż. Anna Wojakowska
Zakład Proteomiki Biomedycznej
Pracownia Spektrometrii Mas
Instytut Chemii Bioorganicznej PAN
e-mail: astasz@ibch.poznan.pl

RECENZJA

rozprawy doktorskiej mgr inż. Wojciecha Sikory zatytułowanej

„ Machine learning-based workflow for the analysis of MALDI-TOF mass spectrometry cancer data"

wykonanej w Katedrze Inżynierii i Analizy Eksploracyjnej Danych,
Wydział Automatyki, Elektroniki i Informatyki Politechniki Śląskiej,
pod kierunkiem prof. dr hab. inż. Joanny Polańskiej w roli promotora

Podstawa wykonania recenzji

Recenzję wykonano na zlecenie Rady Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej z dnia 29.05.2022 r.

Przedstawienie sylwetki kandydata do stopnia doktora

Pan Wojciech Sikora uzyskał tytuł magistra informatyki 16 listopada 2017 roku na Politechnice Śląskiej. Od 2018 roku realizował interdyscyplinarne studia doktoranckie w zakresie przetwarzania i analizy danych. Podczas studiów doktoranckich Pan Wojciech Sikora odbył staż naukowy, był współautorem publikacji naukowych oraz brał udział w zagranicznych konferencjach. Posiada również doświadczenie jako programista, tworzył i utrzymywał systemy zażądania współpracą międzyoperatorską dla firmy Orange Polska w roku 2016. W latach 2020-2022 zajmował stanowisko eksperta testera w ramach projektu z zakresu wykorzystania techniki sztucznej inteligencji realizowanego przez Giełdę Papierów Wartościowych w Warszawie.

Uzasadnienie podjęcia tematu rozprawy doktorskiej

Podjęta przez mgr inż. Wojciecha Sikorę tematyka pracy doktorskiej dotyczy opracowania rzetelnej ścieżki wstępnego przetwarzania danych proteomicznych zarejestrowanych techniką obrazowania molekularnego w oparciu o spektrometrię mas (MSI: *ang. - mass spectrometry imaging*). W wyniku prowadzonych eksperymentów obrazowania molekularnego metodami spektrometrii mas uzyskujemy znaczne ilości danych, zarejestrowane w formie widm masowych, charakteryzujące się dużą wymiarowością. Analiza danych uzyskanych przy wykorzystaniu techniki MSI jest znaczącym wyzwaniem od strony analizy bioinformatycznej zarejestrowanych widm masowych i obejmuje wiele następujących po sobie etapów, w tym korekcję linii bazowej, eliminację szumu, detekcję i ekstrakcję cech, dekonwolucję, integrację i wyrównanie danych, identyfikację obwiedni izotopowych. W wyniku przeprowadzonej wieloetapowej procedury przetwarzania widm masowych uzyskujemy zredukowaną

liczbę cech, odpowiadających składnikom molekularnym badanej tkanki, które następnie możemy korelować z jej cechami histopatologicznymi. Ma to szczególne znaczenie w przypadku analizy danych pochodzących od pacjentów nowotworowych, w przypadku których wyodrębnione cechy, stanowią potencjalnych kandydatów na diagnostyczne, prognostyczne lub predykcyjne sygnatury molekularne. Stad więc właściwie przeprowadzony proces przetwarzania danych zarejestrowanych w wyniku eksperymentów obrazowania molekularnego w konsekwencji prowadzi do poprawnej identyfikacji poszukiwanych składników molekularnych (białek, metabolitów czy lipidów) tkanek. Należy podkreślić, że taka analiza wymaga posiadania odpowiedniej wiedzy eksperckiej, właściwych skryptów dla każdego etapu procesowania danych oraz odpowiedniej mocy obliczeniowej komputera. Istnieje niewiele gotowych rozwiązań w postaci oprogramowania do preprocesingu danych zarejestrowanych technikami MSI, które by uwzględniały właściwe przeprowadzenie wszystkich niezbędnych kroków wstępnej analizy i przetwarzania widm z eksperymentów MALDI-MSI. Z tego względu, opracowanie kompleksowego rozwiązania służącego analizie danych z obrazowania molekularnego metodami spektrometrii mas stanowi ważny i uzasadniony problem badawczy.

Ocena merytoryczna pracy i jej struktury

Przedłożona do recenzji praca doktorska została napisana w języku angielskim. Ma postać manuskryptu obejmującego łącznie 114 stron i składającego się z następujących części: wprowadzenie wraz z celem pracy, przegląd literatury, materiały i metody, wyniki wraz z dyskusją, wnioski, wykaz bibliografii liczący 78 pozycje oraz streszczenie w języku angielskim i polskim. Ponadto, praca zawiera 68 rycin oraz 6 tabel. Brak jest wykazu użytych skrótów. Rozprawę czyta się dobrze, zarysowany jest wyraźny trend badawczy stanowiący następujące po sobie problemy obliczeniowe, których rozwiązanie kolejno przedstawia i interpretuje Doktorant w swojej pracy. Uwagę zwraca brak wyraźnie oddzielonej części teoretycznej od części badawczej, które to przeplatają się stanowiąc w każdym rozdziale swoista bazę i w dalszej kolejności odpowiedź dla podjętego wyzwania badawczego.

We wstępie został uzasadniony istotny problem podjętych badań, jako konieczność opracowania skutecznego zestawu narządzi służących kompleksowej analizie wielowymiarowych zestawów danych pochodzących z eksperymentów MALDI-MSI. Poprawne przeprowadzenie procesu wstępnej analizy i integracji danych z MSI, prowadzi do wyodrębnienia cech molekularnych, które skorelowane z danymi morfologicznymi i histopatologicznymi tkanki są podstawą do identyfikacji poszukiwanych sygnatur molekularnych.

Celem pracy było stworzenie narzędzia obliczeniowego, umożliwiającego zredukowanie wielowymiarowych danych uzyskanych z obrazowania molekularnego MALDI-MSI, oraz wyodrębnienie cech, które mogą być użyte do stworzenia klasyfikatora metodami uczenia maszynowego. Algorytmy testowano na czterech zbiorach danych, uzyskanych w wyniku przeprowadzenia eksperymentów obrazowania molekularnego MALDI-MSI na tkankach guzów pochodzących od pacjentów ze zdiagnozowanym nowotworem regionu głowy i szyi, leczonych w Narodowym Instytucie Onkologii w Gliwicach. W trakcie prowadzonych prac weryfikowano 3 hipotezy: (1) identyfikacja pojedynczych sygnałów w widmach masowych może być przeprowadzona za pomocą analizy całego widma, poprzez podzielenie go na części a następnie modelowaniu ich mieszaninami normalnymi (gaussowskimi); (2) informacja o przestrzennej dystrybucji danych z MSI, może być wykorzystana do zmniejszenia

ich wymiarowości przy jednoczesnym zachowaniu jakości danych; (3) identyfikacja najważniejszych cech dla danych heterogennych jest możliwa i skuteczna dzięki wnioskowaniu na podstawie wielu modeli jednostkowych.

Doktorant we wprowadzeniu słusznie zaznaczył istotną konieczność dogłębnej znajomości natury danych pochodzących ze spektrometrii mas w celu ich właściwej analizy. W rozdziale dotyczącym podstaw techniki MSI we właściwy sposób opisane są główne pojęcia z zakresu metod spektrometrii mas, które stosowane są w przypadku obrazowania molekularnego. Z drobnych nieścisłości, użyte zostało nieprecyzyjnie pojęcie „orbitrap" jako metoda, nie zaś analizator mas, którym w rzeczywistości jest, jak dalej opisuje to Doktorant na stronie 21. Zabrakło również podkreślenia istotnej informacji z punktu widzenia natury analizowanych w rozprawie danych, dotyczącej rodzaju jonów jakie powstają w trakcie prowadzenia eksperymentów proteomicznych z wykorzystaniem jonizacji typu MALDI. Ponadto rozdział zawiera niewielką liczbę odnośników literaturowych, w podrozdziale 2.3 „Ionization" brak jakiegokolwiek odwołania do literatury tematu. Brakuje również referencji dotyczących rycin, co dotyczy większości rozprawy. W odniesieniu do braku wykazu skrótów, nie wszystkie z nich wyjaśnione są w tekście np. AFADESI (strona 17).

W kolejnych rozdziałach Doktorant opisuje poszczególne etapy procesingu danych zarejestrowanych techniką MALDI-MSI dla próbek tkanek pochodzących od 4 pacjentów z rozpoznanym rakiem regionu głowy i szyi. W poszczególnych rozdziałach dotyczących kolejno rejestracji danych, preprocesingu, modelowania widma, inżynierii cech oraz konstrukcji klasyfikatorów Doktorant każdorazowo rozpoczyna wprowadzeniem teoretycznym, omawia problemy napotykane na każdym etapie analizy danych oraz przedstawia swoją rzetelną metodą rozwiązania danego problemu. Przejście między poszczególnymi składowymi rozdziałów było bardzo płynne i w niektórych przypadkach recenzent musiał wykazać się zwiększoną uważnością aby wychwycić moment gdzie kończy się teoria a zaczyna właściwa praca Doktoranta. Tym niemniej taki sposób przedstawienia wyników okazał się doskonałym pomysłem w obliczu mnogości wyzwań napotkanych przez Doktoranta podczas analizy danych oraz etapów koniecznych do jej przeprocesowania. Z punktu widzenia recenzenta schemat przedstawiający poszczególne etapy procesowania danych ułatwiłby podążanie przez kolejne etapy analizy oraz zilustrował zastosowane kompleksowe podejście do analizy danych zarejestrowanych metodą obrazowania molekularnego MALDI-MSI. Należy jednak podkreślić, że schemat przedstawiony na końcu Rozdziału 6 obrazuje drogę jaką pokonał Doktorant począwszy od surowych danych, poprzez modelowanie widma, filtrowanie sygnałów o niskiej abundancji, wyrównanie sygnałów oraz deizotoping, wraz z redukcją stopnia wymiarowości danych. W wyniku zastosowanych metod modelowania i analizy widm uzyskano znaczną redukcję wymiarowości danych z ponad 100 000 kanałów masowych do 888 cech. Analiza została przeprowadzona dla 4 próbek tkanek. Czy taki schemat może być z powodzeniem zastosowany dla większej liczby prób zarejestrowanych w podobnych warunkach analitycznych. Czy każdorazowo wymagany jest etap optymalizacji metody obliczeniowej?

W końcowych rozdziałach Doktorant opisuje proces uczenia klasyfikatorów na danych przetworzonych oraz dokonuje ich oceny w oparciu o metody badania ważności cech LIME i wartości Shapley'a. Doktorant potwierdził że, identyfikacja najważniejszych cech (potencjalnych biomarkerów) dla danych heterogennych jest możliwa i skuteczna dzięki wnioskowaniu na podstawie wielu modeli jednostkowych. Czy podczas tworzenia klasyfikatora

na etapie podziału danych na zestawy uczący, testowy i walidacyjny były brane pod uwagę % udziały poszczególnych typów tkanki (nowotwór vs norma)?

Podsumowanie

Opracowane przez Doktoranta kompleksowe narzędzie do analizy dużych zbiorów danych z obrazowania molekularnego metodami spektrometrii mas stanowi istotny wkład w rozwój metod wstępnej analizy i przetwarzania danych MS. Metoda modelowania widma mieszaninami gaussowskimi oraz zastosowane metody inżynierii cech, skutkujące znaczną redukcją wymiarowości danych oraz wyodrębnieniem istotnych cech (jonów), stanowi oryginalne narzędzie analityczne dla danych uzyskanych techniką MALDI-MSI. Przedstawioną do recenzji rozprawę doktorską oceniam pozytywnie i doceniam naukową dojrzałość, dociekliwość i ogrom pracy włożonej przez doktoranta na każdym etapie jej realizacji. Pan mgr inż. Wojciech Sikora wykazał się wiedzą z zakresu charakteru analizowanych danych i właściwym warsztatem obliczeniowym, które pozwoliły na stworzenie oryginalnego rozwiązania problemu badawczego.

Wnioski

W mojej opinii przedstawiona praca spełnia kryteria stawiane rozprawie doktorskiej określone w aktualnie obowiązujących regulacjach prawnych. Wnoszę do Rady Naukowej Inżynieria Biomedyczna Politechniki Śląskiej o dopuszczenie mgr inż. Wojciecha Sikorę do dalszych etapów postępowania o nadanie stopnia doktora.

Dr hab. inż. Anna Wojakowska

Anna Wojakowska