Silesian University of Technology
Faculty of Automatic Control, Electronics
and Computer Science

PhD Thesis

Methods for similarity analysis of low complexity regions in protein sequences

Patryk Jarnot

Supervisor: dr hab. inż. Aleksandra Gruca, prof. PŚ.
Co-supervisor: dr hab. Marcin Grynberg

Gliwice 2023

Streszczenie szczegółowe

Regiony o niskiej złożoności (ang. Low Complexity Regions - LCRs) to często spotykane fragmenty białek o niewielkiej różnorodności aminokwasów. Definicja ta jest nieprecyzyjna, natomiast naukowcy są zgodni co do tego, że są to proste sekwencje złożone z homopolimerów, krótkich powtórzeń tandemowych (ang. Short Tandem Repeats - STRs) i nieregularnych fragmentów charakteryzujących się niskim zróżnicowaniem aminokwasów [25, 27, 5]. Ich dokładna liczba w dużej mierze zależy od zastosowanej metody identyfikacji. Przykłady różnych typów LCR-ów pokazano na rysunku 0.1. Choć wyglądają jak proste sekwencje, które są nośnikiem bardzo małej ilości informacji to te fragmenty występują bardzo często w białkach i dlatego powinniśmy rozważyć ich ważność. Liczne badania naukowe wykazały już ich ważne właściwości funkcjonalne i strukturalne. Na przykład wykazują właściwości wiążące i istnieją w domenach, które są odpowiedzialne za układanie struktur amyloidowych [33, 21, 14]. Przez długi czas naukowcy uważali jednak, że LCR-y to biologicznie nieistotne fragmenty białek, które wyewoluowały w sposób neutralny [11]. Badali głównie regiony o wysokiej złożoności (ang. High Complexity Regions - HCRs), a nawet maskowali LCR-y, aby usprawnić wyszukiwanie sekwencji homologicznych [29]. Prowadzi to do hipotezy, że powszechnie dostępne narzędzia do grupowania i wyszukiwania białek zostały zaprojektowane z myślą o HCR-ach i nie są one dostosowane do analizy podobieństw między LCR-ami.

Naukowcy rozpoczęli badanie LCR-ów, gdy odkryli, że regiony te są główną przyczyną ewolucyjnie nieistotnych dopasowań sekwencji białek w metodach wyszukujących podobieństwa w sekwencjach białkowych. Dlatego Wootton i Federhen zaprojektowali metodę SEG w swoim artykule na temat statystyk regionów o niskiej złożoności [35]. Metoda ta została dodana do BLAST-a jako rozwiązanie częstych fałszywych trafień [4]. Promponas et al. założyli, że tylko najczęściej występujące reszty w LCR-ach prowadzą do niewłaściwego dopasowania sekwencji i opracowali metodę CAST. Metoda wykrywa LCR-y, ale maskuje tylko nadreprezentatywne aminokwasy w celu

>sp|P14922|CYC8_YEAST General transcriptional corepressor CYC8
MNPGGEQTIM EQPAQQQQQQ QQQQQQQQQQ AAVPQQPLDP LTQSTAETWL SIASLAETLG
DGDRAAMAYD ATLQFNPSSA KALTSLAHLY RSRDMFQRAA ELYERALLVN PELSDVWATL
GHCYLMLDDL QRAYNAYQQA LYHLSNPNVP KLWHGIGILY DRYGSLDYAE EAFAKVLELD
PHFEKANEIY FRLGIIYKHQ GKWSQALECF RYILPQPPAP LQEWDIWFQL GSVLESMGEW
QGAKEAYEHV LAQNQHHAKV LQQLGCLYGM SNVQFYDPQK ALDYLLKSLE ADPSDATTWY
HLGRVHMIRT DYTAAYDAFQ QAVNRDSRNP IFWCSIGVLY YQISQYRDAL DAYTRAIRLN
PYISEVWYDL GTLYETCNNQ LSDALDAYKQ AARLDVNNVH IRERLEALTK QLENPGNINK
SNGAPTNASP APPPVILQPT LQPNDQGNPL NTRISAQSAN ATASMVQQQH PAQQTPINSS
ATMYSNGASP QLQAQAQAQA QAQAQAQAQA QAQAQAQAQA QAQAQAQAQA QAQAQAHAQA
QAQAQAQAQA QAQAQAQQQQ QQQQQQQQQQ QQQQQQQQQQ QQQQQQQLQP LPRQQLQQKG
VSVQMLNPQQ GQPYITQPTV IQAHQLQPFS TQAMEHPQSS QLPPQQQQLQ SVQHPQQLQG
QPQAQAPQPL IQHNVEQNVL PQKRYMEGAI HTLVDAAVSS STHTENNTKS PRQPTHAIPT
QAPATGITNA EPQVKKQKLN SPNSNINKLV NTATSIEENA KSEVSNQSPA VVESNTNNTS
QEEKPVKANS IPSVIGAQEP PQEASPAEEA TKAASVSPST KPLNTEPESS SVQPTVSSES
STTKANDQST AETIELSTAT VPAEASPVED EVRQHSKEEN GTTEASAPST EEAEPAASRD
AEKQQDETAA TTITVIKPTL ETMETVKEEA KMREEEQTSQ EKSPQENTLP RENVVRQVEE
DENYDD

Rysunek 0.1: Sekwencja białka z fragmentami o niskiej złożoności zaznaczonymi na czerwono. Zawiera homopolimeryczne powtórzenie glutaminy na początku sekwencji. W środku znajduje się krótkie powtórzenie tandemowe glutaminy i alaniny, po którym następuje powtórzenie homopolimeryczne glutaminy. Na końcu sekwencji znajdują się nieregularne LCR-y. LCR-y zostały zidentyfikowane metodą SEG z domyślnymi parametrami.

wyszukiwania podobieństwa między białkami [26]. Alba et al. zaadaptowali metodę SIMPLE, stworzoną początkowo dla sekwencji DNA, do białek i argumentowali, że metoda ta może być wykorzystana do badania ewolucji i funkcji LCR-ów [1]. Li i Kahveci opublikowali podejście do LCR-ów, w którym uznają powtórzenia w tych regionach za ważny czynnik dla lepszego zrozumienia ich biologicznych ról [22]. Harrison napisał w 2017 roku artykuł o metodzie fLPS, który jest metodą identyfikacji nadreprezentacyjnych aminokwasów w LCR-ach [12]. Metoda ta została stworzona w celu ułatwienia analizy funkcjonalnej w zidentyfikowanych fragmentach białek. Niedawno Cascarina et al. opublikowali metodę LCD-Composer, która wykrywa LCR-y z regularnie rozmieszczonymi wspólnymi resztami aminokwasowymi [6]. Równolegle z postępami badań nad LCR-ami naukowcy zaczęli analizować powtórzenia tandemowe w białkach. Homopolimery i STR-y, które mogą być wyraźne lub zdegradowane przez mutacje, są podzbiorami LCR-ów. Metody zdolne do wykrywania tych fragmentów, to XSTREAM i T-REKS [24, 18]. Patrząc na ewolucję metod identyfikacji LCR-ów, jasne jest, że początkowo miały one na celu ich maskowanie, a od niedawna odkrycie ich biologicznego znaczenia.

Najpopularniejszą metodą wyszukiwania podobnych sekwencji białek jest BLAST [4]. Wykorzystuje on algorytm Smitha-Watermana do dopasowania sekwencji oraz heurytykę, która wykorzystuje części HSP (ang. High Scoring Parts) sekwencji w celu przyspieszenia obliczeń [3, 30]. W przeszłości metoda SEG była używana do maskowania LCR-ów w celu poprawy wyszukiwania homologii [35]. Metoda ta została następnie zastąpiona przez korekcję macierzy punktacji, która zmniejsza znaczenie często występujących aminokwasów [36]. Główną przyczyną nadreprezentatywnych typów aminokwasów w sekwencjach białkowych są LCR-y, w związku z tym punkty w macierzy tych reszt aminokwasów są obniżane. W międzyczasie BLAST zaczął również używać macierzy punktacji z rodzin BLOSUM i PAM, aby usprawnić wyszukiwanie odległych homologii [13]. Następnie opracowano nowe metody, aby znaleźć bardziej odległe zależności ewolucyjne między białkami. Metody te szukają podobnych białek, które są używane do tworzenia dopasowania wielu sekwencji (MSA) i wykorzystują je jako zamiennik macierzy punktacji. Ta grupa metod obejmuje takie metody jak PSI-BLAST, HHblits i HMMER [4, 28, 8]. One również maskują LCR-y, zmniejszając ich wynik obliczany w trakcie dopasowania sekwencji. Na przykład w HHblits domyślny wynik zwiększa znaczenie rzadko występujących reszt aminokwasowych, ponieważ trudniej jest je przypadkowo wprowadzić np. przez mutacje z innego aminokwasu i dlatego uważa się je za ważne w ewolucji białek [31].

Wymienione metody i ich udoskonalenia wspierają hipotezę o kierunku rozwoju i specjalizacji tych metod do wykrywania podobieństw między HCR-ami ignorując lub wręcz wykluczając LCR-y z analiz podobieństwa.

1 Motywacja i cele

W tej pracy postawiłem następujące cele:

- Przeanalizowanie istniejących metod identyfikacji LCR-ów i wizualizacja wyników.
- Analiza dostępnych metod porównania sekwencji do LCR-ów.
- Usprawnienie wyszukiwania podobnych LCR-ów.
- Usprawnienie grupowania LCR-ów pod względem podobieństwa.

Aktualnie istnieje kilka metod identyfikacji LCR-ów. Wiemy, że te metody mogą maskować LCR-y w celu lepszego wyszukiwania homologii i mogą odkrywać różne istotne biologicznie fragmenty białek. Niektóre z nich potrafią wykrywać domeny podobne do prionów bogate w metioninę [2]. Inne mogą zostać wykorzystane do analizy polarnych LCR-ów [23]. Jednakże, według mojej najlepszej wiedzy, w literaturze naukowej brakuje ich szczegółowego porównania, które można by również wykorzystać do połączenia ich w celu uzyskania nowych wyników. W ramach tej pracy porównuję istniejące metody identyfikacji LCR-ów i proponuję nową metodę konsensusową, która wykorzystuje analizowane narzędzia.

Biologiczne role niektórych LCR-ów zostały już odkryte, ale nie są tak dobrze opisane jak HCR-y. Naukowcy mogą odkrywać ich właściwości funkcjonalne i strukturalne za pomocą czasochłonnych i kosztownych metod eksperymentalnych w laboratorium. Możemy natomiast przyspieszyć ten proces, dokonując wstępnych założeń dotyczących właściwości fragmentu białka na podstawie wnioskowania z innych podobnych fragmentów, których funkcje biologiczne są już znane. Jeśli znajdziemy podobne LCR-y o znanej funkcji, możemy zmniejszyć liczbę scenariuszy badawczych i obniżyć koszt metod eksperymentalnych. Dlatego metody komputerowe mają duży potencjał do współpracy z metodami eksperymentalnymi. Obecnie możemy wykorzystać metody identyfikacji LCR-ów do pobierania ich z baz danych, ale brakuje nam metod do ich efektywnego porównywania. W związku z powyższym, głównym

celem tej rozprawy jest sprawdzenie, czy istniejące metody analizy podobieństwa sekwencji białek mogą być wykorzystane do porównania LCR-ów oraz opracowanie nowych metod ich analizy.

2 Najważniejsze wyniki

W niniejszej pracy przedstawiłem nowe metody identyfikacji, wizualizacji i porównania LCR-ów. Metody identyfikacji LCR-ów obejmują metodę konsensusu, która wykorzystuje relacje między innymi metodami, oraz GBSC, która identyfikuje STR-y. Do porównania LCR-ów naukowcy mogą użyć LCR-BLAST, która jest nową modyfikacją BLAST-a, a do grupowania mogą użyć nowo opracowanej metody, którą jest GBSC. Metody te zostały porównane z innymi rozwiązaniami i wykorzystane w analizach biologicznych w celu pokazania ich przydatności.

W efekcie opublikowałem jako pierwszy autor dwa artykuły w czasopismach naukowych. Pierwszy artykuł zatytułowany „PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins" ukazał się w czasopiśmie Nucleic Acid Research. Praca ta łączy wybrane metody identyfikacji LCR-ów w metodę konsensusową i wizualizuje wyniki odnalezionych fragmentów [15]. Drugi artykuł zatytułowany „Insights from analyses of low complexity regions with canonical methods for protein sequence comparison" został opublikowany w Briefings in Bioinformatics [17]. W tym artykule dostosowano parametry HHblits i CD-HIT w celu polepszenia analizy LCR-ów. Ostrzega on również społeczność naukową, że nawet jeśli używamy zoptymalizowanych zestawów parametrów, to nadal musimy być świadomi błędnych dopasowań sekwencji białek spowodowanych podstawowymi założeniami projektowymi w tych metodach do lepszego dopasowania HCR-ów. Opublikowałem również artykuł konferencyjny na International Conference on Man-Machine Interactions (ICMMI 2019) jako pierwszy autor. Artykuł zatytułowany „LCR-BLAST—a new modification of BLAST to search for similar low complexity regions in protein sequences" wprowadza nową modyfikację metody BLAST dla LCR-ów [16]. Proponuję tam optymalny zestaw parametrów do wyszukiwania tych fragmentów oraz dodatkową metrykę jako alternatywę dla E-wartości, która sortuje wyniki niezależnie od długości znalezionych podobieństw. Jestem także współautorem trzech artykułów, w których wykorzystano przedstawione w tej pracy metody analizy podobieństwa sekwencji białkowych. Wspomniane artykuły to:

„Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases" autorstwa Tørresena et al., „Common low complexity regions for SARS-CoV-2 and human proteomes as a potential multidirectional risk factor in vaccine development" autorstwa Gruca et al. oraz „Quantitative conformational analysis of functionally important electrostatic interactions in the intrinsically disordered region of the delta subunit of bacterial RNA polymerase" autorstwa Kubáň et al. Artykuły te zostały opublikowane odpowiednio w Nucleic Acid Research, BMC Bioinformatics i Journal of American Chemical Society [34, 10, 20]. Metoda GBSC, którą zaprojektowałem i opracowałem w ramach tej pracy, jest objęta grantem OPUS nr 2020/39/B/ST6/03447, w którym jestem współwykonawcą. Dodatkowo swoje doświadczenia z analiz już istniejących metod porównywania podobieństw białek wykorzystałam do zaprojektowania nowej metody będącej częścią grantu PRELUDIUM nr 2021/41/N/ST6/01919, którego jestem kierownikiem.

3 Metody porównania fragmetnów o niskiej złożoności

Do określenia funkcji fragmentu białka można skorzystać z zasady przechodniości. Innymi słowy, jeśli chcemy odgadnąć funkcję fragmentu sekwencji aminokwasowej w białku to możemy znaleźć podobną sekwencje o znanej funkcji i biorąc pod uwagę dodatkowe czynniki takie jak lokalizacja albo struktura, wnioskować o tym czy fragment o nieznanej funkcji będzie miał ją podobną. Również posiadanie wielu podobnych sekwencji jest pomocne, jeśli chcemy odkryć biologiczną rolę białka bez adnotacji. Wtedy możemy znacznie obniżyć koszty i czas metod eksperymentalnych do odkrycia rzeczywistej roli danego fragmentu białka. W ten sposób poszukiwanie podobnych sekwencji białek może pomóc w odkryciu nowych funkcji i struktur. Praca ta, pokazuje że BLAST oraz inne metody posiadają funkcjonalności, które uniemożliwiają efektywną analizę regionów o niskiej złożoności. Zidentyfikowałem powyższe funkcjonalności i zaproponowałem ich zmianę wprowadzając LCR-BLAST. Dodatakowo, niedoskonałości w grupowaniu podobnych LCR-ów z metody CD-HIT wykorzystałem do stworzenia nowej metody grupowania sekwencji zawierających krótkie powtórzenia tandemowe.

3.1 LCR-BLAST

W ramach tej pracy przeanalizowałem parametry BLAST-a, zaproponowałem ich optymalne wartości oraz zmodyfikowałem samą metodę. Każdy parametr jest omówiony z ukazaniem jego wpływu na wyniki porównania LCR-ów. Wykorzystałem następujące kombinacje modyfikacji w celu zaprezentowania wyników:

- DEF-BLAST – parametry domyślne
- SHORT-BLAST – zmieniono parametr *task* na *blastp-short*
- COMP-BLAST – wyłączono przeliczanie macierzy punktacji na podstawie składu aminokwasowego.
- SHORT-COMP-BLAST – zmieniono parametr *task* na *blastp-short* i wyłączono przeliczanie macierzy punktacji na podstawie składu aminokwasowego.
- LCR-BLAST – zmieniono parametr *task* na *blastp-short* i wyłączono przeliczanie macierzy punktacji na podstawie składu aminokwasowego. Dodatkowo uproszczono macierz punktacji i wprowadzono dodatkową metrykę do oceny dopasowań.

BLAST oblicza E-wartość jako główną statystykę do porównywania sekwencji białkowych. Na poniższym równaniu pokazano, jak obliczyć tę wartość.

$$E = (n * m)/(2^{S'}) \tag{0.1}$$

Gdzie n to całkowita liczba reszt aminokwasowych w bazie danych, a m to długość sekwencji zapytania. S' to wynik dopasowania po normalizacji, który pokazuje podobieństwo między dwiema sekwencjami białkowymi lub nukleotydowymi. Głównym zadaniem E-wartości jest oszacowanie, czy dwie sekwencje mają wspólnego przodka. Intuicyjnie, odpowiada ona na pytanie: Ile razy można się spodziewać podobnego lub lepszego dopasowania w zadanej bazie sekwencji? Dlatego im mniejsza wartość, tym bardziej prawdopodobne jest, że dwie sekwencje mają wspólnego przodka. Nie mniej jednak LCR-y często są istotnym elementem do tego, żeby białko mogło pełnić swoją funkcję, która nie zależy od ich częstotliwości w bazie [19]. W związku z tym, znormalizowana ocena dopasowania sekwencji będzie bardziej informacyjna dla fragmentów o niskiej złożoności niż E-wartość. Znormalizowaną ocenę dopasowania można opisać za pomocą następującego wzoru.

$$S' = (\lambda * S – ln(K))/ln(2) \tag{0.2}$$

W tym równaniu λ i K służą do uniezależniania znormalizowanego oceny od parametrów użytych do utworzenia macierzy punktacji i kar za przerwy. W efekcie tak powstałe wyniki można porównać między dopasowaniami utworzonymi przy użyciu różnych parametrów. Krótkie LCR-y często odgrywają ważną rolę w białkach, natomiast podczas analizowania podobnych sekwencji ocena dopasowań zgłasza wyniki w następująco [9]: w pierwszej kolejności, zgłasza długie identyczne sekwencje do zapytania, następnie sekwencje z kilkoma niedopasowaniami, aż w końcu zmniejsza długość dopasowania wraz ze wzrostem liczby mutacji w długich dopasowaniach. Badacz, który analizuje domeny o niskiej złożoności, może być zainteresowany krótszymi i idealnymi dopasowaniami zamiast długimi, ale bardziej zmutowanymi. Jednym ze sposobów na uniezależnienie oceny dopasowania od jego długości jest podzielenie jego oceny przez jego długość, co można opisać następującym wzorem.

$$M = S'/L \tag{0.3}$$

Gdzie M to ocena uśredniona, S' ocena znormalizowana, a L to długość dopasowania. W celu zweryfikowania równania, porównałem je z wynikami zebranymi za pomocą E-wartości.

3.2 GBSC

GBSC identyfikuje STR-y, skanując sekwencję i budując grafy podobne do De Bruijna, które reprezentują odnalezione powtórzenia [7]. Do identyfikacji STR-ów, GBSC iteruje po k-merach sekwencji w kolejności, w jakiej się one pojawiają. W każdej iteracji tworzy węzeł z bieżącego k-mera i łączy poprzedni węzeł z bieżącym, tworząc między nimi krawędź. Każda krawędź ma czas życia i wagę z początkowymi wartościami odpowiednio 0 i 1. Czas życia wzrasta na wszystkich istniejących krawędziach za każdym razem, gdy algorytm przechodzi do następnego k-mera. Jeśli czas życia krawędzi osiągnie próg określony przez użytkownika, znika. W przypadku, gdy istnieje już krawędź między bieżącym a poprzednim k-merem, czas życia krawędzi jest resetowany do 0, a jej waga wzrasta o 1. Jeśli zanikająca krawędź ma przypisaną wagę powyżej progu określonego przez użytkownika, wówczas wszystkie te krawędzie tworzą graf reprezentujący znaleziony STR. Pierwsze i ostatnie k-mery należące do grafu określają granice zidentyfikowanego STR-a. Dla nieskończonego progu życia algorytm tworzy graf De Bruijna sekwencji zawierający wagi, a następnie usuwa krawędzie o wadze poniżej progu. Wynikiem tego procesu są zidentyfikowane STR-y

i grafy reprezentujące każdy fragment zawierający powtórzenia.

Grupowanie w GBSC jest kolejnym krokiem w analizie STR-ów, który opiera się na zidentyfikowanych fragmentach i przypisanych im modelach. Mając fragmenty STR z przypisaną ich reprezentacją grafową, algorytm grupuje sekwencje według odpowiednich grafów. Zapewnia, że każda grupa zawiera sekwencje składające się z podobnego wzorca. Z drugiej strony pozwala na zróżnicowanie sekwencji w klastrach, które jest wprowadzane przez losowe reszty aminokwasowe w powtarzających się wzorach. Takie losowe reszty czasami podążają za ustalonym wzorcem w domenie STR, dlatego można je uznać za integralną część STR-a [32]. Inną powszechną konfiguracją STR-ów jest sytuacja, w której jeden typ STR występuje obok innego typu. Przykład takich sąsiadujących STR-ów można zobaczyć na rysunku 0.1, gdzie po powtórzeniach QA następuje domena poli-Q. GBSC obsługuje sąsiadujące STR-y, przypisując je do grup z oddzielnymi i połączonymi typami. W przypadku, gdy po STR-rze A następuje STR B, to te STR-y przyłączą się do trzech grup: A, B i A-B. Te operacje sprawiają, że GBSC jest w stanie grupować fragmenty białek zawierających szeroką gamę STR-ów.

4 Podsumowanie i wnioski

W ramach tej pracy przedstawiłem metody analiz LCR-ów w białkach oraz porównałem istniejące rozwiązania. Na początku porównałem istniejące metody identyfikacji LCR-ów i pokazałem, jak je połączyć, aby uzyskać nowe wyniki za pomocą konsensusu. Wybrane metody, wraz z metodą konsensusu, zostały również wykorzystane do opracowania podejścia wizualizacyjnego do eksploracji LCR-ów. Naukowcy mogą wykorzystywać te wizualne analizy eksploracyjne do formułowania nowych hipotez dotyczących właściwości biologicznych białek. LCR-y były przez długi czas ignorowane, co doprowadziło do wysunięcia hipotezy, że istniejące metody analizy podobieństwa białek są przeznaczone głównie do analizy HCR sekwencji białkowych. W związku z tym następnie przeanalizowałem trzy najpowszechniejsze metody porównywania sekwencji białek, aby sprawdzić, czy hipoteza ta jest poprawna. Wyniki potwierdziły tę hipotezę, pokazując, że metody te zostały opracowane dla domen białkowych o dużej złożoności i nie są wystarczająco wydajne, aby porównać LCR-y, nawet jeśli zoptymalizujemy ich parametry. Należy to wziąć pod uwagę przy stosowaniu tych metod do poszukiwania podobnych sekwencji białkowych za-

wierających fragmenty o małej złożoności. Następnie pokazałem, jak dostosować parametry BLAST-a, aby lepiej wyszukiwać podobne LCR-y i wprowadziłem nową metodę LCR-BLAST, która jest modyfikacją BLAST-a do fragmentów białek o niskiej złożoności. Wykazałem, że każda z proponowanych modyfikacji zmienia metodę w taki sposób, że daje ona lepsze wyniki dla LCR-ów. Następnie przedstawiłem GBSC, czyli nową metodę identyfikowania i grupowania STR-ów. Metoda ta umożliwia analizę szerokiego zakresu STR-ów, od bardzo czystych do wysoce zdegenerowanych. Na koniec przedstawiłem trzy badania, w których wykorzystałem w praktyce LCR-BLAST i GBSC. Badania te są związane z błędami składania sekwencjonowania w powtórzeniach tandemowych, oddziaływaniami elektrostatycznymi w polimerazie RNA i wspólnymi LCR-ami znajdujących się w proteomach ludzkich i SARS-CoV-2. W pierwszym badaniu wykorzystałem GBSC do pokazania różnic we fragmentach STR pomiędzy wersjami sekwencji białek oraz różnorodność STR-ów w wybranych taksonomiach. Do analizy domeny K-D/E w polimerazie RNA, użyłem LCR-BLAST i GBSC, aby znaleźć ten motyw w innych białkach. Pokazałem również, że GBSC jest w stanie wykryć nieregularne LCR-y na przykładzie motywu D/E. W ostatnim badaniu, dotyczącym SARS-CoV-2, obie metody wykorzystano do znalezienia wspólnych LCR-ów w ludzkich i wirusowych proteomach, aby ostrzec naukowców przed możliwym ryzykiem selekcji pewnych epitopów do leków i szczepionek. Uważam, że wniosłem cenny wkład naukowy w wiedzę o LCR-ach i metodach ich analizy, który będę kontynuował w kolejnych latach.

Bibliografia

[1] Albà, M. M., Laskowski, R. A., Hancock, J. M. Detecting cryptically simple protein sequences using the simple algorithm. *Bioinformatics*, 18(5):672–678, 2002.

[2] Aledo, J. C. A census of human methionine-rich prion-like domain-containing proteins. *Antioxidants*, 11(7):1289, 2022.

[3] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[4] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[5] Battistuzzi, F. U., Schneider, K. A., Spencer, M. K., Fisher, D., Chaudhry, S., Escalante, A. A. Profiles of low complexity regions in apicomplexa. *BMC evolutionary biology*, 16(1):1–12, 2016.

[6] Cascarina, S. M., King, D. C., Osborne Nishimura, E., Ross, E. D. Lcd-composer: an intuitive, composition-centric method enabling the identification and detailed functional mapping of low-complexity domains. *NAR genomics and bioinformatics*, 3(2):lqab048, 2021.

[7] De Bruijn, N. G. A combinatorial problem. *Proc. Koninklijke Nederlandse Academie van Wetenschappen*, wolumen 49, strony 758–764, 1946.

[8] Eddy, S. R. Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195, 2011.

[9] Espinosa Angarica, V., Ventura, S., Sancho, J. Discovering putative prion sequences in complete proteomes using probabilistic representations of q/n-rich domains. *BMC genomics*, 14(1):1–17, 2013.

[10] Gruca, A., Ziemska-Legiecka, J., Jarnot, P., Sarnowska, E., Sarnowski, T. J., Grynberg, M. Common low complexity regions for sars-cov-2 and human proteomes as potential multidirectional risk factor in vaccine development. *BMC bioinformatics*, 22(1):1–18, 2021.

[11] Haerty, W., Golding, G. B. Low-complexity sequences and single amino acid repeats: not just "junk" peptide sequences. *Genome*, 53(10):753–762, 2010.

[12] Harrison, P. M. flps: Fast discovery of compositional biases for the protein universe. *Bmc Bioinformatics*, 18(1):1–9, 2017.

[13] Henikoff, S., Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[14] Hughes, M. P., Sawaya, M. R., Boyer, D. R., Goldschmidt, L., Rodriguez, J. A., Cascio, D., Chong, L., Gonen, T., Eisenberg, D. S. Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. *Science*, 359(6376):698–701, 2018.

[15] Jarnot, P., Ziemska-Legiecka, J., Dobson, L., Merski, M., Mier, P., Andrade-Navarro, M. A., Hancock, J. M., Dosztányi, Z., Paladin, L., Necci, M., i in. Platoloco: the first web meta-server for visualization and annotation of low complexity regions in proteins. *Nucleic acids research*, 48(W1):W77–W84, 2020.

[16] Jarnot, P., Ziemska-Legiecka, J., Grynberg, M., Gruca, A. Lcr-blast—a new modification of blast to search for similar low complexity regions in protein sequences. *International Conference on Man–Machine Interactions*, strony 169–180. Springer, 2019.

[17] Jarnot, P., Ziemska-Legiecka, J., Grynberg, M., Gruca, A. Insights from analyses of low complexity regions with canonical methods for protein sequence comparison. *Briefings in Bioinformatics*, 23(5):bbac299, 2022.

[18] Jorda, J., Kajava, A. V. T-reks: identification of tandem repeats in sequences with a k-means based algorithm. *Bioinformatics*, 25(20):2632–2638, 2009.

[19] Kiledjian, M., Dreyfuss, G. Primary structure and binding activity of the hnrnp u protein: binding rna through rgg box. *The EMBO journal*, 11(7):2655–2664, 1992.

[20] Kubáň, V., Srb, P., Štégnerová, H., Padrta, P., Zachrdla, M., Jaseňáková, Z., Šanderová, H., Vítovská, D., Krasny, L., Koval', T., i in. Quantitative conformational analysis of functionally important electrostatic interactions in the intrinsically disordered region of delta subunit of bacterial rna polymerase. *Journal of the American Chemical Society*, 141(42):16817–16828, 2019.

[21] Kumari, B., Kumar, R., Chauhan, V., Kumar, M. Comparative functional analysis of proteins containing low-complexity predicted amyloid regions. *PeerJ*, 6:e5823, 2018.

[22] Li, X., Kahveci, T. A novel algorithm for identifying low-complexity regions in a protein sequence. *Bioinformatics*, 22(24):2980–2987, 2006.

[23] Martin, E. W., Mittag, T. Relationship of sequence and phase separation in protein low-complexity regions. *Biochemistry*, 57(17):2478–2487, 2018.

[24] Newman, A. M., Cooper, J. B. Xstream: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC bioinformatics*, 8(1):1–19, 2007.

[25] Pizzi, E., Frontali, C. Low-complexity regions in plasmodium falciparum proteins. *Genome Research*, 11(2):218–229, 2001.

[26] Promponas, V. J., Enright, A. J., Tsoka, S., Kreil, D. P., Leroy, C., Hamodrakas, S., Sander, C., Ouzounis, C. A. Cast: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics*, 16(10):915–922, 2000.

[27] Radó-Trilla, N., Albà, M. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC evolutionary biology*, 12(1):1–10, 2012.

[28] Remmert, M., Biegert, A., Hauser, A., Söding, J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012.

[29] Shin, S. W., Kim, S. M. A new algorithm for detecting low-complexity regions in protein sequences. *Bioinformatics*, 21(2):160–170, 2005.

[30] Smith, T. F., Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[31] Söding, J. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21(7):951–960, 2005.

[32] Sorushanova, A., Delgado, L. M., Wu, Z., Shologu, N., Kshirsagar, A., Raghunath, R., Mullen, A. M., Bayon, Y., Pandit, A., Raghunath, M., i in. The collagen suprafamily: from biosynthesis to advanced biomaterial development. *Advanced materials*, 31(1):1801651, 2019.

[33] Stowell, J. A., Wagstaff, J. L., Hill, C. H., Yu, M., McLaughlin, S. H., Freund, S. M., Passmore, L. A. A low-complexity region in the yth domain protein mmi1 enhances rna binding. *Journal of Biological Chemistry*, 293(24):9210–9222, 2018.

[34] Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., i in. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic acids research*, 47(21):10994–11006, 2019.

[35] Wootton, J. C., Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry*, 17(2):149–163, 1993.

[36] Yu, Y.-K., Altschul, S. F. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, 21(7):902–911, 2005.