Prof. dr hab. Michał Ciborowski
Uniwersytet Medyczny w Białymstoku
Centrum Badań Klinicznych
Laboratorium Metabolomiki i Proteomiki

RECENZJA

rozprawy doktorskiej mgr. Łukasza Prussa pt.:

„Bioinformatic analysis in targeted metabolomics, automating the process of translating raw NMR spectrum signals into qualitative and quantitative data for use in the analysis of disease states"

wykonanej w Katedrze Biochemii, Biologii Molekularnej i Biotechnologii Politechniki Wrocławskiej pod kierunkiem prof. dr. hab. Piotra Młynarza oraz dr Kai Milanowskiej-Zabel pełniącej rolę promotora pomocniczego.

Przedłożona mi do oceny rozprawa doktorska przedstawia opracowanie otwartego i ogólnodostępnego narzędzia bioinformatycznego do przetwarzania jednowymiarowych protonowych widm NMR, które zostało zwalidowane za pomocą danych metabolomicznych z otwartego repozytorium, obejmujących widma surowicy pacjentów z dysautonomią rodzinną i ich zdrowych krewnych.

Badania metabolomiczne to złożony, wieloetapowy proces przetwarzania i analizy danych. Programy do opracowywania surowych danych spektralnych są często tworzone przez producentów urządzeń pomiarowych, a w konsekwencji są zwykle kompatybilne jedynie z formatem danych generowanych przez urządzenia danego producenta. Ponadto, nierzadko programy te są płatne (jednorazowo lub wymagają wykupienia subskrypcji czy płatnych aktualizacji). Innym aspektem jest wszechstronność takich narzędzi. Dostępne narzędzia zwykle koncentrują się na pojedynczych etapach opracowywania czy analizy danych, np. przetwarzaniu widm, identyfikacji metabolitów, analizie statystycznej czy analizie szlaków metabolicznych. Brakuje ogólnodostępnego kompleksowego narzędzia, przy pomocy którego otrzymane dane spektralne można poddać przetwarzaniu, analizie statystycznej oraz analizie ścieżek biologicznych.

Przedstawione przed Doktoranta narzędzie *Nextflow Automatization and Standardization for Qualitative and Quantitative ^{1}H NMR Metabolomic*, w skrócie NASQQ, integruje istniejące rozwiązania do przetwarzania widm, opracowywania danych, identyfikacji i kwantyfikacji metabolitów, jak również jedno-

i wielowymiarowej analizy statystycznej oraz analizy ścieżek biologicznych. Stworzone narzędzie, w którym wykorzystywane są m.in. zaawansowane modele uczenia maszynowego, standaryzuje metody przypisywania sygnałów, redukuje błędy operacyjne i zwiększa powtarzalność, a także skutecznie łączy surowe dane widmowe z interpretacją biologiczną. Do walidacji narzędzia Doktorant użył publicznie dostępnego zestawu danych z repozytorium Metabolights, czym wpisuje się w inicjatywę FAIR, która określa jakie wymogi powinny spełniać dane badawcze. Akronim FAIR oznacza: „findable" (możliwy do znalezienia), „accessible" (dostępny), „interoperable" (interoperacyjny) i „reusable" (możliwy do ponownego wykorzystania). Na uwagę zasługuje również fakt, że rozprawa powstała w ramach programu Ministra Nauki i Szkolnictwa Wyższego „doktorat wdrożeniowy", którego celem jest tworzenie warunków do rozwoju współpracy podmiotów systemu szkolnictwa wyższego i nauki z otoczeniem społeczno-gospodarczym. Wspomniana współpraca została nawiązana pomiędzy Politechniką Wrocławską, a spółką akcyjną Ardigen SA, specjalizującą się w badaniach *in silico*, łączących bioinformatykę ze sztuczną inteligencją.

Oceniana rozprawa została napisana w języku angielskim, obejmuje 161 stron maszynopisu i podzielona została na cztery główne rozdziały, tj.: wstęp, metodologia, wyniki i dyskusja. Ponadto, na początku rozprawy umieszczone zostały: streszczenia w języku angielskim i polskim, spis treści, wykaz skrótów oraz minirozdział „Motivation and thesis outline", zaś na końcu spis figur i tabel, lista równań, spis literatury oraz materiały uzupełniające.

W minirozdziale „Motivation and thesis outline" mgr Łukasz Pruss przedstawił swoją motywację stojącą za podjęciem się stworzenia tego narzędzia bioinformatycznego. Wyznaczył cel główny oraz trzy cele szczegółowe, które precyzują jakie rozwiązania zostaną wprowadzone by zrealizować cel główny. Doktorant postawił również trzy hipotezy, które zamierzał zbadać. W dalszej części przedstawiony został zarys informacji zawartych w poszczególnych rozdziałach głównych oraz lista publikacji, które powstały jako część pracy doktorskiej.

Wstęp składa się z czterech podrozdziałów. W pierwszym z nich Doktorant opisał miejsce metabolomiki w biologii systemowej, historię rozwoju metabolomicznych, skupiając się na wprowadzeniu poszczególnych technik analitycznych i rozwoju dedykowanych narzędzi obliczeniowych, a także przedstawił koncepcję działania jądrowego rezonansu magnetycznego (NMR). Następnie skupił się na budowie spektrometru NMR i przedstawił jednowymiarową analizę protonową NMR. Wstęp urozmaicony został wieloma schematami, równaniami matematycznymi i rycinami, które ułatwiają odbiór prezentowanych treści.

Odrębny podrozdział poświęcony został metabolomice i jej niezaprzeczalnej roli w badaniach nad różnymi chorobami. Opisane zostały badania nad chorobami autoimmunologicznymi i opornością na antybiotyki w kontekście złożonego współdziałania predyspozycji genetycznych i czynników środowiskowych, a także potencjał łączenia różnych omik w celu kompleksowego podejścia badawczego. Znalazł się tu również krótki opis dysautonomi rodzinnej, rzadkiej choroby genetycznej charakteryzującej się upośledzonym rozwojem neuronów i postępującą degeneracją. Przybliżenie czytelnikowi tej choroby wynika z tego, że dane metabolomiczne ^{1}H NMR zarejestrowane dla próbek osocza pacjentów z dysautonomią rodzinną oraz grupy kontrolnej (zdrowi krewni) zostały wykorzystane do kompleksowego przetestowania narzędzia NASQQ. Wspomniane dane zostały pobrane z ogólnodostępnego repozytorium.

W dalszej części wstępu przedstawione zostały metody obliczeniowe stosowane w metabolomice. Autor pokrótce opisał genezę bioinformatyki, języki programowania stosowane do tworzenia narzędzi bionformatycznych używanych do analizy danych metabolomicznych, a także przedstawił schematy pracy z danymi oraz przykłady systemów zarządzania przepływem pracy stosowanych w bioinformatyce, w tym Nextflow, z którego Doktorant korzystał przy tworzeniu NASQQ. Następnie opisane zostały poszczególne procesy wstępnego przetwarzania danych NMR i potencjalne podejścia statystyczne z uwzględnieniem metod uczenia maszynowego i *deep learning*. Doktorant wymienił popularne algorytmy klasyfikacji uczenia maszynowego, takie jak: lasy losowe czy KNN, a także techniki walidacji krzyżowej. Na koniec tego podrozdziału Doktorant omówił analizę ścieżek biologicznych, m.in. przy użyciu analizy wzbogacenia czy grafów wiedzy. Przedstawił również potencjalne możliwości łączenia danych omicznych i ich wspólną interpretację.

Ostatnia część wstępu omawia najnowocześniejsze bazy danych i narzędzia bioinformatyczne. Doktorant najpierw przedstawił w tabeli dostępne bazy danych, oprogramowania i pakiety R, a następnie skupił się na najpopularniejszych repozytoriach danych NMR, oprogramowaniach do analizy widm NMR, zarówno komercyjnych, jak i tych ogólnodostępnych (*open source*) oraz ich funkcjonalnościach. Znalazła się tu również wzmianka o genezie inicjatywy FAIR.

Rozdział metodologiczny mgr Łukasz Pruss rozpoczął od opisu stworzonego narzędzia i graficznego przedstawienia harmonogramu opracowywania i analizy danych z podziałem na poszczególne etapy całego procesu. Następnie przeszedł do implementacji Nextflow i szczegółowego opisu stworzonych modułów NASQQ dla trzech głównych etapów: „Przetwarzanie widm", „Analiza danych" i „Interpretacja biologiczna". Doktorant przy tworzeniu narzędzia wykorzystał istniejące już pakiety R i skrypty Pythona (kompletna lista komponentów oprogramowania została przedstawiona w tabeli), jednakże zaznacza, że główna innowacja polegała na stworzeniu spersonalizowanych skryptów od podstaw oraz ich dostosowanie do integracji z architekturą Nextflow. Na każdym etapie pracy przeprowadzano testy na rzeczywistych danych

laboratoryjnych, zarówno własnych jak i tych pobranych z otwartych repozytoriów potwierdzając solidność i niezawodność tworzonych protokołów. Warto podkreślić, że wszystkie skrypty oraz moduły są publicznie dostępne w repozytorium NASQQ GitHub.

W kolejnym rozdziale Doktorant opisał zastosowanie NASQQ do opracowania i analizy pobranych z bazy Metabolights danych metabolomicznych ^{1}H NMR z projektu dot. dysautonomii rodzinnej. Szczegółowo przedstawił poszczególne etapy analizy danych zestawiając wyniki w postaci 28 rycin i 5 tabel (w tym część jako materiały uzupełniające). Ze względu na rzadkość występowania tej choroby było to międzynarodowe badanie wieloośrodkowe, w ramach którego wygenerowanych zostało pięć pakietów surowych danych pomiarowych. Mgr Łukasz Pruss przeprowadził etap przetwarzania danych oddzielnie dla każdej partii danych, a następnie zbadał czy w danych tych występuje tzw. *batch effect*, czyli różnice w poziomach sygnałów wynikające z analizy próbek w różnych sekwencjach analitycznych, a nie z różnic biologicznych. Doktorant zauważył, że próbki grupują się ze względu na partię danych, a poziomy kilku metabolitów istotnie różnią się w zależności od tego w jakiej sekwencji analitycznej były oznaczane. Wpływ sekwencji analitycznej na dane jest znanym problemem badań metabolomicznych, a aby go zniwelować stosowane są różne metody korekcji danych. W przypadku NASQQ wykorzystany został wielowymiarowy moduł zaprojektowany do obsługi dużych zestawów danych metabolomicznych obejmujących różne punkty czasowe, stany pacjentów i choroby. Pozwala on na określenie względnego znaczenia metabolitów przy użyciu wartości Shapleya, co pomaga odróżnić istotne metabolity od artefaktów. Analizy statystyczne pozwoliły na uzyskanie łącznie 56 istotnych statystycznie metabolitów, z których 20 najbardziej istotnych statystycznie wg. analizy wielowymiarowej zostało poddanych analizie szlaków biologicznych. Analiza ta wskazała na potencjalne zaburzenia takich szlaków jak cykl Krebsa, metabolizm argininy i proliny, metabolizm beta-alaniny, szlak sygnałowy glukagonu, transportery ABC, szlak sygnałowy mTOR i metabolizm węgla, a także metabolizm węgla w szlaku nowotworowym.

Rozdział „Dyskusja” również składa się z podrozdziałów. W pierwszym z nich Doktorant podsumował stworzone narzędzie i jego użyteczność do analizy danych NMR. Wskazał, że NASQQ to kompleksowe narzędzie do opracowywania i analizy danych NMR, które zapewnia kontrolę nad każdym etapem procesu analitycznego. Zapewnia ono dostęp do tabel przetworzonych danych i wizualnych reprezentacji w postaci rysunków, umożliwiając użytkownikom łatwe śledzenie postępów i identyfikowanie rozbieżności na poszczególnych etapach analizy. Porównanie pierwotnie przetworzonych widm uzyskanych dla próbek surowicy pobranych od uczestników badania nad dysautonomią rodzinną z widmami uzyskanymi przy użyciu przygotowanego narzędzia ujawniło, że przy odpowiedniej konfiguracji parametrów, NASQQ może generować wyniki o porównywalnej jakości. Odnotowano natomiast rozbieżności w identyfikacji metabolitów, która

wynikała z wykorzystania innych bibliotek referencyjnych. Nie wszystkie metabolity występujące w bibliotece wykorzystanej przez autorów publikacji pokrywały się z metabolitami znajdującymi się w bibliotece użytej przez Doktoranta, stąd brak pełnej spójności w zidentyfikowanych metabolitach. Mam w tym miejscu pytanie do Doktoranta. Czy stosowaną przez Pana bibliotekę można uzupełniać o kolejne metabolity? Następnie Autor podkreślił konieczność równoległego stosowania modeli uczenia maszynowego, co pozwala zwiększyć czułość i swoistość wykrywania metabolitów, umożliwiając identyfikację szerszego zakresu związków biologicznie istotnych oraz integrację informacji na temat wielu metabolitów w celu bardziej wiarygodnego zidentyfikowania zaburzonych szlaków metabolicznych. Podsumowując, autor wskaza, że wykorzystanie NASQQ pozwoliło na dokładniejsze poznanie zmian w szlakach metabolicznych związanych z występowaniem dysautonomii rodzinnej, co przybliża nas do zrozumienia funkcjonalnego tła tej choroby.

Autor wskazał również na ograniczenia stworzonego narzędzia. Pomimo jego ogólnodostępności, NASQQ działa obecnie tylko w jednym formacie danych (Bruker) oraz jest ograniczony do analiz osocza, surowicy czy moczu. Ponadto, dane muszą być odpowiednio dobrane (jednorodny program impulsów), a część parametrów w modułach jest zautomatyzowana i nie ma możliwości ich dostosowywania w czasie rzeczywistym. Kolejnym ograniczeniem jest biblioteka referencyjna, która obecnie obejmuje jedynie 191 związków, a także niespójność informacji zawartych w bibliotece z danymi KEGG, na podstawie których wykonywana jest analiza szlaków metabolicznych, co skutkuje wykluczeniem części metabolitów z analizy szlaków. Doktorant wskazał również na niewykorzystany, jego zdaniem, potencjał NextFlow do jednoczesnego wykonywania wielu procesów, a także na brak interfejsu graficznego i dokładnych wyjaśnień mechaniki operacyjnej NextFlow w repozytorium GitHub, co może początkowo utrudniać mniej zaawansowanym technicznie użytkownikom efektywne korzystanie z dostępnych rozwiązań.

Następny podrozdział Doktorant poświęcił na omówienie dalszego rozwoju stworzonego narzędzia. Rozszerzenie potoku analitycznego o możliwość obsługiwania innego formatu bądź rodzaju danych metabolomicznych, utworzenie spersonalizowanej biblioteki referencyjnej, czy rozszerzenie o narzędzia i podejścia multiomiczne to tylko niektóre przykłady podane przez Doktoranta.

Na koniec Doktorant odniósł się do postawionych celów i hipotez rozprawy. Wskazał, że założony cel główny oraz cele poboczne zostały spełnione, o czym świadczy skutecznie przeprowadzona zautomatyzowana analizy danych metabolomicznych ^{1}H NMR przy użyciu zaawansowanych metod bioinformatycznych i uczenia maszynowego składających się na protokół NASQQ.

Bibliografia, którą posiłkował się Doktorant przygotowując niniejszą rozprawę obejmuje 209 pozycji piśmiennictwa, w większości są to prace anglojęzyczne, a prawie 98% z nich zostało opublikowanych w ostatnim dwudziestoleciu. Oprócz artykułów naukowych oraz monografii cytowane są liczne odnośniki do stron internetowych zawierających instrukcje, biblioteki bądź repozytoria modułów, pakietów czy oprogramowania.

Podsumowując, uważam, iż rozprawa doktorska Pana mgr. Łukasza Prussa odpowiada na aktualne potrzeby w obszarze badań wielkoskalowych, a szczególnie metabolomiki. Opracowany protokół bioinformatyczny uzupełnia lukę w kontekście dostępności do narzędzia pozwalającego na kompleksowe opracowanie danych ^{1}H NMR, od surowych danych do analizy i interpretacji biochemicznej. Jednocześnie ogólnodostępność opracowanego rozwiązania jest ważnym aspektem w kontekście dążenia do w pełni otwartej nauki. Jak zauważył Doktorant, pomimo obecnych ograniczeń, NASQQ może być użytecznym narzędziem, szczególnie do analiz między-laboratoryjnych. Planowane dalsze ulepszenia jedynie zwiększą wartość proponowanego rozwiązania bioinformatycznego.

Oceniana rozprawa jest bardzo dobrze napisana pod względem merytorycznym, jednak z poziomu recenzenta muszę wspomnieć o zauważonych błędach stylistycznych i powtórzeniach. Niektóre z błędów edytorskich wynikają zapewne z dynamiki w życiu (zawodowym i prywatnym), która dotyka obecne społeczeństwo. Dostrzeżone błędy to np.: strona 23 – trzy zdania dotyczące wyników z przedstawianej publikacji są zduplikowane. Kilkukrotnie też pojawiały się powtórzenia sformułowań w następujących po sobie zdaniach. Pragnę jednak podkreślić, że wymienione drobne błędy edytorskie nie podważają wartości merytorycznej przedstawionej pracy i przeprowadzonych badań.

Biorąc pod uwagę całość rozprawy doktorskiej stwierdzam, iż Rozprawa doktorska mgr. Łukasza Prussa spełnia wymogi stawiane rozprawom doktorskim zgodnie z art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. 2023 poz. 742), w związku z czym zwracam się do Rady Dyscypliny Naukowej Nauki Chemiczne Politechniki Wrocławskiej z wnioskiem o nadanie mgr. Łukaszowi Prussowi stopnia doktora w dziedzinie nauk ścisłych i przyrodniczych w dyscyplinie nauki chemiczne.

Ponadto, ze względu na potencjał aplikacyjny stworzonego narzędzia i jego nowatorski charakter, zwracam się z prośbą do Rady Dyscypliny Naukowej Nauki Chemiczne Politechniki Wrocławskiej z wnioskiem o wyróżnienie rozprawy.

Białystok, dn. 22.11.2024r.

Z-ca DYREKTORA
ds. Badań Metabolomicznych i Proteomiki
Centrum Badań Klinicznych
C Borowski
prof. dr hab. Michał Ciborowski