prof. dr hab. Krzysztof Kazimierczuk

email: k.kazimierczuk@cent.uw.edu.pl

Warszawa 18.11.2024

Recenzja rozprawy doktorskiej Pana mgr. inż. Łukasza Prussa:

"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."

Powołanie na recenzenta otrzymałem na podstawie uchwały Rady Dyscypliny Naukowej Nauki Chemiczne Politechniki Wrocławskiej z 23 września 2024 r. Przedłożona do recenzji dysertacja została zrealizowana w ramach programu „Doktorat wdrożeniowy" Ministerstwa Nauki i Szkolnictwa Wyższego. W pracy został zaproponowany program NASQQ, będący potokiem przetwarzania danych z eksperymentów metabolomicznych realizowanych przy użyciu spektroskopii magnetycznego rezonansu jądrowego (NMR).

Metabolomika jest jednym z zastosowań spektroskopii NMR, które bardzo gwałtownie rozwinęły się w XXI wieku. Wielkoskalowe badania składu próbek pochodzenia biologicznego, np. moczu czy surowicy krwi, pochodzących od setek pacjentów pozwalają na wykrycie markerów poszczególnych chorób i powiązanie ich z zaburzeniami konkretnych szlaków metabolicznych. Naukowcy pracujący w tej dziedzinie tworzą również metody klasyfikacji pacjentów oparte na traktowaniu widm jako „odcisków palca", bez szczegółowej interpretacji zmian składu próbek. W obu scenariuszach konieczne jest zastosowanie zaawansowanych metod przetwarzania sygnałów i analizy statystycznej wyników. Ponieważ w jednym badaniu rejestruje się widma setek próbek, potrzebne są

zautomatyzowane metody przetwarzania danych. Na rynku istnieje kilka odpowiednich rozwiązań, takich jak MetaboAnalyst czy Chenomx. Pan mgr inż. Łukasz Pruss, autor rozprawy, postawił sobie za cel stworzenie konkurencyjnego pakietu o otwartym i ogólnodostępnym kodzie. Cel był ambitny, gdyż rozwiązanie miało być kompletne i realizować wszystkie kroki przetwarzania i analizy danych, od transformacji surowego sygnału ze spektrometru do wizualizacji istotnych ścieżek metabolicznych.

Opracowane przez mgr. inż. Prussa rozwiązanie jest oparte na środowisku Nextflow i integruje działanie różnych modułów wykorzystywanych do przetwarzania, analizy statystycznej i interpretacji danych. Wybór środowiska i zastosowanych modułów jest poprzedzony wstępem, dobrze opisującym i uzasadniającym użyte metody. Ważną część pracy zajmuje opis zastosowania programu do analizy zestawu widm surowicy krwi pobranej od pacjentów chorych na dysautonomię rodzinną (ang. *Familial Dysautonomia*) i ich zdrowych krewnych.

Miałem okazję przetestować program w wersji 1.0 oraz 1.1 w środowisku Linux i wysoko oceniam jego działanie. Samą rozprawę doktorską uważam za dobrze napisaną i dowodzącą znajomości dziedziny, choć mam sporo zastrzeżeń dotyczących wstępu teoretycznego podstawy spektroskopii NMR. Zasady stosowania metod zawartych w programie są dobrze opisane i uzasadnione. Praca zawiera doskonale napisany wstęp, prawidłowe odnośniki literaturowe (również w podpisach do ilustracji). Pragnę podkreślić, że treść pracy wykracza daleko poza opis działania programu. Nie jest „instrukcją obsługi", ale ciekawą rozprawą o tym, jakie funkcje i dlaczego powinno posiadać oprogramowanie do przetwarzania danych metabolomicznych. Mogę ją polecić każdemu, kto chciałby podjąć się podobnego zadania lub zrozumieć na czym polega taka analiza.

Na szczególne uznanie zasługują rozdziały I.3 "Computational methods in metabolomics" oraz I.4 "Databases and state-of-the-art tools", które stanowią przegląd dostępnych rozwiązań. Podobało mi się również uczciwe przedstawienie ograniczeń rozwiązania w rozdziałach IV.1 „General conclusions of pipeline usage for metabolomic analysis" i IV.2 „Utilized methodologies limitations".

Oprócz pochwał, mam jednak również kilka krytycznych uwag ogólnych oraz szereg szczegółowych spostrzeżeń dotyczących drobnych niedociągnięć.

Po pierwsze, program powinien być zwalidowany przy użyciu zestawu danych, dla którego znane są prawidłowe wyniki analizy statystycznej. Mógłby to być nawet zestaw stworzony sztucznie. Dane, których użyto do zademonstrowania działania programu, tzn. widma surowicy krwi chorych na dysautonomię rodzinną pochodzą z publikacji Cheney et al. i w oryginalnej pracy zostały przeanalizowane za pomocą programu Chenomx NMR Suite z nieco innym wynikiem. Przykładowo, autorzy wykryli mocznik i ksantynę, których obecności NASQQ nie potwierdził. Trudno powiedzieć, czyj wynik jest właściwy. Doktorant bardzo krótko komentuje rozbieżności, pisząc (na stronie 112), że nie jest celem pracy powtórzenie wyników innych zespołów. Pozostawia to pewien niedosyt.

Po drugie, w pracy zabrakło przedstawienia możliwości wdrożenia programu na szeroką skalę. Wydaje się, że doktorat wdrożeniowy powinien być przyczynkiem do takiej działalności, zwłaszcza że tak duży pakiet programistyczny wymaga sporego zaangażowania czasowego w wykrywanie błędów, wsparcie użytkowników i dalszy rozwój. Z doświadczenia wiem, że podtrzymywanie działania programów w długiej perspektywie wymaga sił i środków, często przekraczających możliwości jednej osoby. Chętnie dowiedziałbym się, czy i jakie są plany używania NASQQ w przemyśle oraz jego dalszego rozwoju.

Po trzecie, program działa dość dobrze przy standardowych ustawieniach parametrów przetwarzania danych, ale użytkownik, który chciałby je zmienić, natrafia na pewne trudności. Przykładowo, nie da się wybrać metody fazowania sygnału ani funkcji apodyzacji. Wydaje mi się, że tego typu modyfikowalnych parametrów w pliku params.yml mogłoby być więcej. Ponadto sądzę, że moduł przetwarzania danych można by było z powodzeniem zastąpić biblioteką nmrglue, powszechnie używaną do tego celu. Jeśli istnieją jakieś przeciwskazania, np. warunki licencji, to dobrze byłoby umożliwić posługiwanie się w NASQQ widmami (nie sygnałami FID) przetworzonymi w innych programach. Pewne zaawansowane rozwiązania, np. filtry sygnału rozpuszczalnika, automatyczne fazowanie itp. mogą być w nich lepsze, niż w programie autora.

Szczegółowe błędy, które zauważyłem w rozprawie, wymieniam poniżej.

1) Na str. 8 autor przedstawia hipotezy badawcze. Jedną z nich jest „*1) At each stage of the automated pipeline, there is a set of parameters that significantly affect the accuracy of the*

results obtained". Jeśli traktować to twierdzenie ściśle jako hipotezę badawczą to wydaje się ona trywialna, a jej weryfikacja niewiele wnosząca.

2) Rozdział I.1 dotyczący podstaw fizycznych spektroksopii NMR uważam za najmniej dopracowaną część dysertacji. Opis jest dość ogólny i zawiera kilka konkretnych błędów:

   a) W równaniu 1.1 B_0 to indukcja („*induction*") pola magnetycznego, a nie „siła" („*strength*") jak pisze autor.

   b) Równanie 1.2 jest błędne, powinno być: $\mu = \gamma I$

   c) Zdanie: „For a given nucleus, the spins can exist in two states: + ½ and − ½, corresponding to two potential orientations of the magnetic moment" jest nieprawdziwe (spin istnieje w wielu stanach, to pomiar zwraca tylko dwie wartości) oraz neizrozumiałe w tym miejscu tekstu (autor dopiero dalej pisze, że rozważa szczególny przypadek spinu 1/2).

   d) Nie jest jasne co oznaczają różowe i okrągłe czarne strzałki na Rys. 3. Podpis pod rysunkiem jest zbyt skrótowy. Cały rysunek jest zresztą nieścisły merytorycznie. Na którym z rysunków z odnośnika [34] autor go oparł?

   e) Autor pisze i pokazuje na Rys. 4, że pochłonięcie energii przez moment magnetyczny powoduje przejście ze stanu o spinie +1/2 do stanu o spinie -1/2. Jest to nieco mylące, gdyż w ogólności może być odwrotnie (zależnie od współczynnika żyromagnetycznego).

   f) *dt* w równianiu 1.4 nie jest „odstępem między punktami czasu w sygnale FID", jak twierdzi autor.

   g) Zdanie „*Surrounding the magnet is a vacuum chamber that maintains the required environment, including liquid nitrogen and liquid helium for cooling purpose*" jest niejasne. Czy komora próżniowa zawiera ciekły azot i hel?

   h) Nazwa sekwencji impulsowej CPMG jest podana bez rozwinięcia (na dodatek z literówką).

i) Na str. 22 pojawia się twierdzenie: „*However, due to its complexity, deconvolution is often considered an auxiliary approach in spectroscopic studies.*” Dekonwolucja jest dość prostą operacją, jednak niestabilną numerycznie

3) Przy opisie programu TopSpin firmy Bruker warto dodać, że istnieją spektrometry tej firmy dedykowane do metabolomiki, z dedykowanym oprogramowaniem (tzw. platforma IVDr)

4) Dobrze byłoby dodać wyjaśnienie czym jest „container”.

5) Zdanie w Tabeli 4: „*Increase spectral signal-to-noise ratio*” – powinno być uzupełnione o koszt (szerokość linii), zaś. „*Improve the visual representation of spectra by adding zeros”* o wyjaśnienie, gdzie się te zera dodaje.

6) W rozdziale II.2.3 w stwierdzeniu „*calculating the smoothed version of the FID.”* należy zamienić „FID” na „spectrum”.

7) Bardzo brakuje korekty fazy pierwszego rzędu. Oczywiście algorytmy jej automatycznej korekty są dość złożone. Myślę, że jest to jeden z powodów, dla których należałoby dopuścić możliwość użycia zewnętrznych modułów do przetwarzania danych

8) W rozdziale II.2.8 należy zastąpić stwierdzenie „*with their standard peaks traditionally set at 0 ppm.”* przez „*with their methyl group peaks traditionally set at 0 ppm.”*

9) Pod równaniem 2.0 pojawia się stwierdzenie, że β_k to współczynniki wielomianu. Chodzi chyba o β (bez k)?

10) W równaniu 2.2 jest błąd, czy X_{ij} również miało być w operacji liczenia mediany? W przeciwnym razie skraca się.

11) Nie jest jasne, czym w równaniu 2.3 jest β bez indeksu.

12) Nie rozumiem, dlaczego skala czasu na rysunkach z sygnałami FID przedstawiana jest w tysiącach mikrosekund, a nie w milisekundach.

Powyższe drobne niedociągnięcia w żaden sposób nie ujmują znaczenia pracy. Rozwój metabolomiki w Polsce jest ważny i dobrze, że prace takie jak niniejsza powstają. Stworzenie powszechnie dostępnego pakietu analizy danych NMR z eksperymentów metabolomicznych przyczyni się do rozwoju dziedziny i wpłynie stymulująco na rozwój konkurencyjnych pakietów oprogramowania.

Stwierdzam, że rozprawa doktorska spełnia warunki określone w art. 187 ust. 1-2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2023 r. poz. 742 z późn. zm.).

Z poważaniem,

prof. dr hab. Krzysztof Kazimierczuk

Podpisany elektronicznie przez
zysztof Kazimierczuk; Uniwersytet Warszaws
18.11.2024
13:44:21 +01'00'