Wrocław, 14.06.2023 r.

Prof. dr hab. inż. Małgorzata Kotulska

Politechnika Wrocławska

Wydział Podstawowych Problemów Techniki

Katedra Inżynierii Biomedycznej

Recenzja rozprawy doktorskiej: *mgr inż. Patryk Jarnot*

"Methods for similarity analysis of low complexity regions in protein sequences"

1. Charakterystyka rozprawy

Przedstawiona mi do oceny praca doktorska Patryka Jarnota powstała na Wydziale Automatyki, Elektroniki i Informatyki, Politechniki Śląskiej, pod kierunkiem dr hab. inż. Aleksandry Grucy. Drugim promotorem rozprawy był dr hab. Marcin Grynberg. Rozprawa wpisuje się w obszar badań z dziedziny bioinformatyki, ma charakter teoretyczno-obliczeniowy z uwzględnieniem praktycznych implementacji zaproponowanych rozwiązań. Autor opracował nowe metody i algorytmy do wykrywania i analizy podobieństwa obszarów o niskiej złożoności (LCR, ang. *Low Complexity Regions*) i obszarów z powtórzeniami (STR, ang. *short tandem repeat*).

Praca doktorska została napisana w języku angielskim, liczy 137 stron i podzielona jest na 8 rozdziałów. W rozdziale, który otwiera rozprawę, Autor definiuje cele pracy doktorskiej i krótko przedstawia swój dorobek naukowy, który powstał w wyniku jej realizacji. Rozdziały drugi, trzeci i czwarty to wprowadzenie teoretyczne do podjętej tematyki. Najpierw Autor przedstawia w nich metody dedykowane rozpoznawaniu krótkich powtórzeń i obszarów o niskiej złożoności, w których znalazły się: fLPS, CAST, SEG, LCD-Composer, SIMPLE, T-REKS, XSTREAM, GBA, a także klasyczne metody porównywania sekwencji, takie jak BLAST, HHblits, CD-HIT, MMseqs i MCL. Następnie, w rozdziale 3, omawia problem obszarów o niskiej złożoności i porównuje skuteczność i powtarzalność wyników z różnych metod im dedykowanych. W rozdziale 4 omawia klasyczne metody porównywania sekwencji, stosowane do wyszukiwania białek homologicznych i problemy z ich zastosowaniem w przypadku obszarów o niskiej złożoności. W rozdziale 5, Autor przedstawia rozwiązanie, które mogłoby wspomóc zastosowanie BLASTa w takim przypadku, wprowadzając nową metodę nazwaną LCR-BLAST. W rozdziale 6, Autor prezentuje kolejną, całkowicie oryginalną metodę przeznaczoną do wykrywania krótkich powtórzeń, opartą na grafowym podejściu do klastrowania sekwencji. Jej zadaniem jest umożliwienie bardziej precyzyjnego rozpoznawania obszarów sekwencji białkowych z powtórzeniami, w tym również takich o niskiej złożoności. W rozdziale 7 zostały zaprezentowane przykłady zastosowania nowych metod opracowanych przez doktoranta oraz wyniki skuteczności ich działania

na rzeczywistych przykładach biologicznych. Rozdział 8 zamyka rozprawę, podsumowując jej najważniejsze cele oraz osiągnięte wyniki.

2. Problem badawczy i jego znaczenie

Problem jakim zajął się doktorant w swojej rozprawie wpisuje się w nurt najbardziej aktualnych badań naukowych. Obszary o niskiej złożoności okazały się znacznie bardziej istotnym elementem sekwencji białkowych niż dawniej przypuszczano i nie zawsze korzystnie jest je ignorować lub wręcz maskować w analizie filogenetycznej, tak jak to na ogół ma miejsce. Obszary te mogą bowiem spełniać w białku określone funkcje. Badania wykazały, że często powstają, lub pozostają nieusunięte i zoptymalizowane, w efekcie kontrolowanej ewolucji. Zespół, w którym pracował doktorant zauważył, że mimo dostępności wielu metod bioinformatycznych, metody dedykowane wykrywaniu i analizie obszarów o niskiej złożoności nie dają jednoznacznych wyników. Brakuje też systematycznej analizy obszaru stosowalności metod, które potencjalnie mogą być w tym celu wykorzystane. Ponadto, klasyczne metody porównywania sekwencji, służące do znajdowania homologów lub segmentów wskazujących na analogiczne funkcjonalności białek, często w analizie pomijają obszary LCR, a wykorzystanie ich do tego celu wymaga bardzo przemyślanego doboru parametrów i procedury użycia. Dlatego też cele, które realizował doktorant i wyniki uzyskane w rozprawie są ważne i mają bardzo duże potencjalne zastosowanie.

3. Ocena przedstawionej rozprawy, komentarze i pytania

Część teoretyczna pracy, omówienie problemu i przegląd stosowanych metod pokazują wystarczająco głęboką i aktualną wiedzę Autora dotyczącą tematyki podjętej w rozprawie doktorskiej. Bibliografia pracy, zawierająca 137 pozycji literaturowych, jest aktualna i adekwatnie dobrana. Potwierdza to wysoki poziom wiedzy oczekiwany od kandydata na stopień doktora.

Część badawczą rozprawy stanowią trzy powiązane ze sobą projekty. Pierwszy z nich to systematyczny przegląd, analiza oraz porównanie skuteczności dostępnych metod stosowanych w analizie obszarów o niskiej złożoności. Autor wykazuje, że metody nie dają jednoznacznych wyników, często nieco odmiennie wskazując poszukiwane segmenty białkowe, a szczególnie położenie ich granic. Na ogół wynika to z różniących się założeń odnośnie obszaru zastosowania każdej z tych metod, Przyjętych długości i wierności powtórzeń. Różnice te prowadzą też do odmiennych profili aminokwasowych, uzyskanych z poszczególnych metod. Rozwiązaniem praktycznym, ułatwiającym uzyskanie wyniku o dużym stopniu pewności, może być wprowadzenie metody opartej na konsensusie. Autor przedstawił wyniki szczegółowych analiz, poparte przykładami na konkretnych sekwencjach. Prace te zostały uwieńczone opracowaniem własnego narzędzia PlaToLoCo, umożliwiającego analizę sekwencji kilkoma różnymi metodami, z wizualizacją wyników i możliwością wyboru rozwiązania konsensusowego. Narzędzie jest przyjazne w użyciu, z czytelną prezentacją

wyników, i opublikowane w bardzo dobrym czasopiśmie. Ta część pracy doprowadziła do istotnego rezultatu, ułatwiając świadomy wybór narzędzia optymalnego w konkretnej analizie i najbardziej adekwatnego wyniku.

W kolejnej części pracy, Doktorant przeanalizował działanie klasycznych narzędzi do porównywania sekwencji i poszukiwania białek homologicznych. Zauważył, że w domyślnych konfiguracjach świadomie pomijają lub zaniedbują w swojej analizie obszary o niskiej złożoności. Przeprowadził analizę wyników z ich działania z uwzględnieniem rozmaitych konfiguracji parametrów. Ostatecznie wykazał, że są one bardzo mocno oparte na obszarach o wysokiej złożoności (HCR, ang. *High Complexity Regions*") i segmenty typu LCR mają niewielki wpływ na wyniki ich pracy. Zaproponował więc zestaw parametrów i modyfikację macierzy oceny, które podnoszą wrażliwość BLASTa na obszary inne niż HCR. W analizach porównawczych umożliwi to wykorzystanie fragmentów LCR, zwłaszcza krótkich, zwykle eliminowanych ze względu na podwyższenie wartości *e-value*. Uważam, że jest to bardzo potrzebne rozwiązanie i warto byłoby je włączyć również do klasycznego BLASTa w formie narzędzia internetowego.

Ostatni projekt, zawarty w pracy, dotyczy opracowania metody GBSC (*Graph Based on Sequence Clustering*). Metoda ma na celu identyfikację STR w sekwencjach białkowych i grupuje je według podobnych wzorców w sekwencji, wykorzystując podejście oparte na grafach. Dodatkowo może identyfikować i grupować powtórzenia za pomocą zredukowanych alfabetów. Pozornie metoda ta przypomina podejście zastosowane w klasycznej metodzie CD-HIT, jednak grupowanie przebiega inaczej, co daje możliwość innych analiz, pokazujących inne zależności. To bardzo przydatna nowa metoda, która pozwoli wykorzystać specyficzne alfabety zredukowane, które nie muszą się opierać na klasycznych własnościach fizykochemicznych, ale będą mogły odnosić się do specyficznych wzorców odpowiadających różnym szczegółowym problemom.

Ostatecznie, Autor testuje dwie własne metody na białkach z bazy UniProtKB/Swiss-Prot. Testy przeprowadza najpierw na wybranym motywie aminokwasowym, a następnie na sekwencjach białkowych z proteomu wirusa SARS-CoV-2. Analiza ta pokazuje istnienie w proteomach drobnoustrojów obszarów o niskiej złożoności. Autor wskazuje jak brzemienne w skutki może być ich przeoczenie, co łatwo może się zdarzyć, jeżeli korzysta się wyłącznie ze standardowych narzędzi do porównywania sekwencji, na przykład przy opracowaniu szczepionki. To istotna część pracy, która dowodzi poprawności i stosowalności wyników uzyskanych przez Doktoranta.

Część badawcza rozprawy w spójny pokazuje systematyczną analizę działania istniejących metod oraz wkład Autora w ich rozwinięcie oraz tworzenie nowych metod, jak również ich implementacja w postaci gotowych narzędzi. Zaproponowane metody znalazły zastosowanie w opracowanych programach, z których część jest już udostępniona na portalach internetowych, takich

jak wspomniany już portal PlaToLoCo. Powstał też GBSC, przeznaczony do klastrowania krótkich powtórzeń w białkach, w których często występują obszary o niskiej złożoności. Ponadto, Autor opracował zmodyfikowaną wersję BLASTa, tak aby można go było bardziej skutecznie użyć do rozpoznawania i porównywania krótkich powtórzeń o niskiej złożoności. Część badawcza jest przekonująca, a opis wyników klarowny. Wyniki zostały opublikowane w pięciu artykułach w czasopismach naukowych i jednym artykule konferencyjnym. Lista zawiera czasopisma najlepsze z dziedziny, takie jak *Nucleic Acis Research*, *Journal of American Chemical Society*, *Briefings in Bioinformatics* i *BMC Bioinformatics*. Świadczy to o wysokim poziomie zaprezentowanych badań.

Jakkolwiek praca mi się bardzo podoba, to mam też do kilka uwag lub pytań:

- Wydaje mi się, że lepiej byłoby przyjąć nieco inną kolejność treści początkowych rozdziałów. Mianowicie, lepiej byłoby najpierw zdefiniować terminologię i problemy dotyczące obszarów LCR (obecnie rozdział 3), a dopiero później przedstawić listę analizowanych narzędzi (rozdział 2). Szczególnie dotyczy to klasycznych metod dopasowywania sekwencji, które pojawiają się w rozdziale 4, i dopiero wtedy staje się zrozumiałe, dlaczego zostały wcześniej przedstawione.

- Autor analizuje zarówno narzędzia dedykowane wykrywaniu LCR, jak i przeznaczone do wykrywania STR lub TR. Nie wyjaśnia w jakich sytuacjach takie obszary można traktować jednakowo, oczekując od narzędzi podobnych wyników. Warto byłoby jawnie wprowadzić definicje LCR, STR, homopolimerów, i obszarów o niezrównoważonej zawartości aminokwasowej („*compositionally biased*"), wyjaśniając wzajemne zależności pomiędzy nimi oraz precyzując, kiedy te pojęcia oznaczają to samo, a kiedy coś odrębnego. Nie wszystkie muszą być obszarami typu LCR. Chciałabym też, żeby Autor wyjaśnił, czy w analizach odfiltrowywał np. obszary TR, które nie są LCR.

- Analiza różnic między wynikami dotychczas dostępnych metod, które służą do wykrywania obszarów LCR lub STR byłaby pełniejsza, gdyby wzbogacić ją o analizę statystyczną. Szczególnie dotyczy to profili aminokwasowych.

Przedstawione powyżej uwagi nie umniejszają mojej wysokiej oceny rozprawy.

4. Podsumowanie

Stwierdzam, że mgr inż. Patryk Jarnołt zaprezentował bardzo rzetelną i dobrze napisaną rozprawę doktorską rozwiązującą aktualny problem naukowy, która przyczyni się do rozwoju reprezentowanej dyscypliny naukowej. Rozprawa zawiera oryginalne rozwiązanie problemu naukowego, a kandydat wykazał, że zarówno posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka jak i umiejętność prowadzenia pracy naukowej na wysokim poziomie. Jej wysoki poziom potwierdzają publikacje w najlepszych czasopismach z dziedziny, w których Doktorant jest współautorem. Chciałabym też

podkreślić, że praca doktorska jest przygotowana bardzo starannie, zarówno edycyjnie, jak i językowo. Bardzo dobrze się ją czyta – jest przejrzysta i interesująco napisana.

Biorąc pod uwagę omówione powyżej elementy oceny, stwierdzam, że oceniana rozprawa doktorska 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.

Niniejszym, wnioskuję o dopuszczenie autora do kolejnych etapów przewodu doktorskiego oraz do publicznej obrony przedstawionej rozprawy.

Ponadto, biorąc pod uwagę bardzo wysoki poziom merytoryczny rozprawy, dodatkowo potwierdzony publikacjami w najlepszych czasopismach z dziedziny oraz praktyczną implementacją opracowanych metod, wnioskuję o wyróżnienie rozprawy.

Małgorzata Kotulska