SZKOŁA GŁÓWNA GOSPODARSTWA WIEJSKIEGO W WARSZAWIE

16.02.2020

dr hab. Krzysztof Pawłowski
Katedra Biochemii i Mikrobiologii,
Instytut Biologii,
SGGW,
ul. Nowoursynowska 166,
02-776 Warszawa

**RECENZJA**

**rozprawy doktorskiej**
**mgr Mai Małkowskiej**
**zatytułowanej**
**„Opracowanie metody analizy sekwencji polimorficznych i ich funkcjonalności w regionach promotorowych.”**

Rozprawa doktorska mgr Mai Małkowskiej, przygotowana w centrum Onkologii – Instytucie im. Marii Skłodowskiej-Curie w Warszawie pod kierunkiem promotora, dr hab. Lucjana Wyrwicza, opisuje rozwinięcie ciekawej metody przewidywania funkcjonalnego znaczenia mutacji w promotorach genów.

Nie jest zadaniem recenzenta przekonywanie Rady Naukowej, jak ważne w biologii jest zrozumienie zależności między zmiennością genetyczną a regulacją genów. Opisane w rozprawie badania mgr Małkowskiej koncentrują się na metodach bioinformatycznych pomagających badać zmienność sekwencji promotorowych genów. Jest to obecnie bardzo gorący temat badań, wg bazy PubMed tematyki tej dotyczy ponad siedem tysięcy prac. Praca doktorantki jest bardzo wartościowym krokiem w tej dziedzinie kierunku. Wyniki składające się na rozprawę doktorską zostały opublikowane w dobrym i znanym czasopiśmie *PeerJ* (2018, IF=2,3). Mgr Małkowska jest pierwszym autorem tej publikacji. Praca ta uzyskała dotąd jedno cytowanie (wg Google Scholar). Oprócz tej publikacji, mgr Małkowska jest także współautorką artykułów w *Biochimie* (2013, IF=3,2) *Virology Journal* (2013, IF=2,5), *JIMD Reports* (2015, IF=4,3), które łącznie uzyskały 43 cytowania. W takiej sytuacji, gdy wyniki rozprawy zostały sprawdzone przez niezależnych recenzentów międzynarodowych czasopism, rola niżej podpisanego recenzenta jest dość ograniczona.

Rozprawa napisana jest w języku polskim. Rozprawa zawiera około trzydziestostronicowy wstęp teoretyczny, kilkustostronicowy opis metod, około dziesięciostronicowe omówienie wyników i zestawienie literatury (blisko 200 pozycji). W pracy znajduje się ponadto suplement opisujący zmienne zastosowane w modelu oraz obszerniejszy suplement na płycie DVD zawierający bardziej szczegółowe wyniki numeryczne, które nie zmieściły się w rozprawie.

Należy podkreślić, że wszystkie dane oraz kod oprogramowania są udostępnione w repozytorium GitHub. Udostępnianie narzędzi jest obecnie szeroko praktykowanym i bardzo cennym wkładem w rozwój metod bioinformatycznych.

W pracy autorka zastosowała zaawansowane metody uczenia maszynowego. Wykorzystała jako treningowe i testowe zbiory danych duże listy znanych polimorfizmów, których wpływ na patogenezę chorób został potwierdzony eksperymentalnie. Użyty też został niezależny walidacyjny zbiór danych. Do przewidywania znaczenia funkcjonalnego polimorfizmów doktorantka wykorzystała dużą liczbę cech otoczenia polimorfizmu, zarówno określonych doświadczalnie, jak i wynikających wprost z sekwencji DNA bądź obliczonych na jej podstawie. Po zastosowaniu algorytmu selekcji cech okazało się, iż najważniejsze dla przewidywań funkcjonalności są cechy lokalnej struktury DNA. Porównując różne klasyfikatory uczenia maszynowego, stwierdzono, iż najlepsze wyniki osiąga algorytm ekstremalnego gradientowego wzmacniania. Autorka, opracowawszy swoją metodę, dokonała także jej porównania z dostępnymi narzędziami oceny funkcjonalności polimorfizmów w regionach niekodujących genomu. Porównanie to wypadło zdecydowanie na korzyść nowej metody.

Recenzent nie dopatrzył się w rozprawie treści zasługujących na poważną krytykę. Praca jest napisana sprawnym i dość dobrym językiem, ma klarowny układ. Zastrzeżeń merytorycznych recenzent ma niewiele i nie są one bardzo ciężkiego kalibru. We wstępie autorka definiując pojęcie genomu mówi, że jest to ”ogół informacji genetycznej… zapisanej w postaci DNA”, nie wspominając o genomach RNA. Również we wstępie, autorka dzieli ludzkie DNA na geny, elementy regulatorowe, oraz „śmieciowe” DNA. Wypadałoby tu zaznaczyć, nie tylko cudzysłowem, że od ponad dekady wiadomo, iż tzw. „śmieciowe” DNA ma ważne funkcje biologiczne, choć nie są one tak dobrze rozumiane, jak funkcje pozostałych obszarów genomu. Omawiając technikę mikromacierzy DNA autorka pomija zasadnicze zastosowanie tej metody, tj. do analizy porównawczej ekspresji genów w wielu próbkach (a nie tylko w dwóch).

Recenzent radby też przeczytać we wstępie więcej, co wiadomo globalnie o polimorfizmach pojedynczego nukleotydu (SNP) w genach, obszarach regulatorowych

i pozostałych obszarach, np. czy występują różnice w ilości polimorfizmów o znaczeniu funkcjonalnym.

Ponadto do obowiązków recenzenta należy zwrócenie uwagi na pewne niezręczności i błędy językowe, np.:

„w śród" zamiast „wśród"

„populacja Chińska" zamiast „populacja chińska"

„hybrydyzacja Nothern" zamiast „„hybrydyzacja Northern"

„wyrywane" zamiast „wykrywane"

Niektóre zdania są gramatycznie zawiłe i trudno zrozumiałe, np.: „W sytuacjach, gdy mamy do czynienia z dużymi, mocno rozrośniętymi drzewami są duże i mocno rozrośnięte" (strona 18). Również pierwsze zdanie w podrozdziale 1.4.2. jest niezrozumiałe (cztery skomplikowane zdania podrzędne). Niekiedy omyłkowe stosowanie terminów sprawia, że czytelnik nie wie, co autorka miała na myśli, np. „warianty położone na pojedynczym chromosomie tworzyły zbiór treningowy, a pochodzące z innego zbiór treningowy."

Nie zawsze też recenzent zgadza się z tłumaczeniem terminów anglojęzycznych. Prawidłowe rozwinięcie skrótu NCBI powinno brzmieć Narodowe Centrum Informacji Biotechnologicznej, a nie „Narodowe Centrum Biotechnologicznej Informacji" (kolejność wyrazów ma znaczenie). Podobnie, na str. 64 pojawiają się „derywaty", co zapewne miało oznaczać „pochodne" (ang. *derivatives*). Termin „uliniowienia sekwencji" (zapewne tłumaczone z „*sequence alignments*"), choć pojawiający się w polskojęzycznych tekstach, jest zdaniem recenzenta błędny i nielogiczny. Proponowałbym „dopasowania sekwencji"

Dwa razy recenzent natknął się na błędy stylistyczne, niestety występujące często w mowie potocznej: „w przeciągu ostatnich pięciu lat" i „na dzień dzisiejszy".

Ponadto, w podpisie rys. 6 pojawia się zmienna „swoistość", która nie występuje w tekście. Omawiając metody nie wyjaśniono, dlaczego w modelu uwzględniano zmienne dotyczące zawartości par GC, skoro wg autorki nie było istotnych różnic między zbiorami testowym i pozytywnym i negatywnym. Informacja o zmiennych związanych z wiązaniem czynników transkrypcyjnych jest niejasna, np.: „wykorzystano klastry V3", „suma zmiennych wynosiła 12". Niejasne jest też stwierdzenie „16 zmiennych wyrażających obserwowaną a oczekiwaną częstość wystąpień" (str. 41).

Ostatnia grupa uwag recenzenta dotyczy kwestii, w których odczuł on pewien niedosyt, a właściwie niezaspokojoną ciekawość. Przyznać jednak należy, że dyskusja wyników jest rzetelna i wyczerpująca.

Ciekawe jest natomiast, i chętnie recenzent usłyszałby na obronie, czy od czasu złożenia publikacji mgr Małkowskiej do recenzji (ponad 2 lata temu) pojawiły się w literaturze nowe informacje pozwalające skonfrontować wyniki autorki z danymi doświadczalnymi? Czy przynajmniej dla części polimorfizmów, dla których przewiduje ona z dużą istotnością znaczenie funkcjonalne, jest możliwość potwierdzenia przewidywań?

Recenzent jest także ciekaw, czy metoda autorki może być stosowana do innych niż człowiek organizmów? W budowie modelu wykorzystano dane o zachowaniu ewolucyjnym sekwencji wśród ssaków. Czy zdaniem autorki byłoby rozsądne wykorzystanie metody do analiz sekwencji promotorowych innych kręgowców? Ewentualnie innych zwierząt, bardziej odległych ewolucyjnie?

Powyższe uwagi i drobne zastrzeżenia nie umniejszają oceny pracy. Recenzent ocenia rozprawę mgr Małkowskiej bardzo wysoko. Doktorantka wykazała się umiejętnością twórczej, samodzielnej pracy naukowej i przedstawiła wartościowe rozwinięcie nowej metody bioinformatycznej. Dorobek naukowy i rozprawa doktorska mgr Mai Małkowskiej w pełni spełniają ustawowe i zwyczajowe wymagania do ubiegania się o stopień naukowy doktora. Wnoszę zarazem o dopuszczenie doktorantki do dalszych etapów przewodu.

Krzysztof Pawłowski