RDITT-wpi. 07. 07. 2023
M Skwarek

Dr hab. inż. Paweł Piotr Łabaj
Małopolskie Centrum Biotechnologii
Uniwersytet Jagielloński
Gronostajowa 7a, 30-387 Kraków
pawel.labaj@uj.edu.pl

Kraków, 15.06.2023

Recenzja rozprawy doktorskiej

Tytuł rozprawy:	Methods for similarity analysis of low complexity regions in protein sequences
Autor rozprawy:	mgr inż. Patryk Jarnot
Promotor rozprawy:	dr hab. inż. Aleksandra Gruca, prof. P.Ś.
Dziedzina:	nauki inżynieryjno-techniczne
Dyscyplina:	Informatyka Techniczna i Telekomunikacja

Regiony o niskiej złożoności (ang. Low Complexity Regions - LCR) to fragmenty białek o niewielkiej różnorodności aminokwasów. Nazwa ta jest niestety nieprecyzyjna i w piśmiennictwie występują jej różne wariancje i interpretacje. Powoduje to, że charakterystyka nie tylko jakościowa ale też ilościowa różni się znacznie w zależności od przyjętej nomenklatury. To co powoduje, że są to obszary warte zainteresowanie to fakt, że pomimo, że wyglądają jak proste sekwencje zawierające bardzo mało informacji to są one powszechne, a zatem powinniśmy uważać je za ważne fragmenty białek. Ponadto liczne badania w ostatnim czasie wykazały ich znaczenie w kontekście strukturalnego układu białka, a co za tym idzie jego właściwości funkcjonalnych. Niestety, przez długi czas uważano, że LCR są biologicznie nieistotnymi fragmentami białek, które ewoluowały w sposób neutralny i w związku z tym cała uwaga skupiała się na obszarach o wysokiej złożoności odpowiedzialnych za „właściwą funkcjonalności" białek. Takie podejście spowodowało, że bardzo często obszary LCR były maskowane przed właściwą analizą, aby swoją niską złożonością nie „myliły" używanych algorytmów. W rezultacie powszechnie dostępne narzędzia do grupowania

i wyszukiwania białek zostały zaprojektowane pod kątem HCR i w związku z tym nie są wiarygodne do analizy podobieństwa między LCR.

Biorąc pod uwagę błyskawiczny w ostatnich latach rozwój narzędzi bazujących na konieczności porównywania sekwencji w celu określania struktury białka a następnie przewidywania ich funkcji (np. AlphaFold, eggNOG, DeepFRI, ...) sprawa właściwego podejścia do LCR staje się kluczowa. W przedstawionej rozprawie Autor z jednej strony *Porównuje istniejące metody identyfikacji LCR i przedstawia metodę konsensusu, która wykorzystuje analizowane narzędzia* a z drugiej *sprawdza, czy istniejące metody analizy podobieństwa sekwencji białek mogą być wykorzystane do porównywania LCR i przedstawia nową metod ich analizy.* Tutaj chciałbym nadmienić, że drugi z wymienionych celów rozprawy, który jest jednocześnie głównym celem, został przedstawiony w zbyt zachowawczy sposób. Doktorant powinien mocniej zaakcentować unikalny oraz twórczy komponent swojej pracy.

W niniejszej rozprawie Autor przedstawia nowe metody identyfikacji, wizualizacji i porównywania LCR. Nowe metody identyfikacji LCR obejmują metodę konsensusu, która wykorzystuje relacje między innymi metodami, oraz GBSC (Graph Based on Sequence Clustering), która identyfikuje STR (Short Tandem Repeats). Do porównywania LCR udostępniono LCR-BLAST, który jest nową modyfikacją metody metody BLAST zoptymalizowaną pod LCR, a do grupowania nowo opracowaną metodę GBSC. Autor dokonał oceny zaproponowanych metod w porównaniu do innych wiodących rozwiązań na przykładowych zestawach danych i wykazał, że zaproponowane metody są lepiej dopasowane do analizy LCR.

W mojej ocenie na szczególną uwagę zasługuje zaproponowana metoda GBSC, która dopuszcza „wstawki" między powtórzeniami. Zaproponowana metoda grupuje podobne sekwencje według powtarzającego się wzoru, który znajduje podczas identyfikacji. Dzięki temu GBSC poprawnie analizuje również te LCR, które częściowo zawierają określony wzór, a częściowo zawierają losowe sekwencje. Jest to w kontraście do standardowych metod opartych na dopasowaniu, które identyfikują taki „niekompletny" LCR jako różny od „kompletnego".

Wyniki badań bezpośrednio i pośrednio związanych z tematem rozprawy zostały opublikowane w dwóch recenzowanych artykułach gdzie doktorant jest pierwszym autorem oraz w trzech innych artykułach jak również były prezentowane na konferencjach naukowych. Natomiast same badania były w części realizowane w ramach grantu OPUS i PRELUDIUM gdzie doktorant jest (współ-)kierownikiem.

Układ rozprawy jest typowy dla tego typu opracowań. W kolejnych rozdziałach:

- *Introduction* – Autor prowadzi czytelnika przez podstawowe informacje i definiuje cele pracy,

- *Description of sequence analysis methods* – Autor opisuje zastosowane w pracy metody do identyfikacja LCR i metody porównywania podobieństwa sekwencji,
- *Low complexity identification methods* – Autor porównuje wybrane metody identyfikacji LCR oraz demonstruje podejście do wizualnej analizy tych fragmentów,
- *Analysis of canonical methods for protein sequence comparison in the case of LCRs* – Autor porównuje trzy wiodące metody analizy podobieństwa sekwencji białek dla fragmentów o niskiej złożoności (BLAST, HHblits i CDHIT),
- *LCR-BLAST - searching for low complexity regions* – Autor przedstawia metodę LCR-BLAST, która jest nową modyfikacją BLAST do wyszukiwania podobnych LCR,
- *GBSC - clustering short tandem repeats* – Autor opisuje nową metodę metodę identyfikacji i grupowania STR-ów według podobnych powtarzających się wzorów. Przedstawione
- *Application* – Autor opisuje przydatność zaproponowanych metod na podstawie wykonanych i opublikowanych analiz LCR
- *Summary and conclusions* – Autor prezentuje podsumowanie pracy z uwypukleniem głównych osiągnięć

Autor bardzo dokładnie omawia obecny stan wiedzy co świadczy o dużym oczytaniu i dobrym przygotowaniu do podjęcia zagadnień poruszanych w dalszych częściach pracy. Czytelnik może mieć odczucie, że zwłaszcza wprowadzenie nie jest dobrze ustrukturyzowane, jednak w omawianym temacie są tak duże zależności, że musze przyznać, że doktorant bardzo dobrze poradził sobie z tą materią. Trochę zamieszania wprowadza też fakt, że metody takie bardziej ogólne są opisane w rozdziale 2 a te dedykowane do LCR w rozdziale 3. Można było pomyśleć o lepszych tytułach tych rozdziałów, które byłyby bardziej jednoznaczne. Należy jednak przyznać, że pomimo ogromu zaprezentowanego materiału czytelnik jest przeprowadzany przez kolejne rozdziały w bardzo umiejętny sposób. Cały układ pracy jest bardzo dobrze przemyślany przez co czytelnik w żadnym momencie nie traci z oczu celu pracy. Przedstawione wyniki na rzeczywistych danych jasno pokazują przewagę wprowadzanych nowych metod. Pomimo dużej szczegółowości czytelnik nie ma uczucia przytłoczenia czy zagubienia.

Tezy rozprawy są sformułowane jasno i przystępnie oraz są w pełni poparte danymi zawartymi w poszczególnych rozdziałach rozprawy. Choć tak jak nadmieniłem wyżej, doktorant mógł mocniej zaakcentować swój twórczy wkład. Podsumowanie rozprawy jest syntetycznym wykazaniem, że założone w pracy cele zostały osiągnięte. Jednocześnie jasno zostało wskazane, w których analizach poszczególne metody miały zastosowanie.

Jedynym zauważonym niedociągnięciem jest kwestia ryciny 3.2, a konkretnie szczątkowy opis jak ją interpretować. Opis rycin powinien w wyczerpujący sposób objaśniać co i dlaczego na niej jest zaprezentowane.

Rozprawa zawiera 47 szczegółowych rycin oraz 11 tabel, które co do zasady są jasne i czytelne. Piśmiennictwo obejmuje 137 dobrze dobranych i aktualnych pozycji.

Podsumowując, przedstawiona do oceny praca doktorska stanowi bardzo wartościowe uzupełnienie obecnego stanu wiedzy odnośnie metod identyfikacji i analizy LCR poprzez kompleksowe porównanie dostępnych rozwiązań oraz zaproponowanie trzech nowych rozwiązań. W związku z tym praca ta jednocześnie wyznacza kierunek dalszych badań nad znaczeniem LCR. Praca ta w pełni odpowiada warunkom stawianym rozprawom doktorskim oraz wypełnia istotną lukę w obecnym stanie wiedzy. Należy też podkreślić ogrom pracy wykonanej przez Doktoranta jak również fakt, że zaproponowane metody z sukcesem są wykorzystywane do analiz.

Na podstawie powyższej oceny stwierdzam, że wymieniona rozprawa doktorska w pełni odpowiada warunkom stawianym w ustawie Prawo o szkolnictwie wyższym i nauce / Dz. U. z 2022 r. poz. 574, w zakresie nadawania stopni naukowych i na tej podstawie wnoszę do Wysokiej Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej o dopuszczenie mgr inż. Patryka Jarnota do dalszych etapów przewodu doktorskiego. Równocześnie chciałbym zgłosić pod dyskusję przyznanie wyróżnienia. Motywacją jest zarówno ogrom jak i wysoka jakość wykonanej pracy ale także bardzo wysoka jakość przygotowanej rozprawy.

Nie mam wątpliwości, że doświadczenie zgromadzone przez Autora stawia cały zespół badawczy w doskonałej pozycji wśród międzynarodowych grup zajmujących się tą tematyką.

Dr hab. inż. Paweł Piotr Łabaj