Sekwencyjna klasyfikacja struktur białek metodą hierarchiczną

Mateusz Skłodowski[1], Joanna M. Macnar[1,2], Dominik Gront[1]

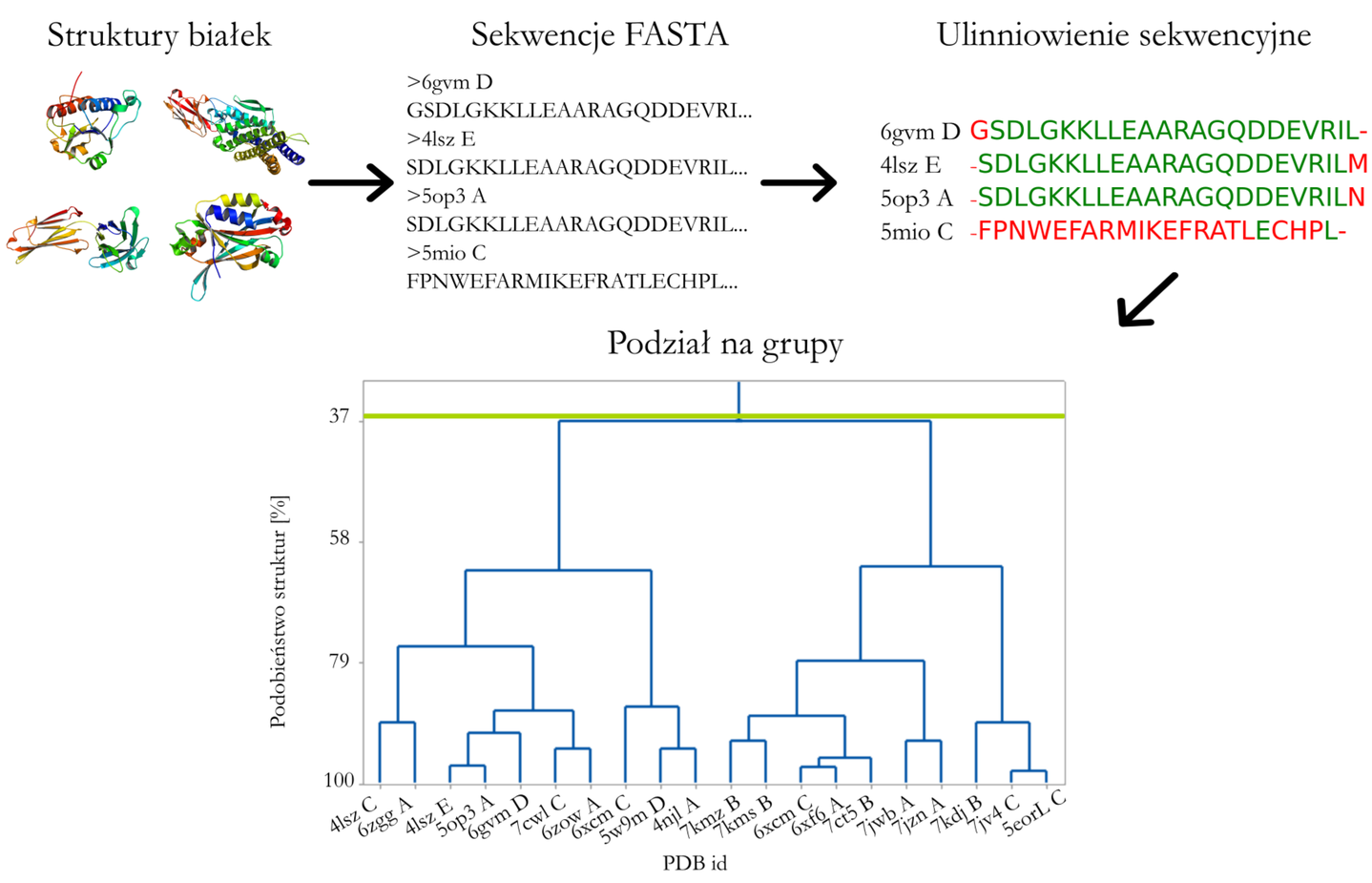
1 Faculty of Chemistry, University of Warsaw, Pasteura 1, Warsaw, Poland

2 College of Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Stefana Banacha 2C, Warsaw, Poland

Wstęp

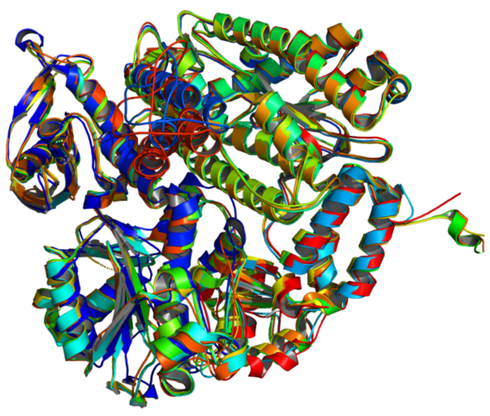
W ostatnich latach obserwujemy gwałtowny rozrost baz danych poświęconych polipeptydom [1]. Towarzyszy temu nierównomierny rozwój badań nad różnymi rodzinami białkowymi oraz ich motywami strukturalnymi. W konsekwencji niektóre grupy są nadreprezentowane, istotnie wpływając na opracowywane na ich podstawie statystyki. Jednym z rozwiązań tego problemu jest wybór reprezentanta, będącego uśrednionionym białkiem dla każdej z grup. W ten sposób powstają bazy białek bez nadmiarowych depozytów (ang. non reduntant), których przykładami mogą być bazy ASTRAL [2] oraz PISCES [3]. Podejście to powoduje jednak stratę informacji o unikatowych motywach i sekwencjach. Drugim podejściem pozwalającym zachować komplet danych jest wykorzystanie systemu wag określających podobieństwo reprezentanta z dowolnym białkiem z grupy [4]. W tym celu można zastosować analizę skupień, korzystając z metod hierarchicznych [5]. To podejście zostało przyjęte w naszej pracy.

Schemat przeprowadzonej analizy



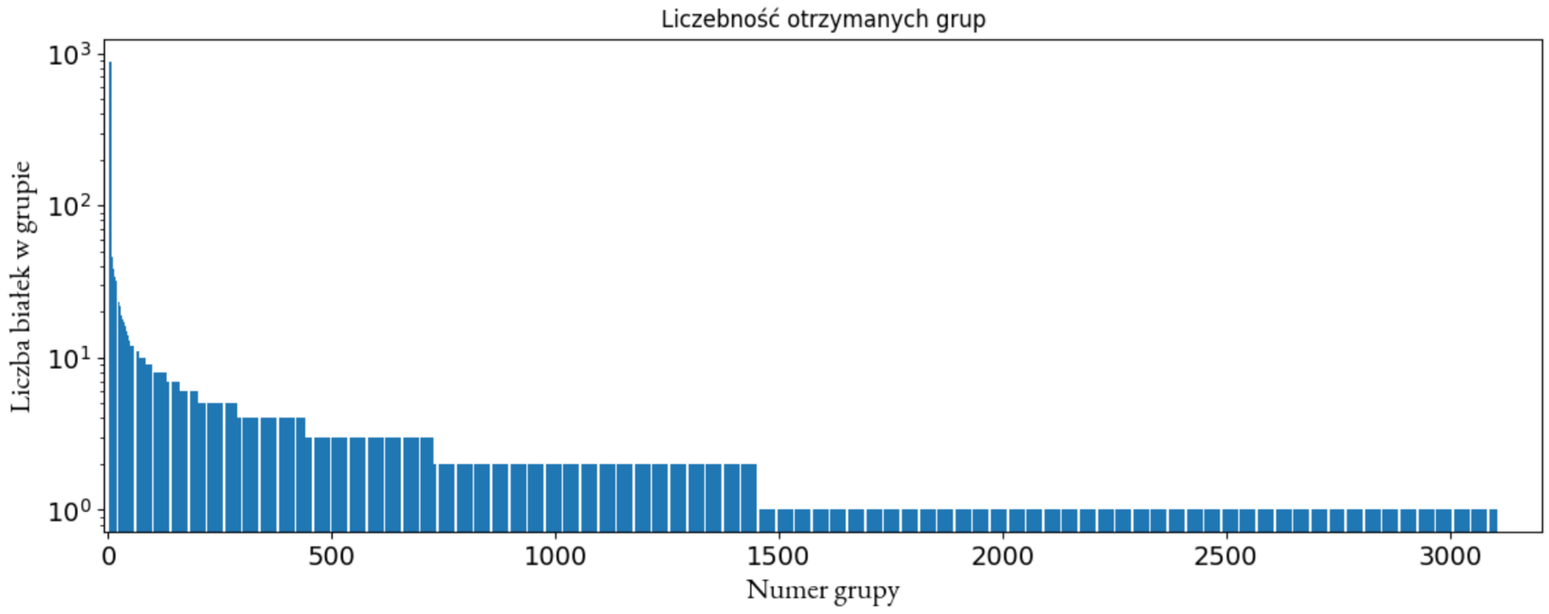
Wyniki

a)



b)

Przykłady nałożonych struktur z dwóch grup: trasferazy DNA (a) oraz hydrolazy (b).



Reprezentant grupy	Przykładowe białka	Liczebność grupy	Opis grupy
7lsi	6aw3, 3i5h	886	Przeciwciała oraz białka transportowe
4dxc	6nfk, 4row	62	Hydrolazy
2b7x	1ssy, 4bv0	58	Hydrolazy oraz białka sygnałowe
6d7g	3kyo, 4zuw	46	Białka głównego układu zgodności tkankowej
3hlv	7jhd, 5kro	45	Białka biorące udział w transkrypcji
2wr7	1mqn, 5vtx	39	Hemaglutyniny
6duf	6uk0, 3hvt	38	Odwrotne transkryptazy
7jwb	7cah, 7ct5	35	Białka Spike koronawirusów
5v8l	5y14, 2cmr	35	Białka związane z wirusem HIV
6mnx	6p0l, 5cm9	34	Wewnątrzkomórkowe białka sygnałowe związane z GDP

Zastosowanie aplikacyjne

Uzyskane wagi oraz reprezentantów można wykorzystać do stworzenia biblioteki szablonów do modelowania porównawczego. Jednocześnie przy użyciu takiej biblioteki możemy wykonać modelowanie przy użyciu wielu szablonów.

Literatura

[1] H. M. Berman et al., “The Protein Data Bank,” Nucleic Acids Research. 2000, doi: 10.1093/nar/28.1.235.

[2] S. E. Brenner, P. Koehl, and M. Levitt, “The ASTRAL compendium for protein structure and sequence analysis,” Nucleic Acids Research. 2000, doi: 10.1093/nar/28.1.254.

[3] G. Wang and R. L. Dunbrack, “PISCES: A protein sequence culling server,” Bioinformatics, 2003, doi: 10.1093/bioinformatics/btg224.

[4] C. Yanover, N. Vanetik, M. Levitt, R. Kolodny, and C. Keasar, “Redundancy-weighting for better inference of protein structural features,” Bioinformatics, 2014, doi: 10.1093/bioinformatics/btu242.

[5] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: An overview,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 2012, doi: 10.1002/widm.53.