

Przewidywanie właściwości sekwencji biologicznych w oparciu o analizę n-gramów

Michał Burdukiewicz

Zakład Genomiki, Uniwersytet Wrocławski

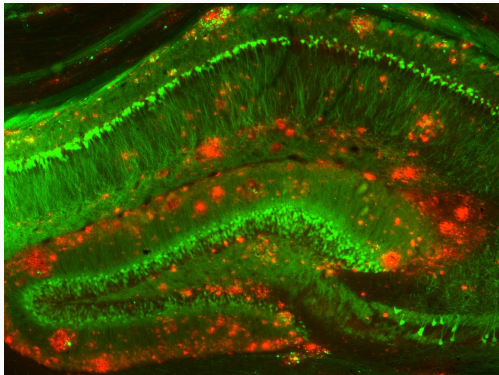
Prace eksperymentalne zazwyczaj poprzedza się analizami komputerowymi, które pozwalają optymalniej zaprojektować dalsze badania.

Przykłady:

- przewidywanie lokalizacji białek w komórce,
- modelowanie struktury przestrzennej białek oraz kwasów nukleinowych,
- wykrywanie miejsc wiązania czynników transkrypcyjnych,
- poszukiwanie sekwencji kodujących białko.

Białka amyloidowe

Białka związane z licznymi chorobami (np. choroby Alzheimera, Parkinsona, Creutzfeldta-Jakoba) tworzące szkodliwe agregaty.

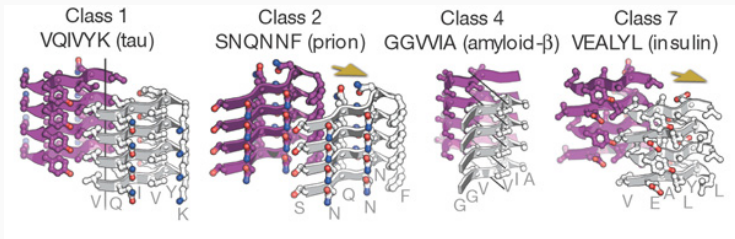


Agregaty amyloidowe (czerwony) wokół neuronów (zielony). Strittmatter Laboratory, Yale University.

Białka amyloidowe

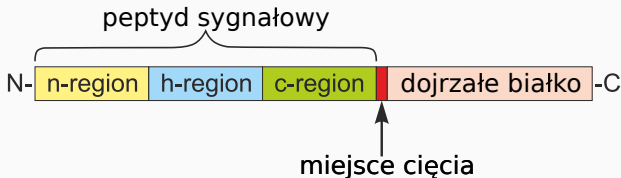
Hot-spots:

- krótkie (6-15 aminokwasów), ale bardzo zróżnicowane fragmenty białek amyloidogennych,
- miejsce inicjacji agregacji amyloidowej,
- formują specyficzne struktury β typu "zamka błyskawicznego".



Sawaya et al. (2007)

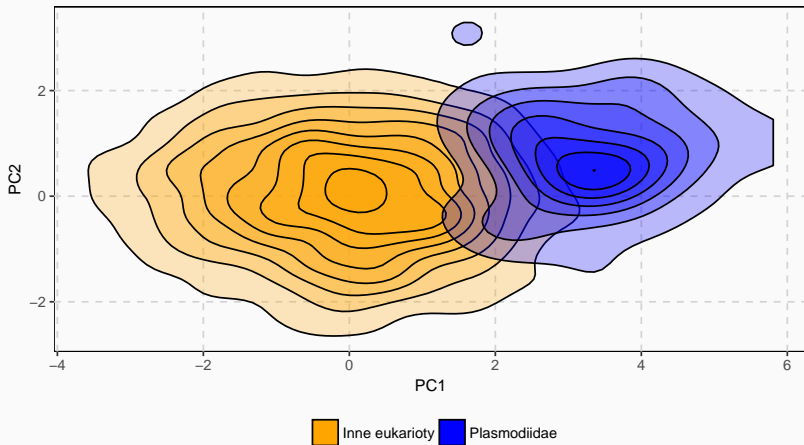
Peptydy sygnałowe



Peptydy sygnałowe:

- krótkie (15-30 aminokwasów) N-końcowe sekwencje,
- występują w białkach układu odpornościowego, strukturalnych, enzymach metabolicznych i hormonach,
- składają się z trzech regionów, gdzie preferowane są aminokwasy o określonych właściwościach fizykochemicznych.
- zróżnicowany skład aminokwasowy peptydów sygnałowych utrudnia ich rozpoznawanie.

Peptydy sygnałowe



Peptydy sygnałowe zarodźców malarii mają skład aminokwasowy różny od peptydów sygnałowych innych eukariontów.

n-gramy (k-tuple, k-mery):

- podsekwencje (ciągłe lub z przerwami) o długości n ,
- uwzględniają otoczenie danej reszty.

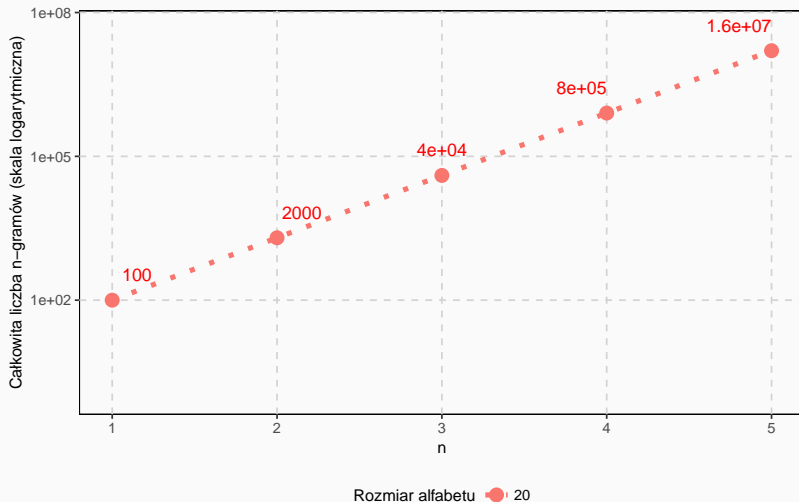
	P1	P2	P3	P4	P5
S1	M	R	K	L	Y

2-gramy: MR, RK, KL, LY

2-gramy (przerwa 1): M – K, R – L, K – Y

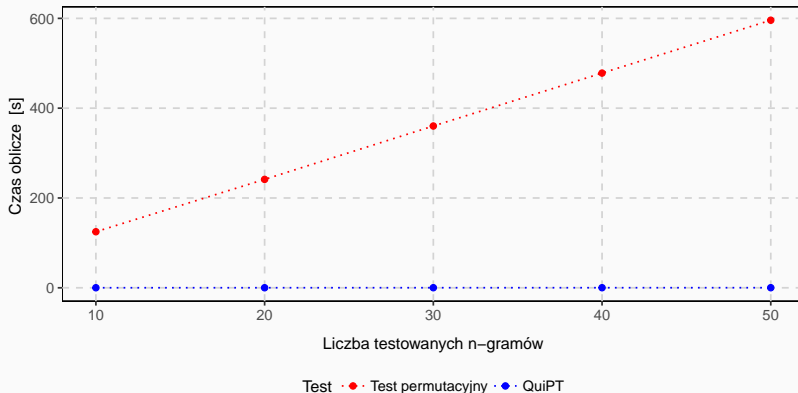
3-gramy: MRK, RKL, KLY

n-gramy



n-gramy tworzą duże i trudne do analizy zbiory danych.

QuiPT (**Q**uick **P**ermutation **T**est) szybko i efektywnie filtruje informatywne n-gramy.



QuiPT jest szybszy niż klasyczne testy permutacyjne i zwraca dokładniejsze p-wartości.

Uprozczone alfabety:

- opierają się na grupowaniu aminokwasów o podobnych właściwościach fizykochemicznych,
- ułatwiają modelowanie i przewidywanie właściwości sekwencji (Murphy et al., 2000),
- tworzą łatwiej interpretowalne modele.

Dwie sekwencje zupełnie różne pod względem składu aminokwasowego mogą być identyczne pod względem właściwości reszt.

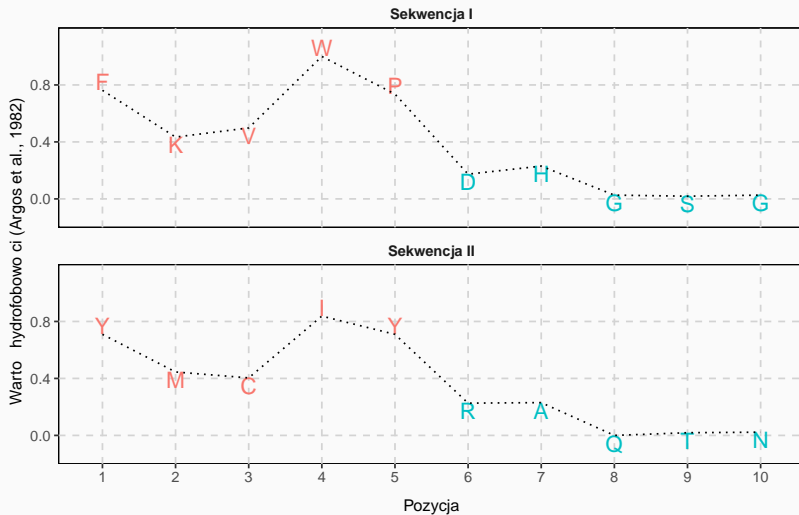
Sekwencja I:

FKVWPDHGSG

Sekwencja II:

YMCIYRAQTN

Uprozczone alfabety



Uproszczone alfabety

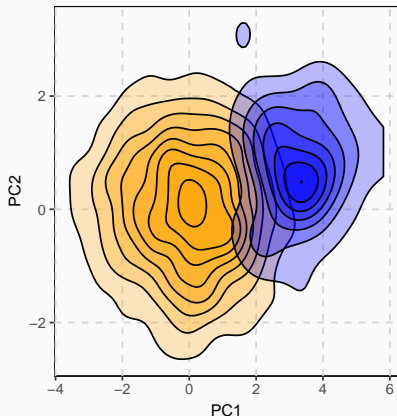
Nr podgrupy	Aminokwasy
1	C, I, L, K, M, F, P, W, Y, V
2	A, D, E, G, H, N, Q, R, S, T

Sekwencja I: FKVWPDHGSG → 1111122222

Sekwencja II: YMCIIYRAQTN → 1111122222

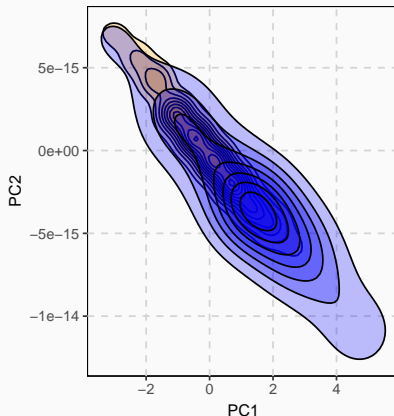
Uprozczone alfabety

Pełny alfabet



■ Inne eukarioty ■ Plasmodiidae

Zredukowany alfabet



■ Inne eukarioty ■ Plasmodiidae

PCA częstości pojedynczych aminokwasów w peptydach sygnałowych innych eukariotów i zarodźców malarii.

AmyloGram: oparte o redukcję alfabetów i kodowanie n-gramowe narzędzie do predykcji białek amyloidogennych (Burdukiewicz et al., 2016).

Amyloidy

FILLTQA

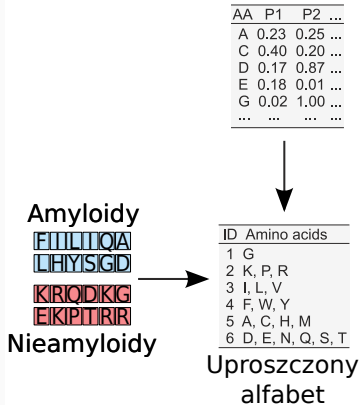
LHYSGD

KRQDKG

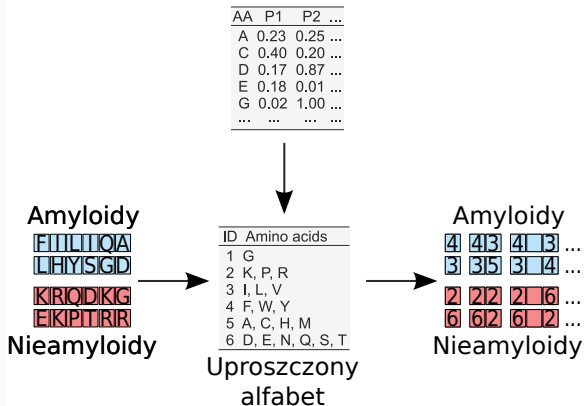
EKPTRR

Nieamyloidy

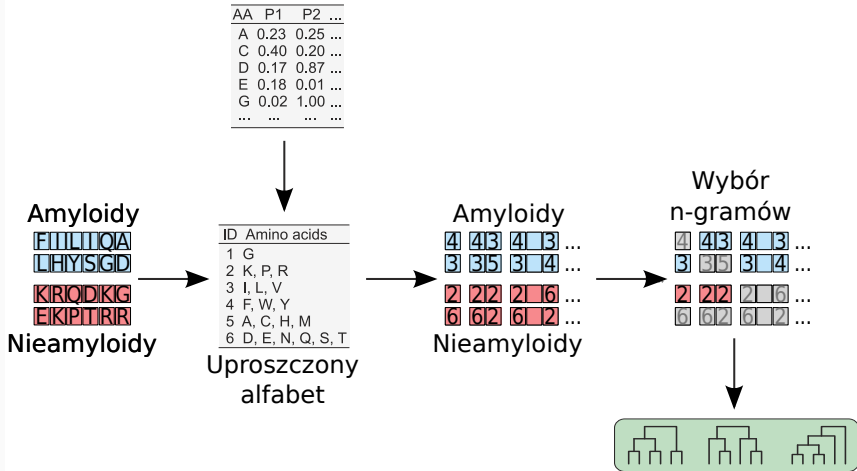
AmyloGram



AmyloGram



AmyloGram



Porównanie z innymi klasyfikatorami

Klasyfikator	AUC	MCC
AmyloGram	0.8972	0.6307
PASTA 2.0 (Walsh et al., 2014)	0.8550	0.4291
FoldAmyloid (Garbuzynskiy et al., 2010)	0.7351	0.4526
APPNN (Família et al., 2015)	0.8343	0.5823

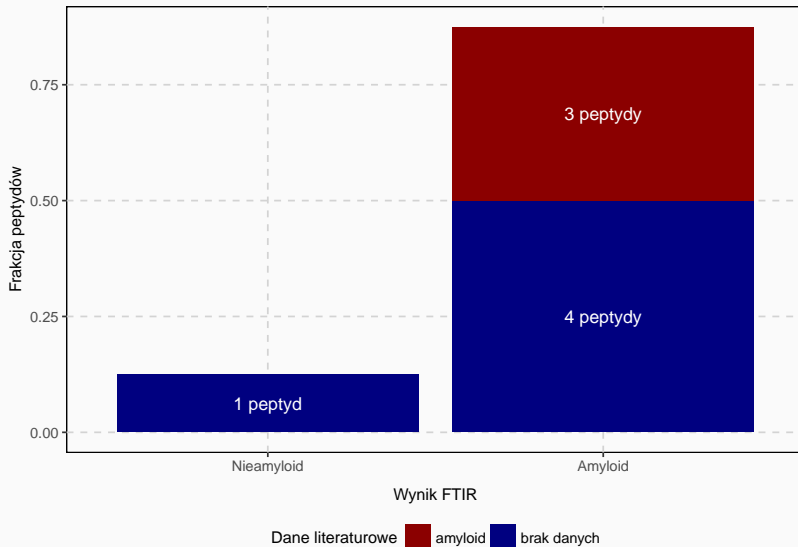
AUC (Area Under the Curve): miara jakości predykcji (1: idealny dobry klasyfikator, 0: idealnie zły klasyfikator).

MCC (Matthew's Correlation Coefficient): miara jakości predykcji (1: idealny dobry klasyfikator, -1: idealnie zły klasyfikator).

AmyloGram porównano z innymi klasyfikatorami na zewnętrznym zbiorze danych *pep424*.

1. Wszystkie nieamyloidowe peptydy z bazy AmyLoad zanalizowano AmyloGramem.
2. Wybrano 8 peptydów z najwyższym prawdopodobieństwem amyloidogenności.
3. Peptydy zanalizowano przy pomocy spektroskopii fourierowskiej (FTIR).

Walidacja eksperymentalna



1. Stworzono algorytm efektywnie selekcionujący informatywne n-gramy reprezentujące sekwencje aminokwasowe.
2. Porównano metody poszukujące uproszczone alfabetu aminokwasowe.
3. Opracowaną metodologię zastosowano do przewidywania białek amyloidogennych tworząc pakiet **R** i web server AmyloGram (<http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/>).

1. Zastosowanie opracowanej metodologii do przewidywania lokalizacji subkomórkowej białek.
2. Upublicznienie rozwijanych metod w postaci pakietu *biogram* w środowisku programistycznym i statystycznym **R**.

Literatura

- Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2016). Prediction of amyloidogenicity based on the n-gram analysis. Technical Report e2390v1, PeerJ Preprints.
- Família, C., Dennison, S. R., Quintas, A., and Phoenix, D. A. (2015). Prediction of Peptide and Protein Propensity for Amyloid Formation. *PLOS ONE*, 10(8):e0134679.

References II

- Garbuzynskiy, S. O., Lobanov, M. Y., and Galzitskaya, O. V. (2010). FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics (Oxford, England)*, 26(3):326–332.
- Murphy, L. R., Wallqvist, A., and Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3):149–152.
- Sawaya, M. R., Sambashivan, S., Nelson, R., Ivanova, M. I., Sievers, S. A., Apostol, M. I., Thompson, M. J., Balbirnie, M., Wiltzius, J. J. W., McFarlane, H. T., Madsen, A., Riek, C., and Eisenberg, D. (2007). Atomic structures of amyloid cross-spines reveal varied steric zippers. *Nature*, 447(7143):453–457.

Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. (2014).
PASTA 2.0: an improved server for protein aggregation
prediction. *Nucleic Acids Research*, 42(W1):W301–W307.