Laboratory of Bioinformatics and Computational Genomics LB!GO
Faculty of Mathematics and Information Science, Warsaw University of Technology
ul. Koszykowa 75, 00-662 Warsaw, Poland

Laboratory of Functional and Structural Genomics LFSG
Centre of New Technologies, University of Warsaw
Banacha 2c Street, 02-097 Warsaw, Poland

mobile: +48504726203, e-mail: Dariusz.Plewczynski@pw.edu.pl, www: https://plewczynski-lab.org

Warszawa,
17.07.2023

Warszawa, 17/07/2023

Prof. dr hab. Dariusz Plewczyński
Laboratorium Bioinformatyki i Genomiki Obliczeniowej,
Wydział Matematyki i Nauk Informacyjnych,
Politechnika Warszawska
Laboratorium Genomiki Funkcjonalnej i Strukturalnej
Centrum Nowych Technologii
Uniwersytet Warszawski

RECENZJA rozprawy doktorskiej mgra Patryka Jarnota

Methods for similarity analysis of low complexity regions in protein sequences

Ukończonej na Wydziale Automatyki, Elektroniki i Informatyki
Politechniki Śląskiej

pod opieką
promotor dr hab. inż. Aleksandry Grucy, prof. Politechniki Śląskiej
oraz
ko-promotora dr hab. Marcina Grynberga

zgłoszonej do
Rady Dyscypliny Naukowej Informatyki Technicznej i Telekomunikacji
w dziedzinie Nauki Techniczne
Politechniki Śląskiej

Sekwencje białkowe o niskiej złożoności (ang. low complexity regions, LCRs) to często spotykane fragmenty sekwencji białek o niewielkim zróżnicowaniu aminokwasowym. Definicja złożoności jest mocno nieprecyzyjna, ale naukowcy zgadzają się, że są to sekwencje proste składające się z homopolimerów, krótkich powtórzeń tandemowych (STR) i nieregularnych fragmentów charakteryzujących się niskim zróżnicowaniem aminokwasowym. Mimo że wyglądają jak proste sekwencje zawierające bardzo mało informacji, są one często spotykane w białkach i dlatego powinniśmy traktować je jako istotne funkcjonalnie fragmenty białek. Liczne badania już wykazały ich unikalne właściwości strukturalne (np. wykazują właściwości wiążące i występują w domenach, które tworzą agregaty amyloidowe). Mimo to przez długi czas naukowcy uważali, że LCR są biologicznie nieistotnymi fragmentami sekwencji białek, które ewoluowały w sposób neutralny bez presji ewolucyjnej. Badacze skupiali się na badaniu regionów o wysokiej złożoności (ang. high complexity regions, HCRs), typowo maskowano regiony LCR, aby poprawić wyszukiwanie homologii między białkami.

Dostępne narzędzia do grupowania i wyszukiwania białek zostały zaprojektowane pod kątem HCR, dlatego nie uniemożliwiają przeprowadzenia wiarygodnej analizy podobieństwa między sekwencjami LCR. Regiony LCR są przyczyną błędnych trafień sekwencji białkowych w wyszukiwarkach, co zmotywowało naukowców do prac nad modelami statystycznymi do analizy podobieństwa tych regionów. Pierwsza metoda SEG zaproponowana przez Woottona i Federhena została dodana do narzędzia BLAST umożliwiając zmniejszenie fałszywych trafień spowodowanych obecnością LCRs. Z kolei Promponas i współautorzy opracowali metodę CAST, która wykrywa składowe w regionach LCR i maskuje je w celu poprawy wyszukiwania podobieństwa między białkami. Alba i współautorzy zastosowali metodę SIMPLE, początkowo stworzoną dla sekwencji DNA, do białek, argumentując, że może być ona używana do badania ewolucji LCRs i ich funkcji biologicznych. Li i Kahveci zaproponowali inne podejście do LCR, które rozpoznaje powtórzenia w regionach LCR i traktuje je jako ważny czynnik dla lepszego zrozumienia funkcji biologicznej białek. Powyższe metody jak i inne dostępne narzędzia bioinformatyczne ulepszyły identyfikacje LCR oraz umożliwiają badanie ich roli biologicznej w szerszej skali, korzystając z ugruntowanych modeli statystycznych lub heurystyk bazujących na optymalizacji wielu parametrów.

Przedstawiona mi do recenzji praca jest wynikiem udanej analizy teoretycznej w paradygmacie bioinformatyki. Autorowi udało się twórczo połączyć innowacyjną metodologię teoretyczną z biologicznie istotnym problemem badawczym skupiającym się na głębszym zrozumieniu roli regionów o niskiej złożoności oraz rozwijaniu metod ich porównywania. Budowa modeli obliczeniowych w celu lepszego porównywania sekwencji białek, co umożliwia zrozumienie procesów ewolucji organizmów żywych oraz lepiej identyfikuje funkcje biologiczne tych regionów, stanowi przykład wysoko interdyscyplinarnej dziedziny, jaką jest bioinformatyka. Podejście bazujące na metodach obliczeniowych i starannie przygotowanych danych doświadczalnych stanowi fundament nowoczesnej biologii molekularnej, a także wspomaga rozwój nowych algorytmów informatyki technicznej w odpowiedzi na podstawowe problemy i wyzwania genomiki.

W myśl wymagań Ustawy o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki z dnia 14 marca 2003 r. (Dz.U. 2017 poz. 1789, z późn. zm.) oraz Rozporządzenia Ministra Nauki i Szkolnictwa Wyższego z dnia 19 stycznia 2018 r. w sprawie szczegółowego trybu i warunków przeprowadzania czynności w przewodzie doktorskim, w postępowaniu habilitacyjnym oraz w postępowaniu o nadanie tytułu profesora (Dz.U. 2018 poz. 261), jak również art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478, 619, 1630), przedmiotem mojej oceny jest oryginalność rozwiązanego problemu naukowego, ogólna wiedza teoretyczna Kandydata w dziedzinie informatyki technicznej i bioinformatyki, a także umiejętność samodzielnego prowadzenia pracy naukowej.

Przedmiot mojej oceny, czyli rozprawa doktorska, jest w pełni zgodna z warunkami określonymi w powyższych Ustawach, a nawet moim zdaniem przekracza zwyczajowe wymagania. Autor prezentuje wyjątkową oryginalność sformułowania unikalnego problemu naukowego oraz jego rozwiązania przy użyciu zaawansowanych metod informatycznych, szeroką wiedzę teoretyczną z zakresu biologii molekularnej i informatyki technicznej, a także umiejętność prowadzenia wytrwałej i twórczej pracy naukowej. Praca Doktoranta wykazuje również dużą samodzielność badawczą, mimo istotnego wsparcia ze strony obu współ-promotorów. Nie mam żadnych poważnych uwag do rozprawy doktorskiej, jakość przedstawienia metodologii i wyników badań, wyjątkowego dorobku publikacyjnego Kandydata. Całość rozprawy w pełni zasługuje na zaakceptowanie przez Radę Dyscypliny, a nawet jestem przekonany na wyróżnienie w stosowny sposób.

Rozprawa doktorska Pana mgr Patryka Jarnota została przygotowana w Katedrze Sieci i Systemów Komputerowych na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej w Gliwicach pod kierownictwem dra hab. Aleksandry Grucy, profesor uczelni, oraz dr hab. Marcina Grynberga z Instytutu Biochemii i Biofizyki Polskiej Akademii Nauk.

Praca liczy sobie 137 stron wraz z odniesieniami do artykułów istotnych dla zrozumienia dziedziny badawczej, jak również opublikowanych przez Kandydata ze współautorami w punktowanych czasopismach naukowych.

W niniejszej pracy Autor postawił sobie następujące cele: (i) analiza istniejących metod identyfikacji LCR i wizualizacja ich wyników; (ii) analiza istniejących metod analizy podobieństwa LCR; (iii) ulepszenie wyszukiwania podobnych LCR; (iv) ulepszenie grupowania podobnych LCR. Skupia się na analizie regionów o niskiej złożoności w sekwencjach białkowych, przy wykorzystaniu różnych metod i narzędzi oraz opracowaniu nowych algorytmów z dziedziny klasycznej bioinformatyki, ciekawie rozszerzonej o nowe wyzwania związane z analizą powtarzalności aminokwasów oraz określaniu ich roli w strukturze białek i ich kompleksów.

Motywacją pracy jest brak odpowiedniego instrumentarium do analizy bardzo podobnych sekwencji białek. Zdaniem recenzenta jest to ważny i często niedoceniany problem badawczy związany z mechanizmami ewolucyjnymi i charakterem mutacji. Od niedawna co prawda pojawiło się kilka metod identyfikacji LCR, metody te mogą również maskować LCR w celu poprawy wyszukiwania białek homologicznych oraz określenia ich funkcji. Jednakże, jak dotąd brakuje w literaturze naukowej szczegółowego porównania tych metod, które mogłoby być również wykorzystane do połączenia ich w celu uzyskania nowych wyników. W ramach swojej pracy doktorskiej Kandydat porównuje istniejące metody identyfikacji LCR i proponuje metodę konsensusową, która wykorzystuje aktualnie dostępne narzędzia, dodatkowo prezentując w sposób bardzo przejrzysty wyniki i wnioski ze swoich badań.

Funkcja biologiczna niektórych LCR jest już dobrze poznana, ale nadal brakuje dobrego opisu i bardziej sformalizowanych metod do ich wyszukiwania, analizy i wizualizacji nowych regionów LCR, których jest bardzo dużo w badach danych. Paradygmat który realizuje doktorant umożliwia rozszerzenie wyników szczegółowych otrzymanych dzięki czasochłonnym i kosztownym doświadczeniom biologii molekularnej na nowe przypadki za pomocą metod in silico. Umożliwia to dokonanie przewidywania właściwości fragmentu białka o

niskiej złożoności na podstawie obserwacji dotyczących innych podobnych fragmentów, których rola jest już znana na podstawie eksperymentów. Proponując metody porównywania regionów LCR umożliwiamy więc przewidywanie ich własności co znacząco obniża koszt związany z metodami doświadczalnymi. Eksperymenty często są nisko-przepustowe, daleko im do masywnego stosowania dla dużych zbiorów białek. Trudno nie zgodzić się z Doktorantem, że to właśnie metody in silico mogą istotnie wspomóc metody doświadczalne w analizie LCRs. Obecnie możemy używać istniejących metod identyfikacji LCR do pobierania LCR z publicznie dostępnych baz danych, brakuje nam jednak metod ich efektywnego porównywania i wizualizacji wyników. Dlatego głównym celem doktoratu jest sprawdzenie, czy istniejące metody analizy podobieństwa sekwencji białkowych mogą być używane do porównywania LCR oraz opracowanie nowych metod ich analizy i wizualizacji.

Lista prac naukowych opublikowanych przez Doktoranta w trakcie pracy nad doktoratem obejmuje pięc pozycji powiązanych z tematyką pracy doktorskiej. Poniżej lista publikacji z bazy pubmed opublikowanych przez doktoranta:

- [P1] Insights from analyses of low complexity regions with canonical methods for protein sequence comparison. **Jarnot P**, Ziemska-Legiecka J, Grynberg M, Gruca A. **Brief Bioinform.** 2022 Sep 20;23(5):bbac299. doi: 10.1093/bib/bbac299. PMID: 35914952; Impact Factor: 13.99, Punkty ministerialne: 140;
- [P2] Common low complexity regions for SARS-CoV-2 and human proteomes as potential multidirectional risk factor in vaccine development. Gruca A, Ziemska-Legiecka J, **Jarnot P**, Sarnowska E, Sarnowski TJ, Grynberg M. **BMC Bioinformatics.** 2021 Apr 8;22(1):182. doi: 10.1186/s12859-021-04017-7. PMID: 33832440, Impact Factor: 3.327, Punkty ministerialne: 100;
- [P3] PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. **Jarnot P**, Ziemska-Legiecka J, Dobson L, Merski M, Mier P, Andrade-Navarro MA, Hancock JM, Dosztányi Z, Paladin L, Necci M, Piovesan D, Tosatto SCE, Promponas VJ, Grynberg M, Gruca A. **Nucleic Acids Res.** 2020 Jul 2;48(W1):W77-W84. doi: 10.1093/nar/gkaa339. PMID: 32421769, Impact Factor: 19.16, Punkty ministerialne: 200;
- [P4] Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, **Jarnot P**, Gruca A, Grynberg M, Kajava AV, Promponas VJ, Anisimova M, Jakobsen KS, Linke D. **Nucleic Acids Res.** 2019 Dec 2;47(21):10994-11006. doi: 10.1093/nar/gkz841. PMID: 31584084, Impact Factor: 19.16, Punkty ministerialne: 200;

- [P5] Quantitative Conformational Analysis of Functionally Important Electrostatic Interactions in the Intrinsically Disordered Region of Delta Subunit of Bacterial RNA Polymerase. Kubáň V, Srb P, Štégnerová H, Padrta P, Zachrdla M, Jaseňáková Z, Šanderová H, Vítovská D, Krásný L, Koval' T, Dohnálek J, Ziemska-Legiecka J, Grynberg M, **Jarnot P**, Gruca A, Jensen MR, Blackledge M, Žídek L. **J Am Chem Soc.** 2019 Oct 23;141(42):16817-16828. doi: 10.1021/jacs.9b07837. Epub 2019 Oct 10. PMID: 31550880, Impact Factor: 16.383, Punkty ministerialne: 200;

Doktorant uzyskał samodzielnie grant badawczy pod tytułem „*Metody analizy podobieństwa regionów o niskiej złożoności*", który został przyznany w 2021 roku przez Narodowe Centrum Nauki w konkursie PRELUDIUM (nr 2021/41/N/ST6/01919). Metoda GBSC została z kolei zaprojektowana w ramach grantu OPUS pt. „LCRPlatform: nowe algorytmy i metody identyfikacji, kategoryzacji oraz adnotacji regionów białkowych o niskiej złożoności" o numerze 2020/39/B/ST6/03447, którego kierownikiem był jeden ze współpromotorów (dr hab. Marcin Grynberg).

W swojej pracy doktorskiej Patryk Jarnot przedstawia nowe metody identyfikacji, wizualizacji i porównywania obszarów o niskiej złożoności (LCR). Opracowane metody, takie jak metoda konsensusowa i GBSC, umożliwiają identyfikację LCR oraz analizę krótkich powtórzeń tandemowych (STR). Do porównywania i grupowania LCR można zastosować modyfikację metody BLAST o nazwie LCR-BLAST oraz nowo opracowane narzędzie GBSC. Metody te zostały ocenione poprzez porównanie z innymi rozwiązaniami oraz zastosowane w kilku przykładowych aplikacjach. Patryk Jarnot opublikował dwa artykuły jako pierwszy autor w renomowanych czasopismach naukowych, które prezentują wyniki jego badań (w tym w remonomowanym czasopiśmie Briefings in Bioinformatics). W ramach pracy doktorskiej przeprowadzono również badania we współpracy z innymi autorami, które zostały opublikowane w Nucleic Acid Research, BMC Bioinformatics i Journal of American Chemical Society. Metoda GBSC została opracowana w ramach prac finansowanych przez granty OPUS i PRELUDIUM, w których mgr. Patryk Jarnot brał udział jako współbadacz lub główny badacz.

Praca doktorska mgr. Patryka Jarnota składa się z ośmiu głównych części. "Wstęp" wprowadza czytelnika do tematyki badawczej i określa cele pracy. Rozdział "Metody analizy sekwencji" przedstawia metody identyfikacji LCR i porównywania podobieństwa sekwencji. Sekcja "Metody identyfikacji o niskiej złożoności" porównuje różne metody identyfikacji LCR i prezentuje autorskie podejście do wizualnej analizy tych fragmentów. Rozdział "Analiza kanonicznych metod porównywania sekwencji białek w przypadku LCR" porównuje trzy

nowoczesne i często używane metody analizy podobieństwa sekwencji białek: BLAST, HHblits i CD-HIT. Z kolei sekcja "LCR-BLAST - wyszukiwanie regionów o niskiej złożoności" przedstawia nowatorską modyfikację metody BLAST do wyszukiwania wysoko-podobnych LCR. Sekcja "GBSC - klasteryzacja krótkich powtórzeń tandemowych" opisuje autorską metodę identyfikacji i grupowania STR-ów (ang. short tandem repeats). W rozdziale "Zastosowania" opracowane metody są stosowane w różnych analizach LCR, prezentowane są przykłady i otrzymane wyniki. Praca kończy się rozdziałem "Podsumowanie i wnioski".

W rozdziale pierwszy „Introduction” Patryk Jarnot wprowadza w tematykę swojej pracy doktorskiej, przedstawia motywację i cele pracy, a następnie omawia strukturę pracy i najważniejsze wyniki.

W rozdziale drugim „Description of sequence analysis methods” Doktorant skupia się na opisie metod analizy sekwencji białkowych. Przedstawione są różne metody identyfikacji obszarów o niskiej złożoności (LCR), takie jak fLPS, CAST, SEG, LCD-Composer, SIMPLE, T-REKS, XSTREAM i GBA. Metody te służą do analizy składu, powtórzeń tandemowych i struktury LCR. Kolejnym tematem są metody porównywania podobieństwa sekwencji białek, takie jak BLAST, HHblits, CD-HIT, MMseqs2 i MCL. Opisuje się zastosowanie tych metod w przeszukiwaniu bazy danych białek, klasyfikacji rodzin białek i analizie funkcji białek.

W rozdziale trzecim „Low complexity identification methods” mgr Patryk Jarnot omawia metody identyfikacji obszarów o niskiej złożoności (LCR) w białkach. Przedstawione są różne metody takie jak SEG, CAST, SIMPLE, GBA, XSTREAM, T-REKS, fLPS, LCD-Composer i GBSC. Autor porównuje te metody pod kątem wyników, analizuje ich nakładanie się oraz wskazuje na unikalne cechy każdej z nich. Opisuje również zastosowanie tych metod w identyfikacji powtórzeń tandemowych. Autor przedstawia analizę długości i składu aminokwasowego LCR, wskazując na różnice wynikające z zastosowania różnych metod. Omawia również podobieństwa wyników i możliwość kombinowania metod w celu uzyskania większej precyzji. Wnioskiem z rozdziału jest potrzeba dalszych badań i analiz w celu lepszego zrozumienia i identyfikacji LCR w białkach.

W rozdziale czwartym „Analysis of canonical methods for protein sequence comparison in the case of LCRs” Doktorant analizuje istniejące metody porównywania sekwencji białkowych, szczególnie w kontekście obszarów o niskiej złożoności (LCR). Autor stwierdza, że metody takie jak BLAST, HHblits i CD-HIT, zaprojektowane pierwotnie dla obszarów o wysokiej złożoności (HCR),

wymagają ulepszeń w celu skutecznego porównywania LCR. Wskazuje na niedoszacowanie roli LCR przez BLAST i HHblits, problemy z generowaniem profilu HMM dla LCR oraz niestosowność metryki E-value dla porównywania LCR. Autor zaleca uwzględnienie innych miar, takich jak wynik i tożsamość, oraz globalne porównywanie sekwencji LCR. Również metoda CD-HIT, oparta na grupowaniu, może wymagać ulepszeń. Wnioskiem jest potrzeba opracowania nowych metod lub ulepszenia istniejących w celu skutecznego porównywania LCR.

W rozdziale piątym "LCR-BLAST - searching for low complexity regions" Kandydat opisuje opracowanie narzędzia o nazwie LCR-BLAST, które umożliwia bardziej efektywne wyszukiwanie podobnych regionów o niskiej złożoności (LCR) w sekwencjach białkowych. Autor proponuje ulepszenia dla popularnego narzędzia BLAST, które tradycyjnie skupiało się na regionach o wysokiej złożoności (HCR).

Przeprowadzona analiza obejmuje kilka modyfikacji, w tym wyłączenie obliczania macierzy skorelowania opartej na kompozycji, a także wprowadzenie nowych metryk, takich jak średni wynik i macierz skorelowania identyczności. LCR-BLAST został zoptymalizowany dla krótkich sekwencji poprzez dostosowanie różnych parametrów, takich jak kary za otwarcie i rozszerzenie przerw, próg wartości E, rozmiar słowa i opcje filtrowania.

Autor podkreśla potrzebę opracowania dedykowanej macierzy skorelowania dla LCR, ze względu na ich odmienne preferencje w zakresie aminokwasów. LCR-BLAST został opracowany w oparciu o UniProtKB/Swiss-Prot, a wyniki porównań zostały przedstawione i ocenione.

Wnioskiem jest, że LCR-BLAST stanowi istotną poprawę w porównaniu do istniejących wariantów BLAST dla analizy LCR. Narzędzie to umożliwia bardziej precyzyjne i skuteczne porównywanie sekwencji LCR, wykorzystując zoptymalizowane parametry i nowe miary oceny.

W rozdziale szóstym „GBSC - clustering short tandem repeats" Doktorant opisuje nową metodę o nazwie GBSC, służącą do identyfikacji i grupowania krótkich powtórzeń tandemowych (STR) w sekwencjach białkowych. Metoda ta ma na celu przezwyciężenie ograniczeń istniejących metod identyfikacji regionów o niskiej złożoności (LCR) oraz skupia się na szczególnych wyzwaniach związanych z STR.

GBSC wykorzystuje model grafowy do identyfikacji i grupowania STR w sekwencjach białkowych. W trakcie procesu identyfikacji wykorzystuje się model grafowy typu De Bruijn, który generuje węzły na podstawie k-merów w sekwencji i tworzy krawędzie między kolejnymi k-merami. Krawędzie te mają określony czas życia i wagę, które są manipulowane w trakcie procesu. Jeśli czas życia krawędzi przekracza ustalony próg, krawędź znika. Jeśli waga krawędzi przekracza próg, tworzy się graf reprezentujący STR.

W procesie grupowania wykorzystuje się modele grafowe STR i przypisuje sekwencje do klastrów na podstawie podobieństwa wzorca. GBSC jest w stanie obsłużyć sąsiadujące STR, grupując je jako osobne lub połączone typy.

Wprowadzono również opcję redukcji alfabetu, która polega na łączeniu podobnych aminokwasów. Porównano GBSC z innymi metodami grupowania, takimi jak MMseqs i CD-HIT, używając STR zidentyfikowanych przez GBSC jako danych wejściowych.

Metoda GBSC ma wiele zalet, takich jak obsługa degeneracyjnych powtórzeń, grupowanie sąsiadujących STR i wykorzystanie grafowych modeli do oznaczania zidentyfikowanych i pogrupowanych sekwencji. Jest to znaczące ulepszenie w porównaniu do istniejących metod identyfikacji regionów o niskiej złożoności, które nie dostarczają tak szczegółowych informacji na temat zidentyfikowanych fragmentów.

Wnioskiem jest to, że GBSC oferuje wiele zalet w porównaniu do istniejących metod, zwłaszcza pod względem grupowania powtórzeń degeneracyjnych, obsługi sąsiadujących powtórzeń i szczegółowości dostarczanych informacji dzięki wykorzystaniu modeli grafowych. Metoda ta może być szczególnie pomocna, gdy interpretowalność powtórzeń ma kluczowe znaczenie.

W rozdziale siódmym „Applications” pracy doktorskiej Kandydat przedstawia praktyczne zastosowania dwóch metod porównywania sekwencji białek: LCR-BLAST do wyszukiwania podobnych regionów o niskiej złożoności (LCR) i GBSC do grupowania krótkich powtórzeń tandemowych (STR) na podstawie wzorców powtarzających się. Autor wykorzystuje te metody do analizy błędów rekonstrukcji sekwencji STR, identyfikacji podobnych segmentów do domeny K-D/E z podjednostki delta polimerazy RNA bakteryjnej oraz znajdowania wspólnych epitopów dla SARS-CoV-2 i genomu ludzkiego. Badania obejmują również analizę STR w różnych podziałach taksonomicznych, analizę LCR z

polimerazy RNA bakteryjnej oraz analizę proteomu SARS-CoV-2 w kontekście potencjalnych błędów składania i projektowania szczepionek.

W ostatnim, ósmym rozdziale „Summary and conclusions" autor podsumowuje swoje badania streszczając najważniejsze osiągnięcia.

Wnioski końcowe pracy doktorskiej obejmują:

- Autor przedstawił trzy nowatorskie metody analizy regionów o niskiej złożoności (LCR) w białkach oraz porównał je z istniejącymi rozwiązaniami. Rozpoczął od oceny istniejących metod identyfikacji LCR i pokazał, jak je zintegrować, aby uzyskać nowe wyniki przy użyciu podejścia konsensualnego. Opracował również podejście do wizualizacji, które ułatwia eksplorację LCR i pomaga naukowcom formułować hipotezy dotyczące właściwości biologicznych białek.
- Autor postawił hipotezę, że obecne metody analizy podobieństwa sekwencji białek są głównie projektowane do analizy regionów o dużej złożoności (HCR) sekwencji białkowych, często pomijając LCR. Ich analiza potwierdziła to, pokazując, że istniejące metody nie są tak efektywne w porównywaniu LCR, nawet po optymalizacji. W rezultacie autor dostosował parametry BLAST do lepszego porównywania LCR, co doprowadziło do stworzenia LCR-BLAST, zmodyfikowanej wersji BLAST dla tych fragmentów białkowych.
- Następnie autor przedstawił GBSC, nową metodę identyfikacji i grupowania krótkich powtórzeń tandemowych (STR). GBSC może analizować szeroki zakres STR, od jasnych do silnie zdegenerowanych, obejmujących różne typy LCR.
- Na koniec autor zademonstrował praktyczne zastosowanie LCR-BLAST i GBSC w trzech badaniach: badaniu błędów składania sekwencji po powtórzeniach tandemowych, analizie interakcji w polimerazie RNA oraz identyfikowaniu wspólnych LCR w proteomach ludzkim i SARS-CoV-2. Autor pokazał zdolności GBSC do wykrywania nieregularnych LCR i zwrócenia uwagi na możliwe ryzyko związane z wyborem określonych epitopów dla leków i szczepionek związanych z SARS-CoV-2.
- Wszystkie te algorytmy i metody opracowane w rozprawie doktorskiej zostały opublikowane w pięciu manuskryptach naukowych o wysokim impact factor i punktach ministerialnych, podkreślając potrzebę dalszego rozwoju metod uwzględniających sekwencje LCR w białkach i praktycznego ich wykorzystania w badaniach naukowych.

Pytania do doktoranta:

- Jak można dalej wykorzystać uzyskane zbiory sekwencji białek zawierające segmenty LCR? W jakim zakresie są one pełne i odzwierciedlają wszystkie podtypy LCR? Czy przeprowadzono analizę skupień wszystkich zebranych LCR i wyróżniono klasy funkcjonalne i ich unikalne cechy odróżniające je od siebie?
- Dlaczego w większości przypadków w analizie LCRs używane są tradycyjne macierze podobieństwa aminokwasów (BLOSUM, PAM)? Autor proponuje używanie macierzy identyczności – czy można użyć innych alternatywnych macierzy, takich jak AAindex które reprezentujące podobieństwa fizykochemiczne między sekwencjami białek?
- Czy można wytrenować metody uczenia maszynowego do zadania porównawczego i przyspieszenia analizy białek zawierających LCR? Jakie cechy powinny być użyte w treningu? Czy wystarczy samo podobieństwo ewolucyjne czy też użycie cech fizyko-chemicznych lub strukturalnych miałoby dodatkowe zalety?
- Czy obserwowane są wyróżniające się deskryptory fizyko-chemiczne z bazy AAindex w kontekście ich wzbogacenia w sekwencjach LCR?
- A jeśli tak – to czy do rankingu cech można byłoby użyć metody selekcji cech, jak należy wtedy postawić zadanie klasyfikacji i treningu uczenia maszynowego?
- Które deskryptory z AAindex są wzbogacone w zbiorach aminokwasów z segmentów LCR, czy zadziałałyby one najlepiej w tak postawionej klasyfikacji LCR?

W ramach uwag do pracy doktorskiej:

- Należy ujednolicić wygląd wszystkich wykresów i tabel.
- Należy poprawić drobne błędy interpunkcyjne.
- Warto zamieścić elementy związane z życiorysem badawczym doktoranta – listę publikacji, listę prezentacji ustnych oraz posterów na konferencjach krajowych i międzynarodowych itp.

Wnioski końcowe

W podsumowaniu oceny rozprawy doktorskiej pana mgr Patryka Jarnota pt. "*Methods for similarity analysis of low complexity regions in protein sequences*", pragnę wyrazić moje wysokie uznanie dla przedstawionej pracy. Suma wskaźników impact factor wszystkich publikacji wynosi **72.02**, a suma punktów ministerialnych w Polsce wynosi **840**. Jest to wynik moim zdaniem wyjątkowy, nawet jeśli uwzględnimy że są to publikacje współautorskie.

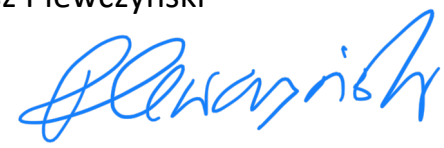
Biorąc pod uwagę czytelność i wartość naukową rozprawy doktorskiej, udane połączenie starannie opisanych narzędzi obliczeniowych, a także biologicznie istotnych eksperymentów i fundamentalnych pytań badawczych związanych z regionami niskiej złożoności w sekwencjach białkowych, uważam rozprawę doktorską pana mgr Patryka Jarnota za znaczący wkład w dziedzinę bioinformatyki, oraz istotny wykład w Informatykę Techniczną i Telekomunikację.

Rozprawa doktorska spełnia warunki określone w Art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478, 619, 1630). Ponadto uważam, że rozprawa ta przewyższa wszystkie powszechne i ustawowe wymagania stawiane rozprawom doktorskim, stanowi oryginalne rozwiązanie problemu naukowego, wykazuje ogólną wiedzę informatyczną kandydata w zakresie LCR/HCR i technik bioinformatycznych, oraz demonstruje zdolność do samodzielnego prowadzenia pracy naukowej.

W związku z powyższym, mam przyjemność przedłożyć Radzie Dyscypliny Naukowej Informatyki Technicznej i Telekomunikacji Politechniki Śląskiej moją recenzję dotyczącą dopuszczenia pana mgr Patryka Jarnota do dalszych etapów przewodu doktorskiego.

Ponadto, zważywszy na wysoki poziom merytoryczny i obszerność rozprawy, jej staranne przygotowanie oraz klarowny sposób prezentacji tematyki badawczej, metodyki i wyników, **wnoszę o wyróżnienie rozprawy stosowną nagrodą**.

Prof. Dariusz Plewczyński



Dariusz Plewczynski, PhD, Professor of Exact and Natural Sciences; Principal Investigator
Phone: +48 22 554 36 54 or +48 22 234 7219
e-mail: d.plewczynski@cent.uw.edu.pl or Dariusz.Plewczynski@pw.edu.pl www: https://plewczynski-lab.org

Laboratory of Functional and Structural Genomics LFSG
Centre of New Technologies, University of Warsaw; Banacha 2c Street, 02-097 Warsaw, Poland
Laboratory of Bioinformatics and Computational Genomics LB!GO
Faculty of Mathematics and Information Science, Warsaw University of Technology; ul. Koszykowa 75, 00-662 Warsaw, Poland