

# Tworzenie skróconych alfabetów aminokwasowych dla białek amyloidogennych

Projekt badawczy Doktoranckiego Koła Naukowego Bioinformatyki

Michał Burdukiewicz, Przemysław Gagat

## 1 Założenia projektu badawczego

Sekwencje białkowe zazwyczaj są opisywane przy wykorzystaniu alfabetu aminokwasowego zawierającego 20 aminokwasów. W przypadku wielu problemów badawczych, np. przewidywania funkcji białka (Longo et al., 2013) lub jego fałdowania (Murphy et al., 2000), znajomość dokładnego składu aminokwasowego nie jest konieczna i trafnych analiz można dokonywać na podstawie skróconego alfabetu aminokwasowego.

Skrócony alfabet aminokwasowy to alfabet w którym aminokwasy na podstawie określonych podobieństw są przypisane do grup (Tab. 1). Zazwyczaj w celu stworzenia takiego alfabetu wykorzystuje się macierze podstawień (Cannata et al., 2002) lub właściwości fizykochemiczne aminokwasów (Stephenson & Freeland, 2013). Istnieją również kryteria pozwalające ustalić optymalną wielkość alfabetu (liczbę grup, do których przypisujemy aminokwasy) (Solis, 2015).

Tab. 1: Przykładowy skrócony alfabet aminokwasowy (Melo & Marti-Renom, 2006). Sekwencja ADPH w tym alfabecie zostanie zapisana jako 1335.

Numer grupy	Aminokwasy
1	A, G
2	C
3	D, E, K, N, P, Q, R, S, T
4	F, I, L, M, V, W, Y
5	H

Nasze wstępne wyniki sugerują, że w przypadku niektórych problemów przewidywania właściwości białek, zastosowanie skróconego alfabetu aminokwasowego znacząco ulepszyć precyzję predykcji. Jednakże taki alfabet musi zostać wybrany spośród wielu potencjalnych skróconych alfabetów. Liczba wszystkich możliwych skróconych alfabetów aminokwasowych  $n_s$  zawierających  $k$  grup jest wyrażona poprzez:

$$n_s = \frac{1}{k!} \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} j^{20}$$

Nawet przy założeniu, że optymalna wielkość skróconego alfabetu jest już znana, to liczba potencjalnych możliwości jest bardzo duża (np. wszystkich alfabetów sześcioelementowych jest  $4.306079 \times 10^{12}$ ).

## 2 Obecny stan badań

### 2.1 Predykcja białek amyloidogennych

Jednym z poprzednich zadań badawczych naszego Koła było opracowanie algorytmu przewidującego amyloidy, białka, których agregaty występują w różnych zaburzeniach neurodegeneracyjnych (m.in. choroby Alzheimera, Creutzfelda-Jacoba). Podczas tworzenia predyktora amyloidogenności AmyloGram wygenerowaliśmy i przetestowaliśmy 18 535 unikalnych skróconych alfabetów o długości od 3 do 6 (ok.  $3.63\% \times 10^{-7}$  wszystkich możliwych skróconych alfabetów o długości od 3 do 6). Spośród sprawdzonych alfabetów wybrano ten, który gwarantował najlepszą predykcję.

Zastosowanie skróconego alfabetu istotnie podwyższyło skuteczność przewidywania amyloidogenności (Tab. 2). Jednakże ze względu na szczupłość sprawdzonego zbioru alfabetów, prawdopodobnie istnieje skrócony alfabet, który pozwala na jeszcze lepsze predykcje oraz precyzyjniejsze opisanie procesu agregacji amyloid.

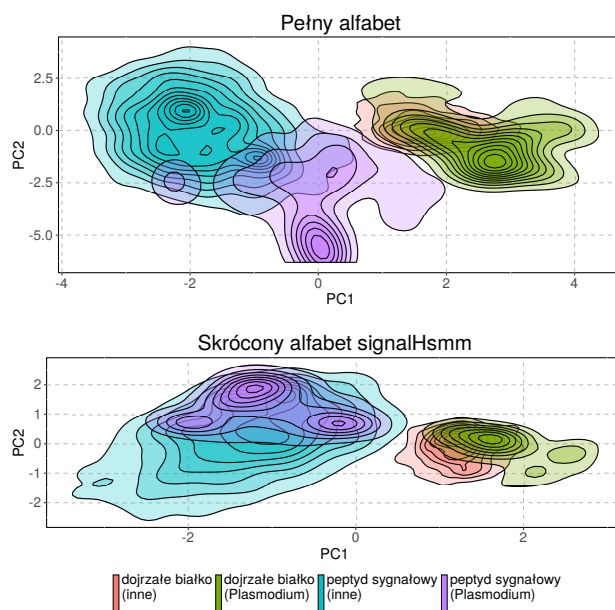
Tab. 2: Porównanie predykcji amyloidogenności na zbiorze testowym *pep424* (Walsh et al., 2014) dla klasyfikatorów uczonych na sekwencjach zapisanych z wykorzystaniem pełnego alfabetu i skróconego. AUC: Area Under the Curve. MCC: Matthew’s Correlation Coefficient.

Classifier	AUC	MCC
Skrócony alfabet (zbiór uczący A)	0.8856	0.6057
Pełny alfabet (zbiór uczący A)	0.8411	0.5427
Skrócony alfabet (zbiór uczący B)	<b>0.8972</b>	<b>0.6307</b>
Pełny alfabet (zbiór uczący B)	0.8581	0.5698
Skrócony alfabet (zbiór uczący C)	0.8728	0.5420
Pełny alfabet (zbiór uczący C)	0.8610	0.5490

### 2.2 Predykcja peptydów sygnałowych u nietypowych organizmów

Zredukowane alfabety aminokwasowe zostały również wykorzystane w naszym programie signalHsmm, który przewiduje występowanie peptydów sygnałowych

u Eukariontów. Zastosowanie skróconego alfabetu aminokwasowego pozwoliło trafnie przewidywać peptydy sygnałowe również u nietypowych grup taksonomicznych takich jak rodzaj *Plasmodiidae* do którego zalicza się zarodziec malarii. Nietypowy skład aminokwasowy takich peptydów sygnałowych utrudnia ich rozpoznawanie przy użyciu innych programów, które wykorzystują pełny alfabet aminokwasowy. Zapisanie sekwencji peptydów sygnałowych *Plasmodiidae* przy użyciu skróconego alfabetu aminokwasowego pozwala na identyfikację reguł decyzyjnych, które łączą je z analogicznymi sekwencjami pochodzących od innych eukariontów (Rys. 1). Nie jest to możliwe przy użyciu pełnego alfabetu aminokwasowego.



Tab. 1: Wyniki PCA dla częstości aminokwasów w peptydach sygnałowych i dojrzałych białkach *Plasmodiidae* oraz innych eukariontów.

### 3 Cel badań

Głównym zadaniem badawczym jest opracowanie nowej metody poszukiwania skróconych alfabetów amyloidogennych. Nowe rozwiązania zostaną zastosowane do problemu przewidywania białek amyloidogennych i nietypowych peptydów sygnałowych.

## 4 Metody

Głównym narzędziem wykorzystywanym w projekcie badawczym jest pakiet *bio-gram* przeznaczony do analizy n-gramowej sekwencji biologicznych. Program ten zawiera liczne narzędzia umożliwiające tworzenie i porównywanie skróconych alfabetów, które mogą być podstawą dla nowego algorytmu opracowanego podczas projektu badawczego.

## 5 Planowane wydatki

Łączny koszt projektu badawczego to 29 802,00 zł.

Kosztorys projektu badawczego.

Nazwa	Koszt
Utworzenie studenckiego klastra obliczeniowego	25 042,00 zł
Akcesoria niezbędne w realizacji zadań badawczych	760,00 zł
Wyjazdy konferencyjne	4 000,00 zł
Łącznie	29 802,00 zł

### 5.1 Utworzenie studenckiego klastra obliczeniowego

Realizacja projektu wymaga stworzenia nowego klastra obliczeniowego na potrzeby przewidzianych zadań badawczych (Tab. 3). Po zakończeniu badań klastr pozostanie cennym narzędziem dla studentów, którzy będą go mogli wykorzystać do obliczeń niezbędnych do napisania prac dyplomowych.

Tab. 3: Koszty utworzenia studenckiego klastra obliczeniowego.

Nazwa	Cena (szt.)	Liczba	Łączna cena
Zasilacze - 500W	230,00 zł	4	920,00 zł
Bateria do UPS APC RBC7	314,00 zł	3	942,00 zł
Napęd optyczny zewnętrzny USB 2.0	150,00 zł	1	150,00 zł
Switch Netgear GS324 24 porty 10/100/1000	500,00 zł	2	1 000,00 zł
Pamięć RAM 2x8GB DDR3	330,00 zł	1	330,00 zł
Jednostka obliczeniowa (komputer)	3 500,00 zł	6	21 000,00 zł
Monitor 24"	700,00 zł	1	700,00 zł
Łącznie:			25 042,00 zł

Tab. 4: Koszty akcesoriów niezbędnych w realizacji zadań badawczych.

Nazwa	Cena (szt.)	Liczba	Łączna cena
Pendrive USB 3.0 - 32 GB	45,00 zł	4	180,00 zł
Słuchawki z mikrofonem Creative	140,00 zł	2	280,00 zł
Plecak na laptopa	150,00 zł	2	300,00 zł
Łącznie:			760,00 zł

## 5.2 Akcesoria niezbędne w realizacji zadań badawczych

Właściwe zrealizowanie projektu badawczego wymaga również dokupienie akcesoriów (Tab. 4), takich jak pamięci USB niezbędne do przenoszenia dużych objętości danych i słuchawki z mikrofonem do prowadzenia rozmów z zagranicznym uczestnikiem projektu. Wykonanie części zadań badawczych nie byłaby możliwa gdyby nie komputery przenośne udostępnione członkom Koła przez Zakład Genomiki. Bezpieczny transport otrzymanego sprzętu wymaga zakupu specjalnych plecaków na laptopy.

## 5.3 Wyjazdy konferencyjne

Wstępne wyniki badań zostaną zaprezentowane podczas 9 Sympozjum Polskiego Towarzystwa Bioinformatycznego (28-30 września 2016, Białystok) oraz European R User Meeting (12-14 października 2016, Poznań). Dofinansowanie umożliwi większej liczbie członków Koła aktywny udział w konferencji i zaprezentowanie nie tylko wyników realizacji zadań badawczych postawionych w tym wniosku, ale również postępów w realizacji prac doktorskich. W ramach projektu łącznie odbędą się cztery wyjazdy (trzy wyjazdy na Sympozjum Polskiego Towarzystwa Bioinformatycznego i jeden wyjazd na European R User Meeting).

Kosztorys wyjazdów zagranicznych.

Nazwa	Cena (szt.)	Liczba	Łączna cena
Dofinansowanie wyjazdu	1 000 zł	4	4 000 zł
Łącznie:			4 000,00 zł

## Literatura

- Cannata, N., Toppo, S., Romualdi, C., & Valle, G. (2002, August). Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. *Bioinformatics (Oxford, England)*, 18(8), 1102–1108.
- Longo, L. M., Lee, J., & Blaber, M. (2013, February). Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. *Proceedings of the National Academy of Sciences of the United States of America*, 110(6), 2135–2139. doi: 10.1073/pnas.1219530110

- Melo, F., & Marti-Renom, M. A. (2006, June). Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. *Proteins*, 63(4), 986–995. doi: 10.1002/prot.20881
- Murphy, L. R., Wallqvist, A., & Levy, R. M. (2000, March). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3), 149–152. Retrieved 2016-01-24, from <http://peds.oxfordjournals.org/content/13/3/149> doi: 10.1093/protein/13.3.149
- Solis, A. D. (2015, December). Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins. *Proteins: Structure, Function, and Bioinformatics*, 83(12), 2198–2216. Retrieved 2016-07-14, from <http://onlinelibrary.wiley.com/doi/10.1002/prot.24936/abstract> doi: 10.1002/prot.24936
- Stephenson, J. D., & Freeland, S. J. (2013, October). Unearthing the root of amino acid similarity. *Journal of Molecular Evolution*, 77(4), 159–169. doi: 10.1007/s00239-013-9565-0
- Walsh, I., Seno, F., Tosatto, S. C. E., & Trovato, A. (2014, July). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, 42(W1), W301–W307. Retrieved 2015-07-24, from <http://nar.oxfordjournals.org/content/42/W1/W301> doi: 10.1093/nar/gku399