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

W niniejszej rozprawie doktorskiej skupiłem się na problemie porównywania regionów o niskiej złożoności (ang. low complexity regions - LCRs) sekwencji białkowych. Regiony te są ważne dla wielu funkcji w białkach, takich jak wiązanie RNA, jednak naukowcy przez długi czas je ignorowali, skupiając się głównie na fragmentach o wysokiej złożoności. Na tej podstawie została sformułowana hipoteza, że metody analizy podobieństwa sekwencji białek są przeznaczone przede wszystkim dla regionów o dużej złożoności i nie są wystarczająco wydajne do porównania LCR. W ramach niniejszej rozprawy potwierdziłem tę hipotezę i zaproponowałem nowe metody efektywnego porównywania LCR.

Pierwszym krokiem w tworzeniu nowych metod analizy podobieństw między LCR było sprawdzenie, czy istniejące metody radzą sobie z nimi prawidłowo. W tym celu wykorzystałem metody BLAST, HHblits oraz CD-HIT do znalezienia podobnych LCR-ów, po czym przeanalizowałem uzyskane wyniki. Wybrane metody są uważane za złoty standard do znajdowania podobnych sekwencji białkowych, a inne metody często bazują na podobnych modelach statystycznych. Wyniki pokazały, że niektóre rozwiązania stosowane w tych metodach są skuteczne tylko do porównania fragmentów białek o wysokiej złożoności i są mało wydajne do porównania LCR nawet po optymalizacji ich parametrów. Spostrzeżenia z przeprowadzonej analizy wykorzystałem do opracowania nowych metod, którymi są LCR-BLAST oraz GBSC. LCR-BLAST wywodzi się z metody BLAST, która została dostosowana do wyszukiwania LCR poprzez dostosowanie parametrów BLAST, uproszczenie macierzy punktacji i dodanie nowej metryki do oceny podobieństwa między sekwencjami. Z kolei GBSC jest nową metodą zaprojektowaną do identyfikacji i grupowania wyraźnych i zdegenerowanych krótkich powtórzeń tandemowych (ang. short tandem repeat - STRs) w sekwencjach białkowych. Oba powyższe typy STR obejmują dużą część sekwencji uważanych za LCR, pozostawiając tylko niewielką część, w której brakuje nawet rozmytych powtórzeń. Pokazałem również, że ten niewielki ułamek sekwencji LCR pominiętych przez GBSC można jeszcze bardziej zmniejszyć, redukując alfabet kanonicznych aminokwasów. GBSC porównałem z innymi metodami identyfikacji LCR oraz z wybranymi metodami grupowania sekwencji białek, żeby pokazać podobieństwa między sekwencjami, które nie są wychwytywane przez istniejące metody. Obie metody zostały wykorzystane w trzech różnych analizach LCR w białkach, aby pokazać ich przydatność. W pierwszym badaniu użyłem GBSC, aby

pokazać, że w bazach danych białek mogą występować błędy składania sekwencji. W kolejnym badaniu użyłem LCR-BLAST i GBSC do znalezienia motywów K-D/E, które są podobne do motywu znalezionego w polimerazie RNA. Wreszcie, obie metody zostały również wykorzystane do znalezienia wspólnych LCR w proteomach SARS-CoV-2 i ludzkim, aby ostrzec, że wybranie pewnych epitopów dla leków i szczepionek może spowodować chorobę autoimmunologiczną.