Wrocław, 18.09.2022 r.

Prof. dr hab. inż. Małgorzata Kotulska
Politechnika Wrocławska
Wydział Podstawowych Problemów Techniki
Katedra Inżynierii Biomedycznej
E-mail: malgorzata.kotulska@pwr.edu.pl

Recenzja rozprawy doktorskiej: *mgr Anny Macioszek*

zatytułowanej

HMM-based method for identifying enrichment in signal from sequencing-based experiments

1. Charakterystyka rozprawy

Przedstawiona mi do oceny praca doktorska pani Anny Macioszek powstała w Instytucie Informatyki na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego, pod kierunkiem dra hab. Bartosza Wilczyńskiego. Rozprawa wpisuje się w obszar badań z dziedziny bioinformatyki, ma charakter teoretyczno-obliczeniowy, z bardzo silnym ukierunkowaniem na rozwiązanie praktyczne i jego implementację.

Praca doktorska została napisana w języku angielskim, liczy 126 stron i podzielona jest na 7 rozdziałów. Rozprawę otwiera wprowadzenie teoretyczne. Autorka przedstawia w nim problem biologiczny badań nad strukturą chromatyny, aktualnie stosowane metody eksperymentalne oraz obliczeniowe (rozdział 1). W rozdziale 2 omówiony jest szczegółowy problem badawczy, któremu poświęcona jest dysertacja. Przedstawione są w nim metody obliczeniowe stosowane w algorytmach typu „*peak-calling*", umożliwiające znajdowanie obszarów sekwencji o dużym znaczeniu funkcjonalnym. W końcówce tego rozdziału, w podrozdziale „*Challenges*" (rozdział 2.3), Autorka wymienia nierozwiązane dotychczas problemy aktualnie stosowanych metod obliczeniowych. Rozdział ten można traktować jako zdefiniowanie celów pracy doktorskiej. Rozdział 3 przedstawia propozycję autorskich zmian w metodyce obliczeniowej i szczegóły implementacji oprogramowania. Rozdziały 4-6 poświęcone są wynikom testów nowej metodologii z uwzględnieniem różnych typów danych, symulowanych i eksperymentalnych. W rozdziale 7 Autorka krótko podsumowuje wyniki swojej pracy.

2. Ocena przedstawionej rozprawy, komentarze i pytania

Wkład autorki przedstawionej rozprawy można podzielić na dwie zasadnicze części: krytyczny przegląd i omówienie dostępnych metod obliczeniowych oraz wyniki własnych prac badawczych.

Część teoretyczna pracy, omówienie problemu i przegląd stosowanych metod pokazuje wystarczająco głęboką wiedzę Autorki dotyczącą tematyki podjętej w rozprawie doktorskiej. Część przeglądowa jest przygotowana starannie, zarówno pod względem ogólnej kompozycji, zawartości, jak i językowo oraz edycyjnie.

W części badawczej pracy Autorka skupiła się na ulepszeniu metodyki, na której oparte są narzędzia do „*peak-calling*". Jako uzasadnienie wyboru takiego celu pracy podaje niewystarczającą skuteczność stosowanych obecnie metod, zwłaszcza w przypadku niektórych typów sekwencji poddanych modyfikacji, na przykład tworzących bardzo długie domeny obejmujące po kilkadziesiąt genów zawartych w pojedynczym piku. W danych mogą się też pojawić inne problemy, dodatkowo utrudniające analizę. Przykładowo, H3K27me3, na którym skupiła się Autorka, nie tylko rozciąga się na bardzo długim odcinku, ale również wykazuje znacznie słabszy sygnał, co wynika z bardzo zwartej struktury chromatyny. Zadanie wykrycia tej domeny dodatkowo utrudnia fakt dynamicznych zmian jej położenia w genomie, różniącego się pomiędzy komórkami z tej samej populacji, co skutkuje mało wyrazistymi i trudnymi do detekcji granicami badanego obszaru. Dla tego typu zadań potrzebne są specjalistyczne i dedykowane narzędzia.

Podjęta w pracy tematyka badawcza jest ważna a efektem bardziej skutecznej analizy sekwencji może być lepsze zrozumienie mechanizmów regulacji ekspresji genów, ich wpływu na funkcjonowanie i zdrowie całego organizmu. Problem, który rozwiązuje doktorantka ma więc duży potencjał naukowy i reprezentuje aktualną i ważną tematykę badawczą.

Rozwiązania prezentowane przez Autorkę pracy spełniają założenia postawione w celu pracy. Autorka wykorzystała podejście oparte na ukrytych modelach Markowa, w których (w odróżnieniu od stosowanych w innych metodach *dwóch* stanów ukrytych) wprowadziła możliwość wystąpienia trzeciego stanu ukrytego „*brak sygnału*", reprezentującego dane mało charakterystyczne. Dzięki zastosowanej metodyce uwzględniona jest też możliwość wystąpienia zależności pomiędzy sąsiadującymi fragmentami, co wydaje się uzasadnione w przypadku danych, dla których tworzona była ta metoda. Opracowane przez Autorkę narzędzie daje też możliwość elastycznego doboru stosowanych parametrów obliczeniowych, na przykład pozwalając na wybór pomiędzy dwoma typami rozkładów (rozkład Gaussa lub ujemny rozkład dwumianowy), a w efekcie możliwość wyboru lepszego modelu rzeczywistych

danych. Efektem wprowadzonych przez Autorkę innowacji jest poprawa dokładności wyników analizy. Skuteczność prezentowanej metody okazała się wyższa od algorytmów zaimplementowanych w dotychczas używanych narzędziach. Szczególnie dotyczyło to danych o długich i mało wyrazistych domenach.

Mam jednak kilka uwag krytycznych dotyczących ocenianej pracy. Najistotniejsza z nich dotyczy sposobu prezentacji przeprowadzonych badań naukowych. Autorka bardzo mało wyraziście prezentuje główne założenia teoretyczne proponowanej przez siebie metody. Rozdział 3, który ją przedstawia, zawiera najpierw teoretyczny opis standardowych metod matematycznych zastosowanych w rozwiązaniu (rozdział 3.1 i jego podrozdziały). Następnie Autorka od razu przechodzi do rozdziału zatytułowanego „*Implementation*" (rozdział 3.2), gdzie w sposób dosyć wymieszany przedstawia równocześnie, i dość ogólnie, wszystkie zaproponowane przez siebie innowacje. Brakuje dyskusji ich genezy, merytorycznego omówienia pomysłów, a także analizy spodziewanych efektów – przedstawionych odrębnie dla każdego nowego elementu metodyki. Następnie, w kolejnych podrozdziałach, Autorka prezentuje praktyczną instrukcję obsługi swojego programu, poświęcając na to sporo miejsca. Zabrakło mi wyraźnego sformułowania tez (lub hipotez) badawczych pracy, które nie miałyby od razu charakteru konkretnej implementacji. Podobnie, w rozdziale *Summary* nie znalazło się szczegółowe omówienie i dyskusja nowych aspektów metody i możliwych konsekwencji ich wprowadzenia, a jedynie stwierdzenie, że program z autorskimi modyfikacjami metody działa lepiej niż inne narzędzia przeznaczone do tego celu.

W dalszej części pracy Autorka testuje swoja metodę na danych symulowanych i eksperymentalnych. Skuteczność porównuje z wynikami uzyskanymi przez inne dostępne metody. Analiza jest przeprowadzona dość wyczerpująco, uwzględniając rozmaity charakter danych, które mogą się pojawić. Jednak również tutaj zabrakło mi analizy na poziomie przyczynowości, w której wyraźnie byłoby sformułowane, które konkretnie zmiany w metodyce powodują pożądany efekt przy określonej charakterystyce analizowanych danych i jak silny jest wkład każdej innowacji (oddzielnie) na poprawę dokładności metody, w porównaniu do analizowanych metod konkurencyjnych. Oczywiście istotny jest łączny efekt działania wprowadzonych ulepszeń, jednak analiza ukierunkowana, bardziej precyzyjna i szczegółowa, umożliwiłaby zrozumienie czemu metody Autorki są lepsze. Pozwoliłaby też badaczom na rozwój innych nowych metod dedykowanych określonym danym. Warto byłoby też wskazać które narzędzia i zaimplementowane w nich metody mogą okazać się najlepsze w przy pewnej określonej charakterystyce danych, i dlaczego tak można przypuszczać. Czy zawsze to będzie metoda Autorki?

Pomimo powyższych uwag krytycznych uważam jednak, że praca doktorska przyczyniła się do rozwoju bioinformatyki i stosowanych w niej metod badawczych. Do najważniejszych osiągnięć Autorki zaliczam opracowanie nowego wariantu metodologii stosowanej w „*peak-calling*", zaimplementowanego i przetestowanego za pomocą autorskiego oprogramowania *HERON*, który umożliwia skuteczniejszą analizę danych o mniej typowej charakterystyce. Wyniki prac badawczych Autorki zostały opublikowane w dwóch artykułach naukowych. Prace bezpośrednio związane z celem doktoratu zostały przedstawione w czasopiśmie *International Journal of Molecular Sciences*. Ponadto, doktorantka jest współautorką publikacji w *Nature Communications*, w której uczestniczyła w analizie elementów regulatorowych chromatyny.

3. Podsumowanie

Stwierdzam, że mgr Anna Macioszek zaprezentowała rozprawę doktorską rozwiązującą aktualny problem naukowy, jak również przyczyniła się do rozwoju reprezentowanej dyscypliny naukowej. Autorka potwierdziła umiejętność prowadzenia samodzielnej pracy naukowej, jak również wystarczającą ogólną wiedzą teoretyczną, niezbędną do uzyskania stopnia doktora z w dyscyplinie informatyka.

Biorąc pod uwagę omówione powyżej elementy oceny, stwierdzam, że rozprawa doktorska mgr Anny Macioszek spełnia wymagania określone w Ustawie z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (tekst jednolity: Dz.U. 2022 r. poz. 574 z późn. zm.) w sprawie szczegółowego trybu i warunków przeprowadzenia czynności w przewodach doktorskich, postępowaniu habilitacyjnym oraz w postępowaniu o nadanie tytułu profesora.

Wnioskuję o dopuszczenie autorki do kolejnych etapów przewodu doktorskiego oraz do publicznej obrony przedstawionej rozprawy.

Małgorzata Kotulska