

ALFABET ŻYCIA: N-GRAMOWA ANALIZA BIAŁEK

MICHAŁ BURDUKIEWICZ

MI² DATA LAB, POLITECHNIKA WARSZAWSKA

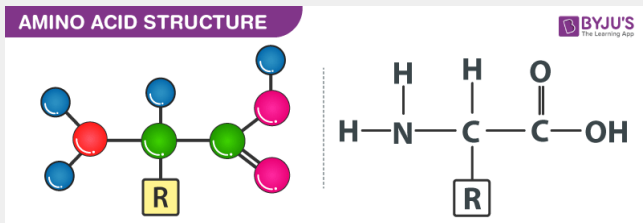
bioinformatyka [gr.-łac.], interdyscyplinarna dziedzina nauki zajmująca się sposobami gromadzenia, przekazywania i przetwarzania informacji w układach biol., gł. na poziomie mezoskopowym (na poziomie makromolekuł);
Źródło: *sjp.pwn.pl*.

Zastosowanie metod uczenia maszynowego do przewidywania właściwości białek.

- 1 Aminokwasy i białka
- 2 n-gramy i uproszczone alfabety
- 3 Przewidywanie amyloidów
- 4 Badania eksperymentalne

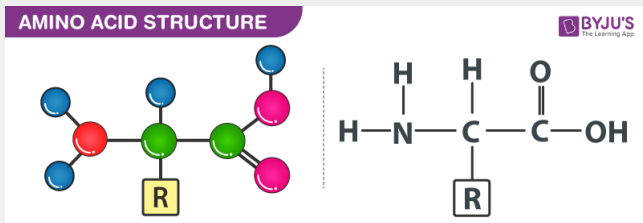
AMINOKWASY I BIAŁKA

AMINOKWASY



Źródło: byjus.com





















AMINOKWASY



Źródło: byjus.com

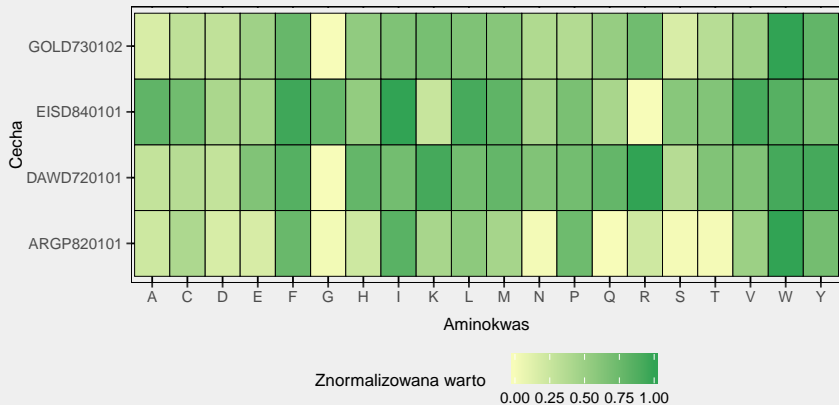
AMINOKWASY

Chart Key: ● ALIPHATIC ● AROMATIC ● ACIDIC ● BASIC ● HYDROXYLIC ● SULFUR-CONTAINING ● AMIDIC ○ NON-ESSENTIAL ○ ESSENTIAL

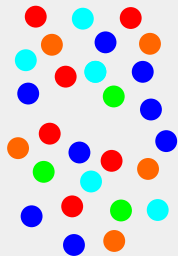
<p><i>Chemical Structure</i> single letter code</p> <p>NAME (A) three letter code DNA codons</p>	<p></p> <p>ALANINE (A) <i>Ala</i> GCT, GCC, GCA, GCG</p>	<p></p> <p>GLYCINE (G) <i>Gly</i> GGT, GGC, GGA, GGG</p>	<p></p> <p>ISOLEUCINE (I) <i>Ile</i> ATT, ATC, ATA</p>	<p></p> <p>LEUCINE (L) <i>Leu</i> CTT, CTC, CTA, CTG, TTA, TTG</p>	<p></p> <p>PROLINE (P) <i>Pro</i> CCT, CCC, CCA, CCG</p>	<p></p> <p>VALINE (V) <i>Val</i> GTT, GTC, GTA, GTG</p>
<p></p> <p>PHENYLALANINE (F) <i>Phe</i> TTT, TTC</p>	<p></p> <p>TRYPTOPHAN (W) <i>Trp</i> TGG</p>	<p></p> <p>TYROSINE (Y) <i>Tyr</i> TAT, TAC</p>	<p></p> <p>ASPARTIC ACID (D) <i>Asp</i> GAT, GAC</p>	<p></p> <p>GLUTAMIC ACID (E) <i>Glu</i> GAA, GAG</p>	<p></p> <p>ARGININE (R) <i>Arg</i> CGT, CGC, CGA, CGG, AGA, AGG</p>	<p></p> <p>HISTIDINE (H) <i>His</i> CAT, CAC</p>
<p></p> <p>LYSINE (K) <i>Lys</i> AAA, AAG</p>	<p></p> <p>SERINE (S) <i>Ser</i> TCT, TCC, TGA, TGG, AGT, AGC</p>	<p></p> <p>THREONINE (T) <i>Thr</i> ACT, ACC, ACA, ACG</p>	<p></p> <p>CYSTEINE (C) <i>Cys</i> TGT, TGC</p>	<p></p> <p>METHIONINE (M) <i>Met</i> ATG</p>	<p></p> <p>ASPARAGINE (N) <i>Asn</i> AAT, AAC</p>	<p></p> <p>GLUTAMINE (Q) <i>Gln</i> CAA, CAG</p>

Źródło: microbenotes.com

AMINOKWASY



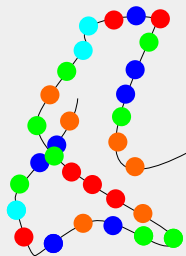
BIAŁKA



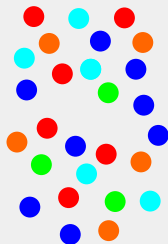
Aminokwasy



Białka (struktura pierwszorzędowa)



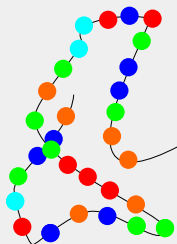
Białka (struktura wyższych rzędów)



Aminokwasy



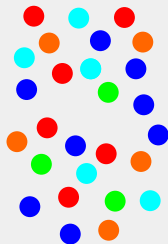
Białka (struktura
pierwszorzędowa)



Białka (struktura
wyższych rzędów)

POZNANE

NIEPOZNANE

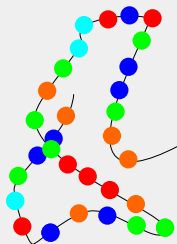


Aminokwasy



Białka (struktura
pierwszorzędowa)

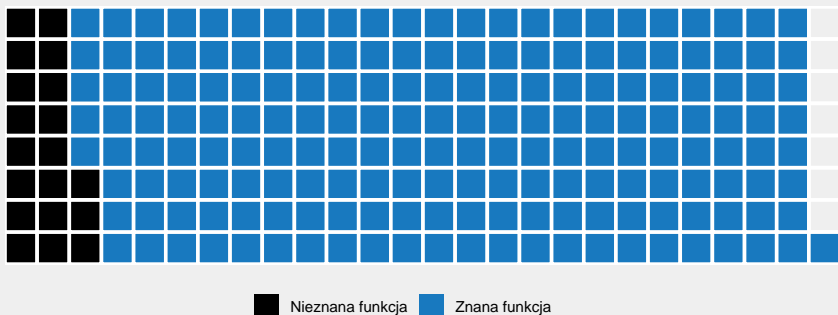
POZNANE



Białka (struktura
wyższych rzędów)

NIEPOZNANE

Struktura wyższych rzędów określa funkcję białka.



1937 ludzkich białek ma nieznaną funkcję (dark proteome)
(Young-Ki Paik et al., 2018).

Zastosowanie metod uczenia maszynowego do przewidywania właściwości białek **na podstawie ich struktury pierwszorzędowej.**

N-GRAMY I UPROSZCZONE ALFABETY

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

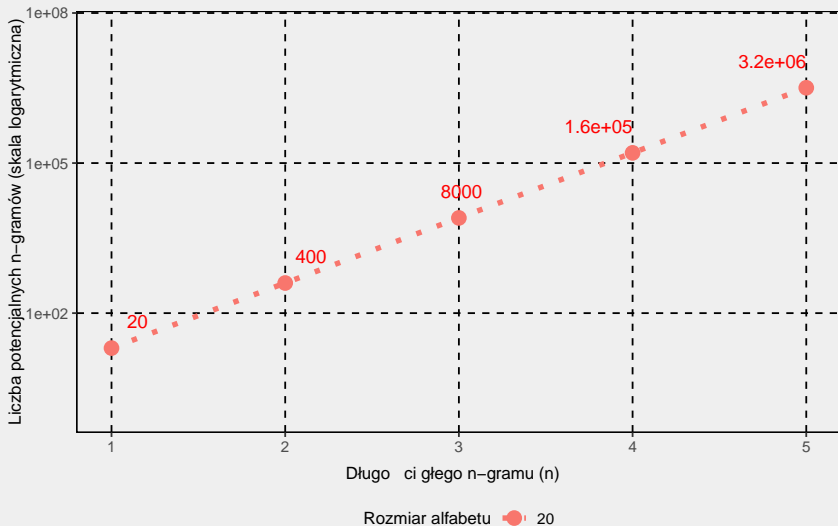
- podsekwencje (ciągłe lub z przerwami) n reszt aminokwasowych lub nukleotydowych,
- bardziej informatywne niż pojedyncze reszty.

Peptyd I: FKVWPDHGSG

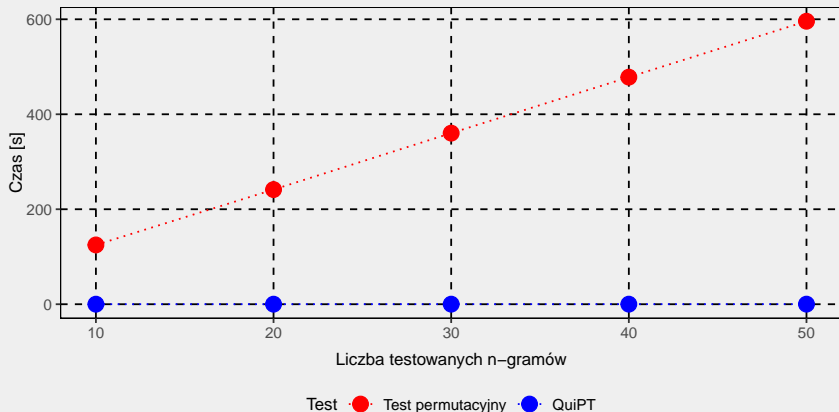
Peptyd II: YMCIIYRAQTN

Przykłady n-gramów uzyskanych z peptydów I i II:

1. 1-gramy: F, Y, K, M,
2. 2-gramy: FK, YM, KV, MC,
3. 2-gramy (nieciągłe): F-V, Y-C, K-W, M-I,
4. 3-gramy (nieciągłe): F-WP, Y-IY, K-PD, M-YR.



Dłuższe n-gramy są bardziej informatywne, ale tworzą większe przestrzenie atrybutów, które są trudniejsze do analizy.



QuiPT (dostępny jako funkcja w pakiecie **biogram**) jest szybszy niż klasyczne testy permutacyjne.

Uproszczone alfabety:

- aminokwasy są grupowane w większe zbiory na podstawie określonych kryteriów,
- łatwiejsze przewidywanie struktur (Murphy et al., 2000),
- tworzenie bardziej uogólnionych modeli.

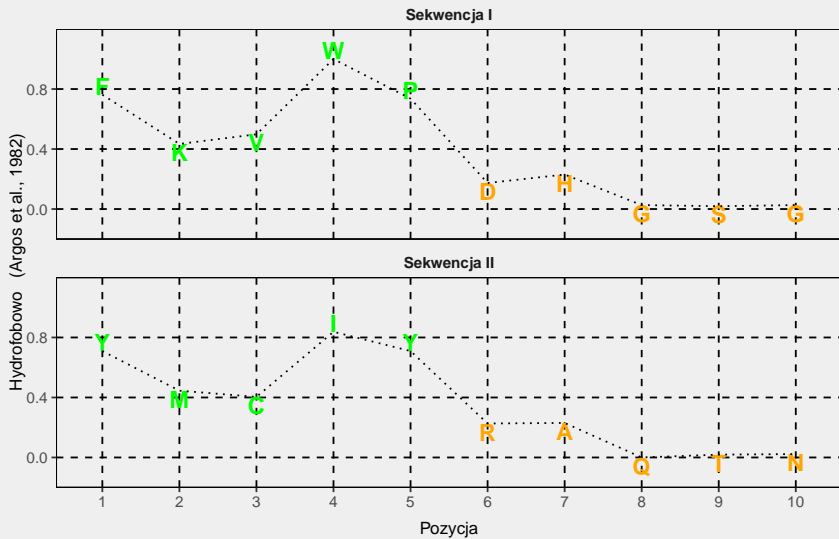
Poniższe peptydy wydają się być całkowicie różne pod względem składu aminokwasowego.

Peptyd I:

FKVWPDHGSG

Peptyd II:

YMCIIYRAQTN



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

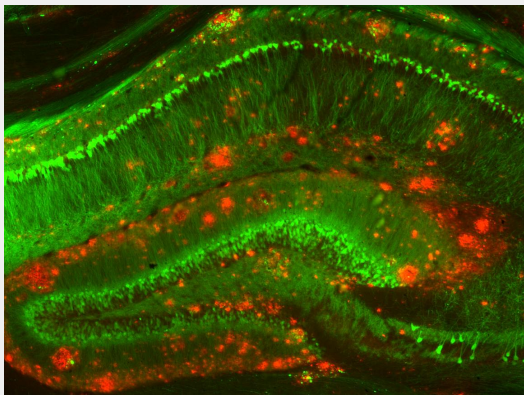
Peptyd I: FKVWPDHGSG → 1111122222
 Peptyd II: YMCIRYAQTN → 1111122222

Zastosowanie metod uczenia maszynowego do przewidywania właściwości białek na podstawie ich struktury pierwszorzędowej **zakodowanej w postaci n-gramów zapisanych w uproszczonym alfabecie.**

PRZEWIDYWANIE AMYLOIDÓW

BIAŁKA AMYLOIDOWE

Agregaty białek amyloidowe występują w tkankach osób cierpiących na zaburzenia neurodegeneracyjne, takie jak choroba Alzheimera i Parkinsona, a także wiele innych schorzeń.

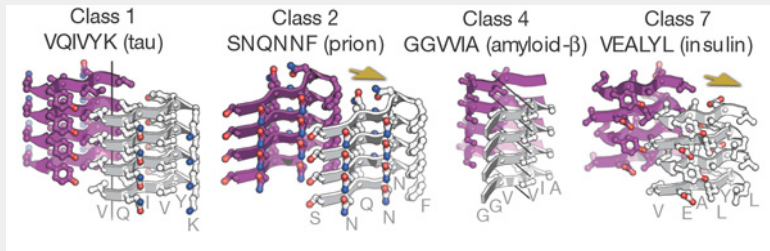


Agregaty amyloidowe (czerwone) wokół neuronów (zielone). Strittmatter Laboratory, Yale University.

BIAŁKA AMYLOIDOWE

Za agregację białek amyloidogennych odpowiedzialne są sekwencje peptydowe o właściwościach amyloidogennych (hot spots):

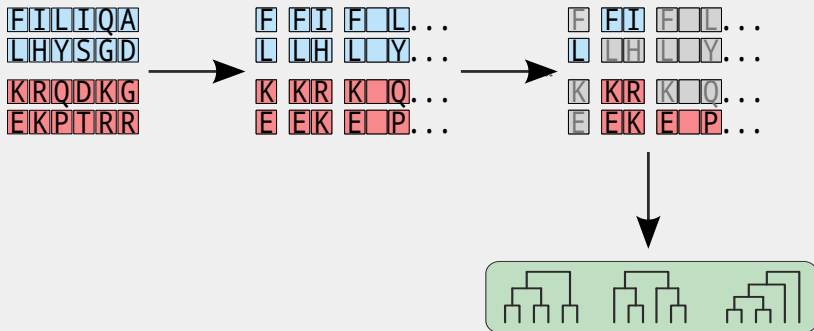
- krótkie (6-15 aminokwasów),
- bardzo zmienny, zazwyczaj hydrofobowy skład aminokwasowy,
- tworzą unikalne β -struktury.



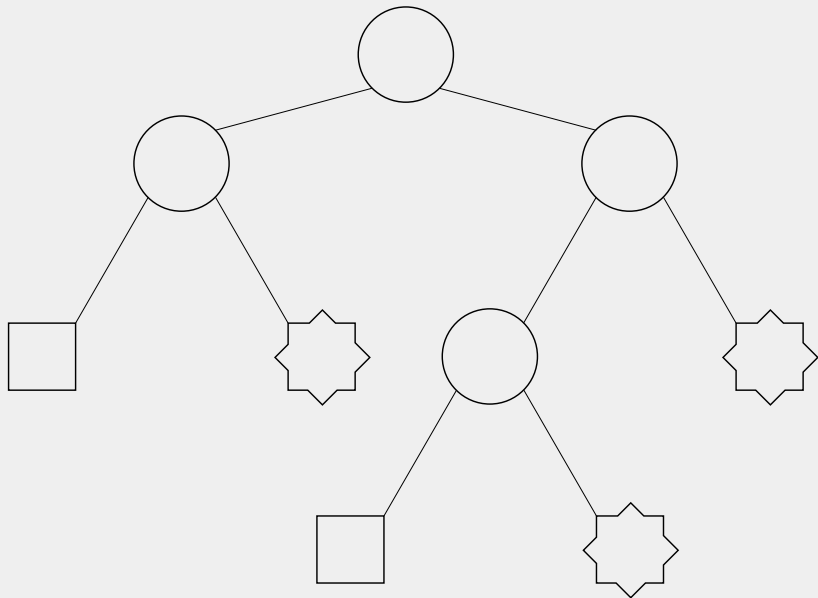
Sawaya et al. (2007)

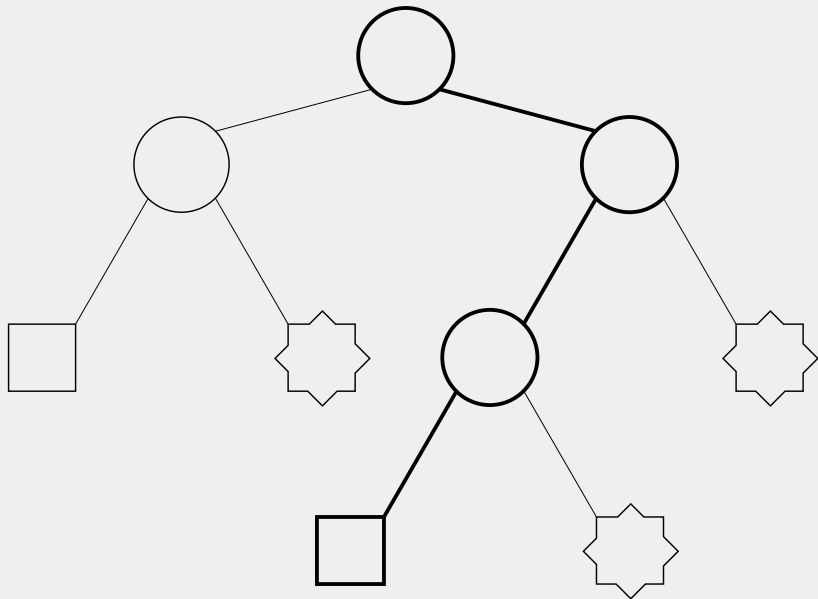
Zastosowanie metod uczenia maszynowego do przewidywania właściwości **amyloidogenności** białek na podstawie ich struktury pierwszorzędowej zakodowanej w postaci n-gramów zapisanych w uproszczonym alfabecie.

AmyloGram: oparte na analizie n-gramowej narzędzie do przewidywania amyloidów (Burdukiewicz et al., 2016, 2017).

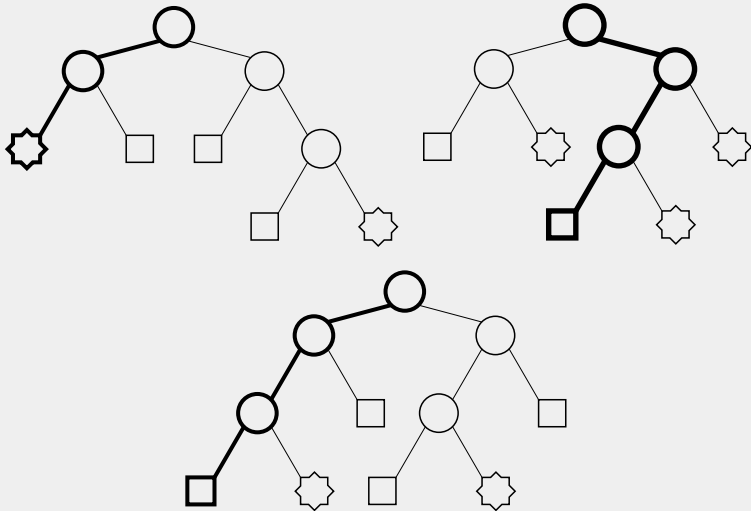


Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961





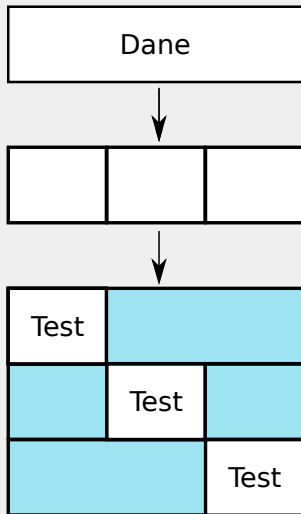
LASY LOSOWE



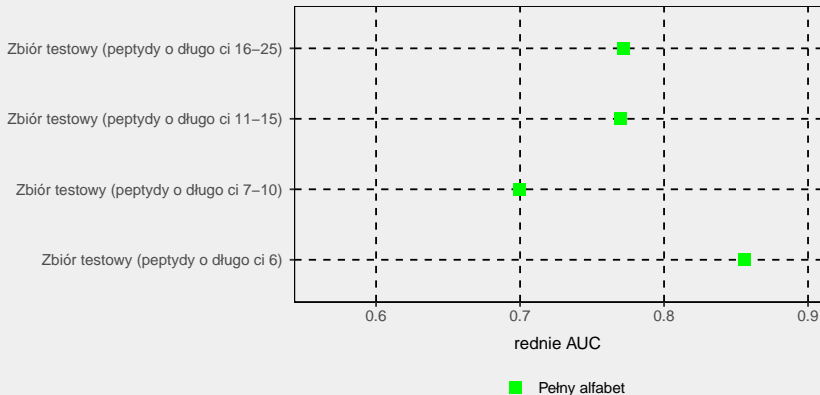
Przewidywania: ☐ ☐ ☒

Ostateczna decyzja: ☐

WALIDACJA KRZYŻOWA



Zbiór treningowy (peptydy o długości 6)



Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

Czy standardowe uproszczone alfabety opracowane dla różnych zagadnień biologicznych pomagają lepiej przewidywać amyloidy?

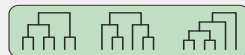
STANDARDOWE UPROSZCZONE ALFABETY

F I I L I I Q A
 L I H Y S G D
 K R Q D K G
 E K P T R R

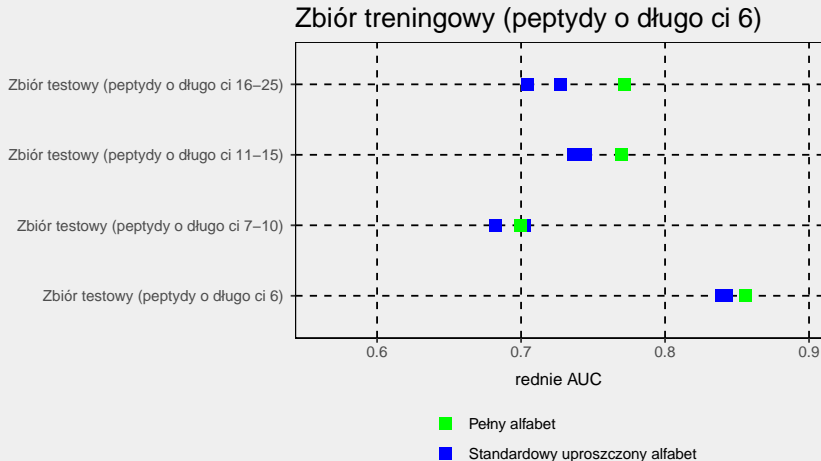
ID	Amino acids
1	G
2	K, P, R
3	I, L, V
4	F, W, Y
5	A, C, H, M
6	D, E, N, Q, S, T

4 4 3 4 3...
 3 3 5 3 4...
 2 2 2 2 6...
 6 6 2 6 2...

4 4 3 4 3...
 3 3 5 3 4...
 2 2 2 2 6...
 6 6 2 6 2...



Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961



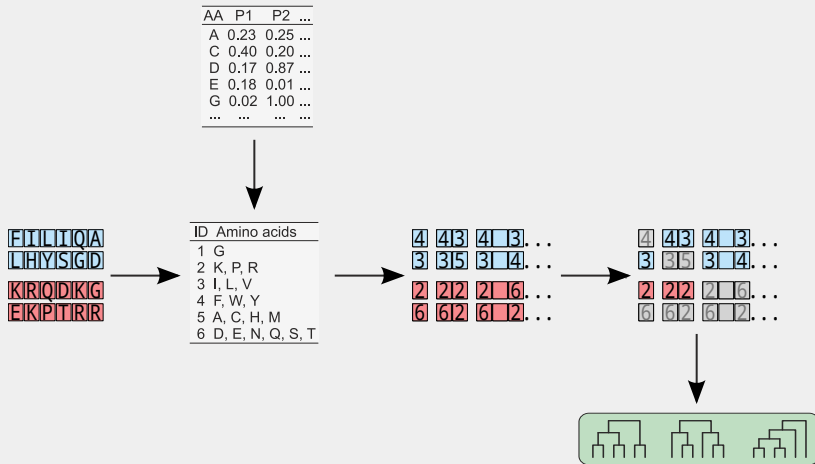
Standardowe alfabety aminokwasowe nie poprawiają jakości predykcji amyloidów.

Burdukiewicz, M., Sobczyk, P., Rödigier, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

NOWE UPROSZCZONE ALFABETY

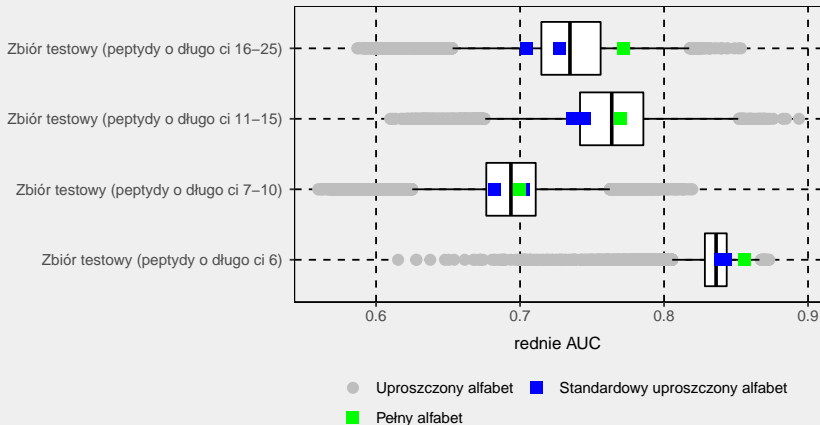
- 17 miar fizykochemicznych wybranych z bazy AAIndex:
 - ▶ rozmiar,
 - ▶ hydrofobowość,
 - ▶ częstość w β -karkach,
 - ▶ zdolność do tworzenia kontaktów.
- 524 284 uproszczonych alfabetów aminokwasowych o różnej wielkości (od 3 do 6 grup).

NOWE UPROSZCZONE ALFABETY



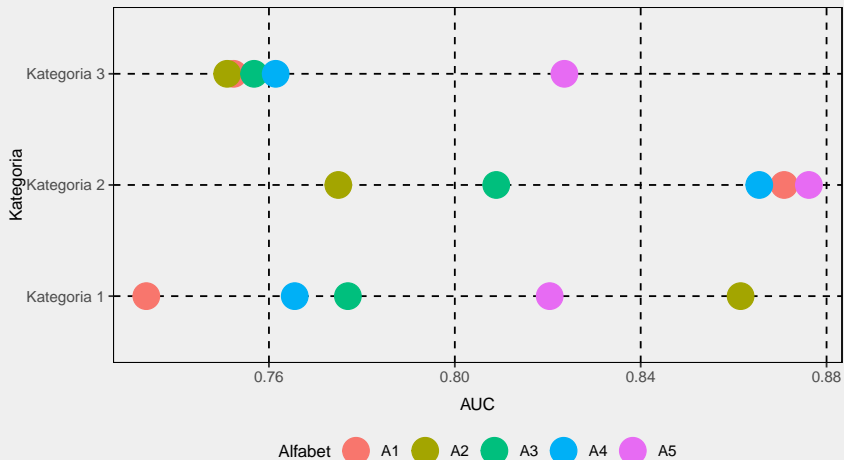
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

Zbiór treningowy (peptydy o długości 6)

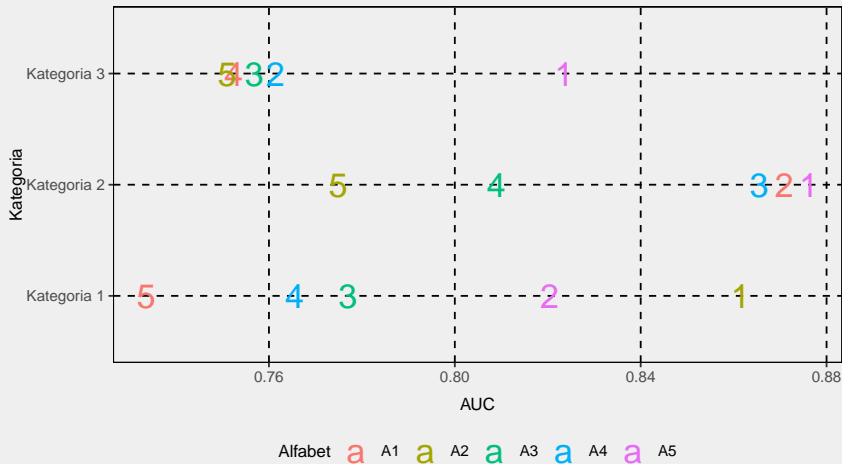


Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

WYBÓR NAJLEPSZEGO ALFABETU

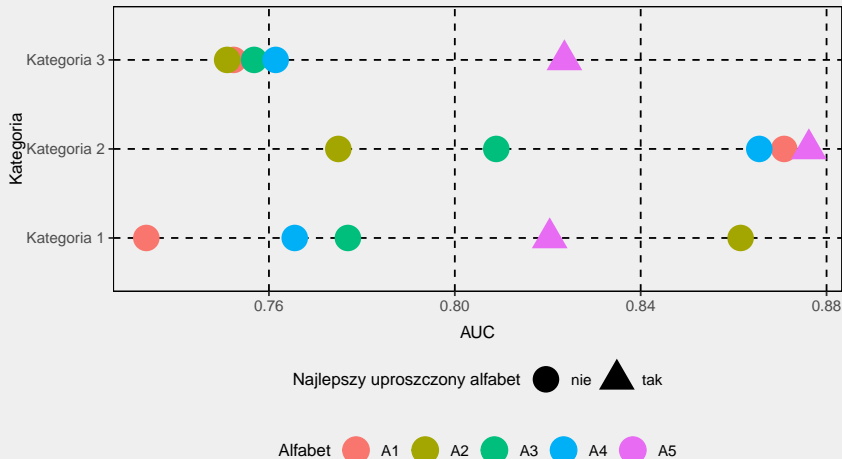


WYBÓR NAJLEPSZEGO ALFABETU



Dla każdej kategorii alfabetu zostały porangowane (ranga 1 dla najlepszego AUC itd.).

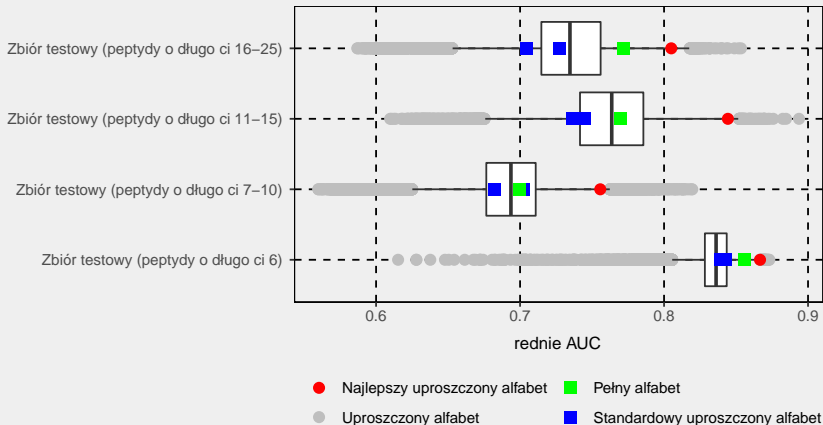
WYBÓR NAJLEPSZEGO ALFABETU



Za najlepszy alfabet uznano alfabet z najmniejszą sumą rang.

NAJLEPSZY UPROSZCZONY ALFABET

Zbiór treningowy (peptydy o długości 6)



Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

NAJLEPSZY UPROSZCZONY ALFABET

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

Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

NAJLEPSZY UPROSZCZONY ALFABET

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

Grupy 3 i 4 - aminokwasy hydrofobowe.

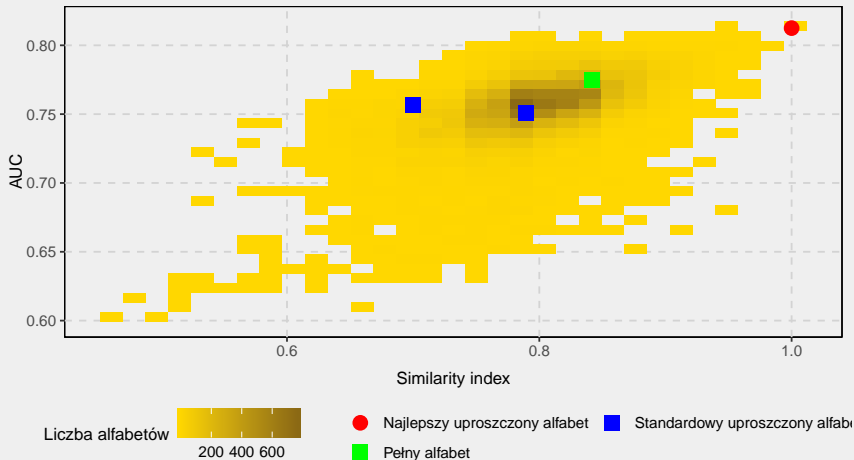
NAJLEPSZY UPROSZCZONY ALFABET

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

Grupa 2 - reszty aminokwasowe zakłócające β -struktury.

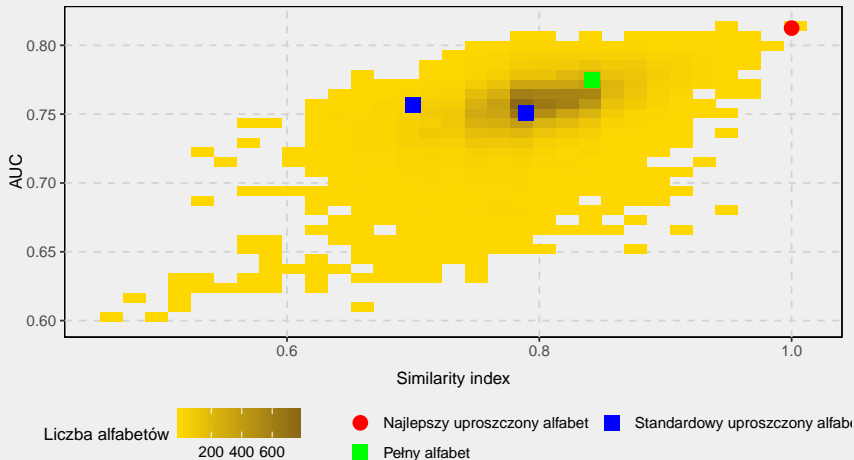
Czy alfabety podobne do najlepszego uproszczonego alfabetu również wspierają przewidywania amyloidów?

SIMILARITY INDEX



Similarity index (Stephenson and Freeland, 2013) mierzy podobieństwo między dwoma uproszczonymi alfabetami (1: identyczne alfabety, 0: zupełnie niepodobne alfabety).

SIMILARITY INDEX



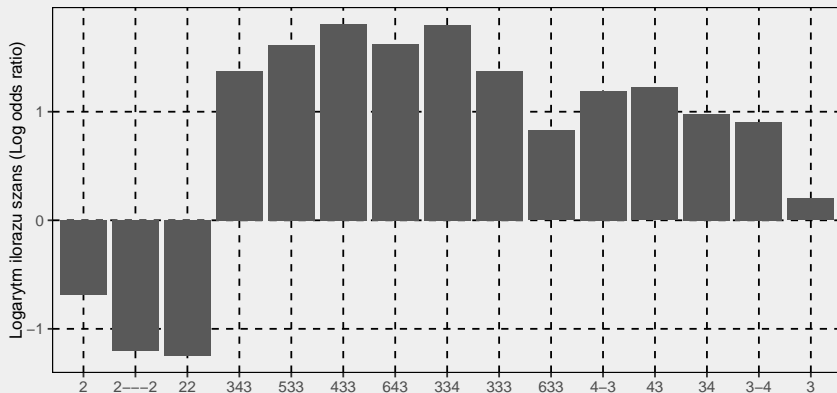
Korelacja między similarity index i średnim AUC jest istotna

($p\text{-value} \leq 2.2^{-16}$; $\rho = 0.51$).

Burdukiewicz, M., Sobczyk, P., Rödigier, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

Czy informatywne n-gramy znalezione przez QuiPT są związane z amyloidogennością?

INFORMATYWNE N-GRAMY



Spośród 65 najbardziej informatywnych n-gramów, 15 (23%) jest również obecnych w motywach aminokwasowych znalezionych ekperymentalnie (Paz and Serrano, 2004).

Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

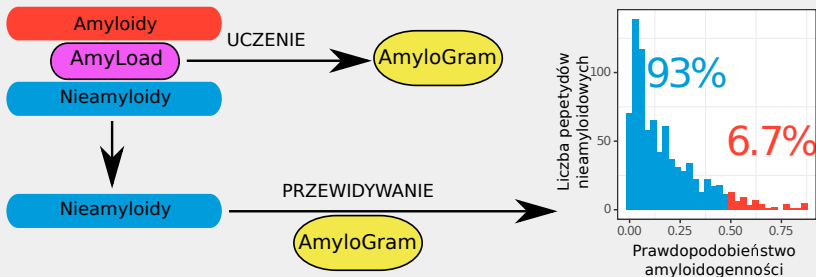
PORÓWNANIE Z INNYMI NARZĘDZIAMI

Program	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

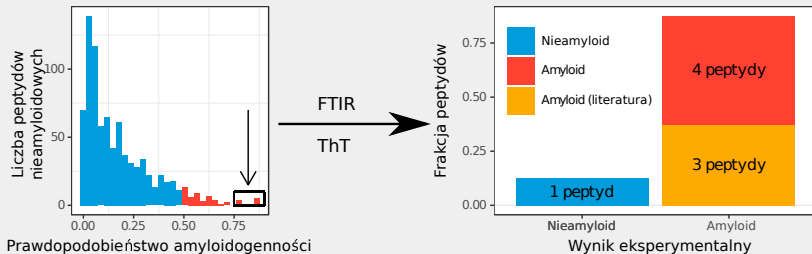
Klasyfikator wytrenowany z wykorzystaniem najlepszego uproszczonego alfabetu, AmyloGram, został porównany z innymi narzędziami do przewidywania amyloidów z użyciem zewnętrznego zbioru danych *pep424*.

BADANIA EKSPERYMENTALNE

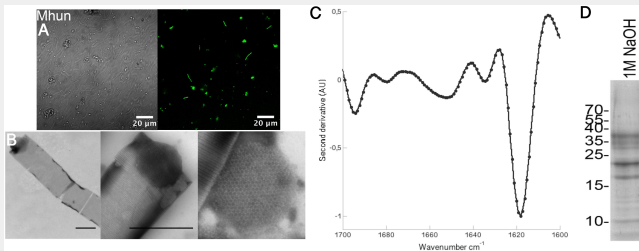
WALIDACJA EKSPERYMENTALNA



WALIDACJA EKSPERYMENTALNA



NOWE BIAŁKO AMYLOIDOWE



Nowy amyloid funkcjonalny produkowany przez *Methanospirillum* sp. (Christensen et al., 2018) został wybrany do analiz *in vitro* dzięki wskazaniom AmyloGramu.

Modele przewidujące właściwości białek mogą opierać się na regułach precyzyjnych zrozumiałych dla biologów i weryfikowalnych eksperymentalnie nie tracąc na swojej skuteczności.

Mentorzy:

- Paweł Mackiewicz (Uniwersytet Wrocławski).
- Małgorzata Kotulska (Politechnika Wrocławska).
- Stefan Rödiger (Brandenburg University of Technology Cottbus-Senftenberg).
- Henrik Nielsen (Technical University of Denmark).
- Lars Kaderali (University of Greifswald).
- Jarosław Chilimoniuk (Uniwersytet Wrocławski).
- Piotr Sobczyk (Politechnika Wrocławska).

Finansowanie:

- Narodowe Centrum Nauki (2015/17/N/NZ2/01845 i 2017/24/T/NZ2/00003).
- COST ACTION CA15110 (Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research).
- KNOW Wrocław Center for Biotechnology.
- Ministerstwo Edukacji i Badań Naukowych Niemiec (InnoProfile-Transfer-Projekt 03IPT611X).

MI² Data Lab (<https://mi2.mini.pw.edu.pl/>), Wydział Matematyki i Nauk Informatycznych, Politechnika Warszawska.



Kontakt: michalburdukiewicz@gmail.com.
Prezentacja: <https://github.com/michbur/dpm>.

REFERENCES I

- Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports, 7(1):12961.
- 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.
- Christensen, L. F. B., Hansen, L. M., Finster, K., Christiansen, G., Nielsen, P. H., Otzen, D. E., and Dueholm, M. S. (2018). The sheaths of methanospirillum are made of a new type of amyloid protein. Frontiers in Microbiology, 9:2729.
- 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.
- Paz, M. L. d. I. and Serrano, L. (2004). Sequence determinants of amyloid fibril formation. Proceedings of the National Academy of Sciences, 101(1):87–92.
- 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., Riekel, C., and Eisenberg, D. (2007). Atomic structures of amyloid cross- spines reveal varied steric zippers. Nature, 447(7143):453–457.

REFERENCES III

- Stephenson, J. D. and Freeland, S. J. (2013). Unearthing the root of amino acid similarity. Journal of Molecular Evolution, 77(4):159–169.
- 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.