

Przewidywanie właściwości sekwencji biologicznych w oparciu o analizę n-gramów

Michał Burdukiewicz

Zakład Genomiki, Uniwersytet Wrocławski

Bioinformatyczne przewidywanie funkcji białek

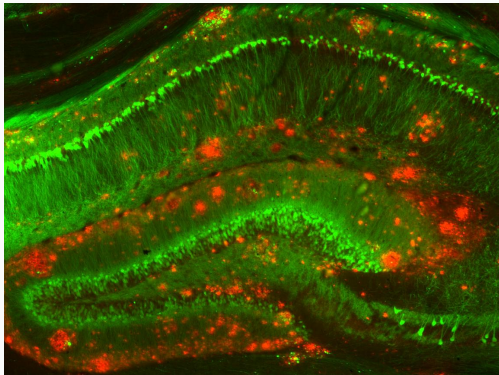
Prace eksperymentalne zazwyczaj poprzedza się analizami komputerowymi, które pozwalają optymalniej zaprojektować dalsze badania.

Przykłady:

- przewidywanie lokalizacji subkomórkowej białek (sygnałów kierujących),
- predykcja struktury drugorzędowej i trzeciorzędowej białek oraz kwasów nukleinowych,
- wykrywanie miejsc wiązania czynników transkrypcyjnych,
- poszukiwanie sekwencji kodujących białko.

Białka amyloidowe

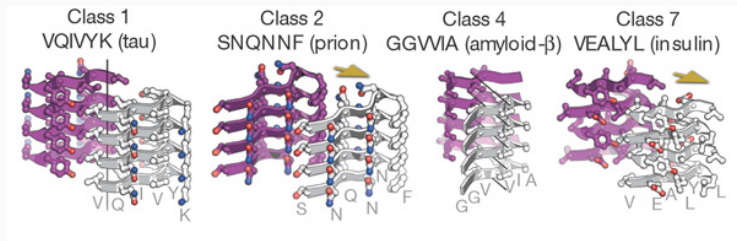
Białka związane z licznymi chorobami (np. choroby Alzheimera, Parkinsona, Creutzfeldta-Jakoba) tworzące szkodliwe agregaty.



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

Białka amyloidowe

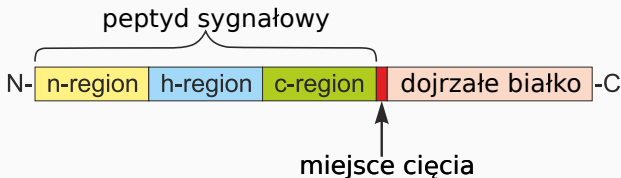
Proces agregacji jest inicjowany w obrębie tzw. hot spots, krótkich (6-15 aminokwasów), ale zróżnicowanych podsekwencji, które występują we wszystkich białkach amyloidowych i formują specyficzne struktury β typu "zamka błyskawicznego" (zipper-like).



Sawaya et al. (2007)

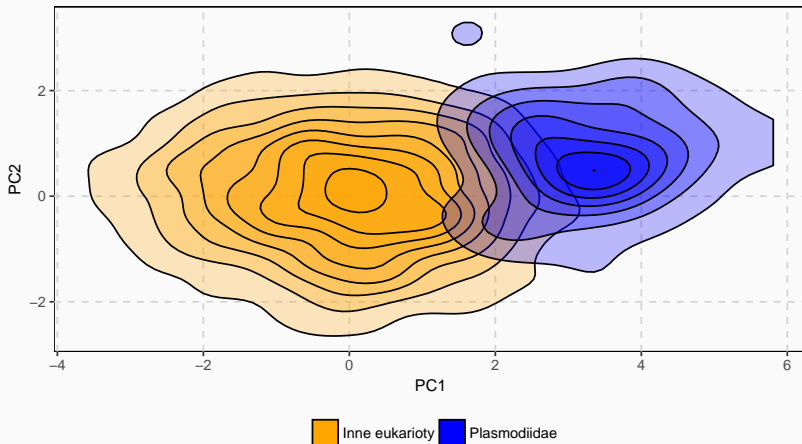
Peptydy sygnałowe

Peptydy sygnałowe to krótkie (15-30 aminokwasów) N-końcowe sekwencje kierujące białko do sekrecji.



Peptydy sygnałowe rozpoczynają się naładowanym dodatnio n-regionem, po którym występuje hydrofobowy h-region i c-region zakończony miejscem cięcia rozpoznawanym przez peptydazę sygnałową.

Peptydy sygnałowe



Peptydy sygnałowe nie wymagają konkretnych aminokwasów, ale reszt o określonych właściwościach fizykochemicznych. Przykładem mogą być peptydy sygnałowe zarodźców malarii, których skład aminokwasowy jest istotnie różny od składu innych peptydów sygnałowych eukariontów.

Metody

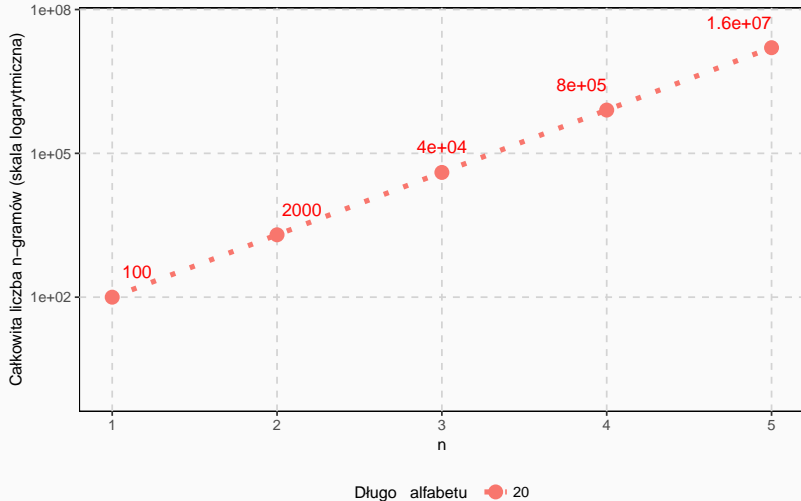
n-gramy

n-gramy (k-tuple, k-mery) to podsekwencje o długości n .

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

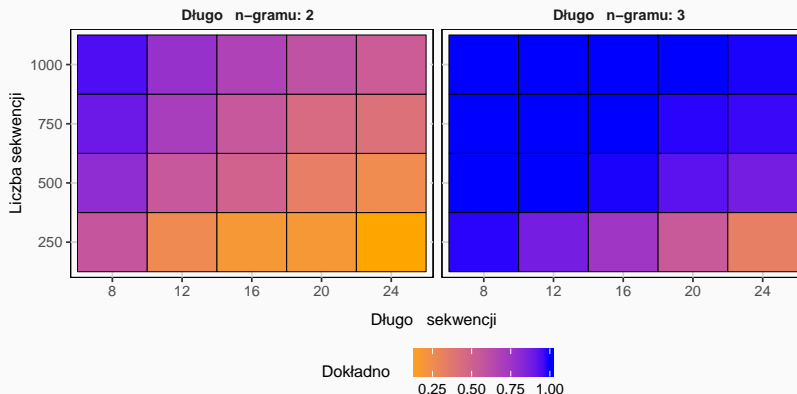
3-gramy (z przerwą między drugim i trzecim elementem): MR - L ,
MG - F , MA - G , ME - G , RK - Y , GL - N , AF - S ,
ER - A

n-gramy



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

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



Dokładność (frakcja poprawnie zidentyfikowanych motywów) QuiPT w testach symulacyjnych.

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

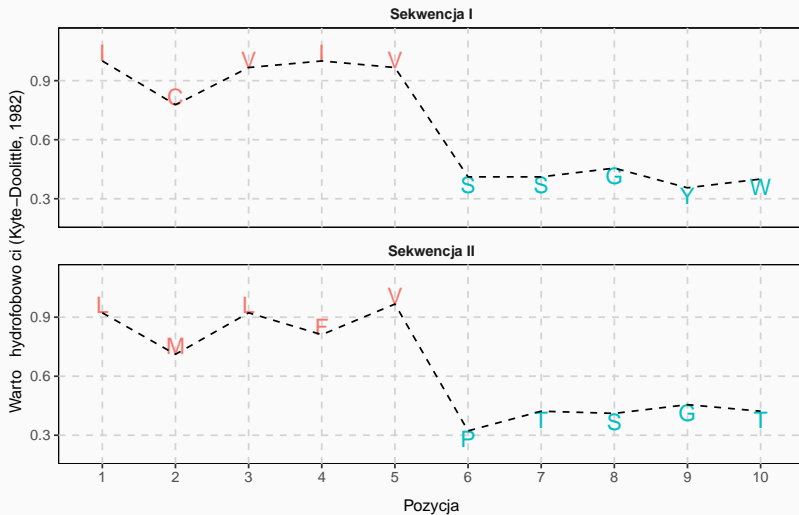
Sekwencja I:

ICVIVSSGYW

Sekwencja II:

LMLFVPTSGT

Uprozczone alfabety



Uproszczone alfabety

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

Sekwencja I:

ICVIVSSGYW

Po uproszczeniu alfabetu:

1111122222

Sekwencja II:

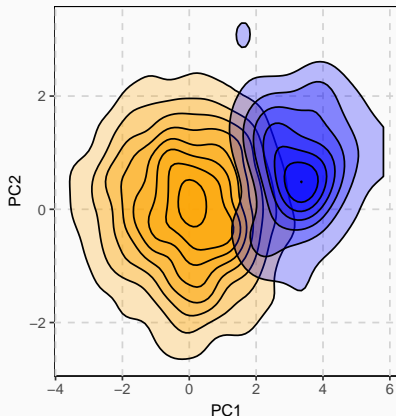
LMLFVPTSGT

Po uproszczeniu alfabetu:

1111122222

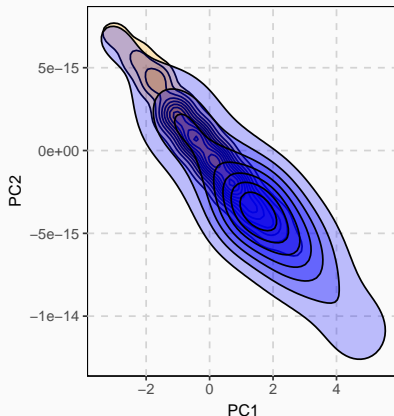
Uprozczone alfabety

Pełny alfabet



■ Inne eukarioty ■ Plasmodiidae

Zredukowany alfabet



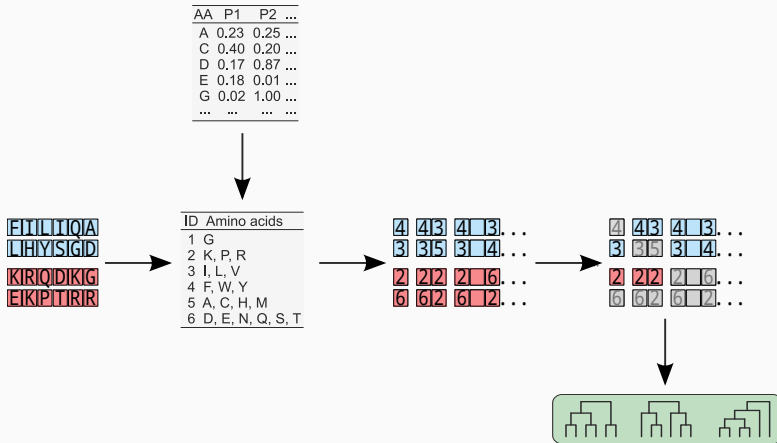
■ Inne eukarioty ■ Plasmodiidae

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

Wyniki

Porównanie z innymi klasyfikatorami

AmyloGram: oparte o redukcję alfabetów i kodowanie n-gramowe narzędzie do predykcji białek amyloidogennych.



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

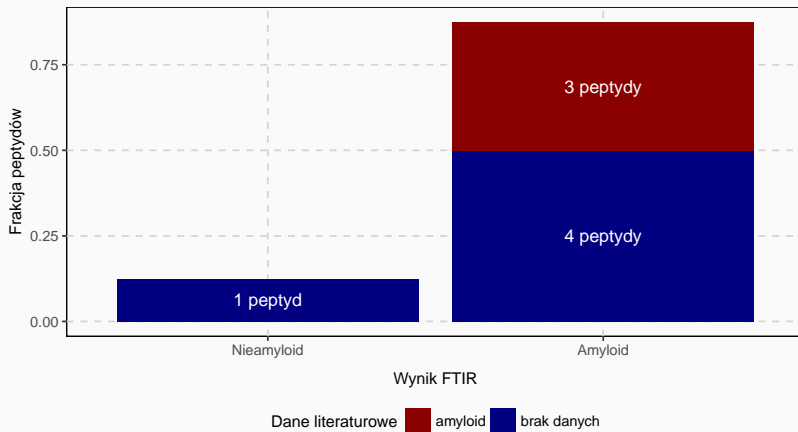
AmyloGram porównano z innymi klasyfikatorami na zewnętrznym zbiorze danych *pep424*.

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).

Walidacja eksperymentalna

Przy pomocy spektroskopii fourierowskiej zweryfikowano **8 peptydów**, które w bazie AmyLoad są oznaczone jako nieamyloidowe, a przez AmyloGram zostały rozpoznane jako amyloidy.



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 efektywny program do przewidywania białek amyloidogennych AmyloGram.

1. Zastosowanie opracowanej metodologii do przewidywania lokalizacji subkomórkowej białek. Porównanie wyników przewidywań z predyktorami opartymi na pełnych alfabetach aminokwasowych i innych informacjach o sekwencjach białek.
2. Zastosowanie opracowanej metodologii do przewidywania lokalizacji subkomórkowej białek. Porównanie wyników przewidywań z predyktorami opartymi na pełnych alfabetach aminokwasowych i innych informacjach o sekwencjach białek.
3. Upublicznienie rozwijanych metod w postaci pakietu *biogram* w środowisku programistycznym i statystycznym **R**.

Literatura

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.

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.

References II

- 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.

n-grams

n-grams (k-tuples) are vectors of n characters derived from input sequence(s).

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

1-grams: M, M, M, M, R, G, A, E

n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

2-grams: MR, MG, MA, ME, RK, GL, AF, ER

n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

3-grams: MRK, MGL, MAF, MER, RKL, GLF, AFG, ERG

n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

2-grams (with a single gap): M-K, M-L, M-F, M-R, R-L,
G-F, A-G, E-G

n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

3-grams (with gaps): M - K - - C, M - L - - I, M - F - - L,
M - R - - G, R - L - - V, G - F - - S, A - G - - L, E - G - - A