Przewidywanie właściwości sekwencji biologicznych w oparciu o analizę n-gramów

Michał Burdukiewicz

Zakład Bioinformatyki i Genomiki, Uniwersytet Wrocławski

Promotor pracy: prof. dr hab. Paweł Mackiewicz
Promotor pomocniczy: dr Paweł Błażej

Plan prezentacji

n-gramy i uproszczone alfabety

Przewidywanie amyloidów

Przewidywanie peptydów sygnałowych

Przewidywanie warunków hodowlanych metanogenów

Badania *in silico* pozwalają efektywniej planować prace eksperymentalne.
Przykłady:

- przewidywanie właściwości białek (np. obecność sekwencji sygnałowych, amyloidogenność),
- przewidywanie warunków hodowlanych mikroorganizmów.

Cel

Opracowanie metodologii analizy sekwencji biologicznych opierającej się na zrozumiałych dla człowieka regułach decyzyjnych.

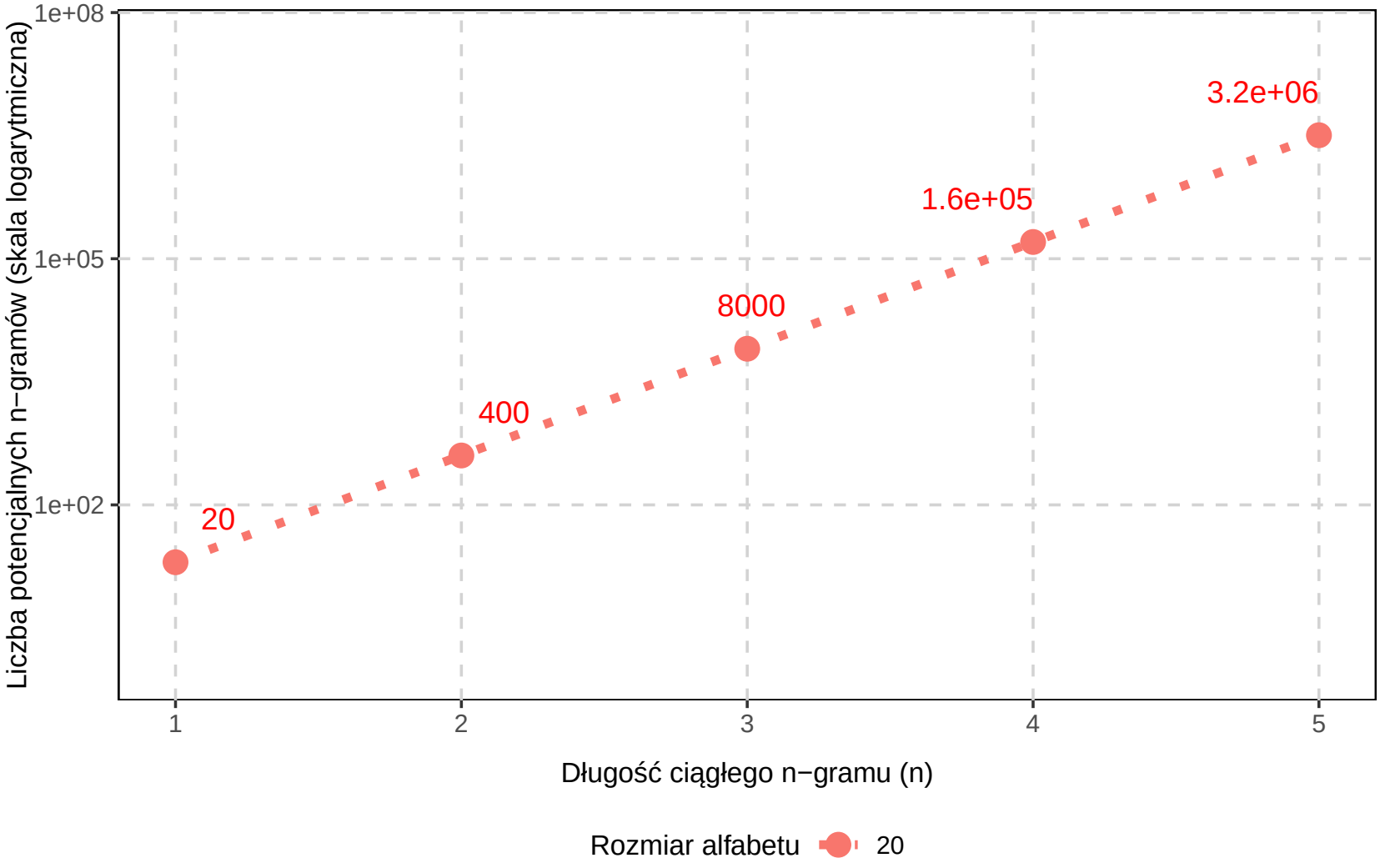
n-gramy (k-tuple, k-mery):

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

Peptyd I: FKVWPDHGSG
Peptyd II: YMCIYRAQTN

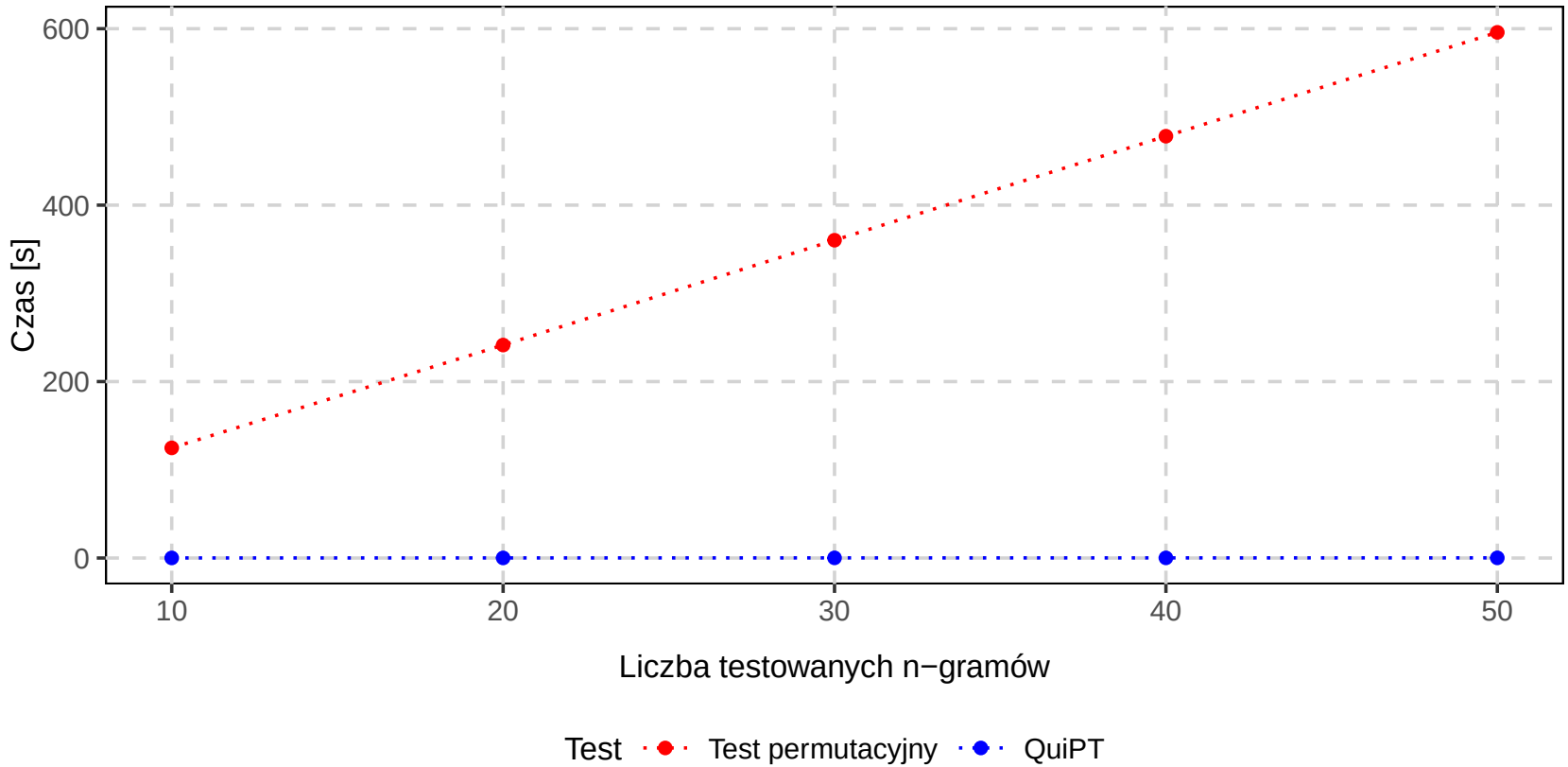
Przykłady n-gramów uzyskanych z peptydów I i II:

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



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

Uproszczone alfabety

Uproszczone alfabety:

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

Uproszczone alfabety

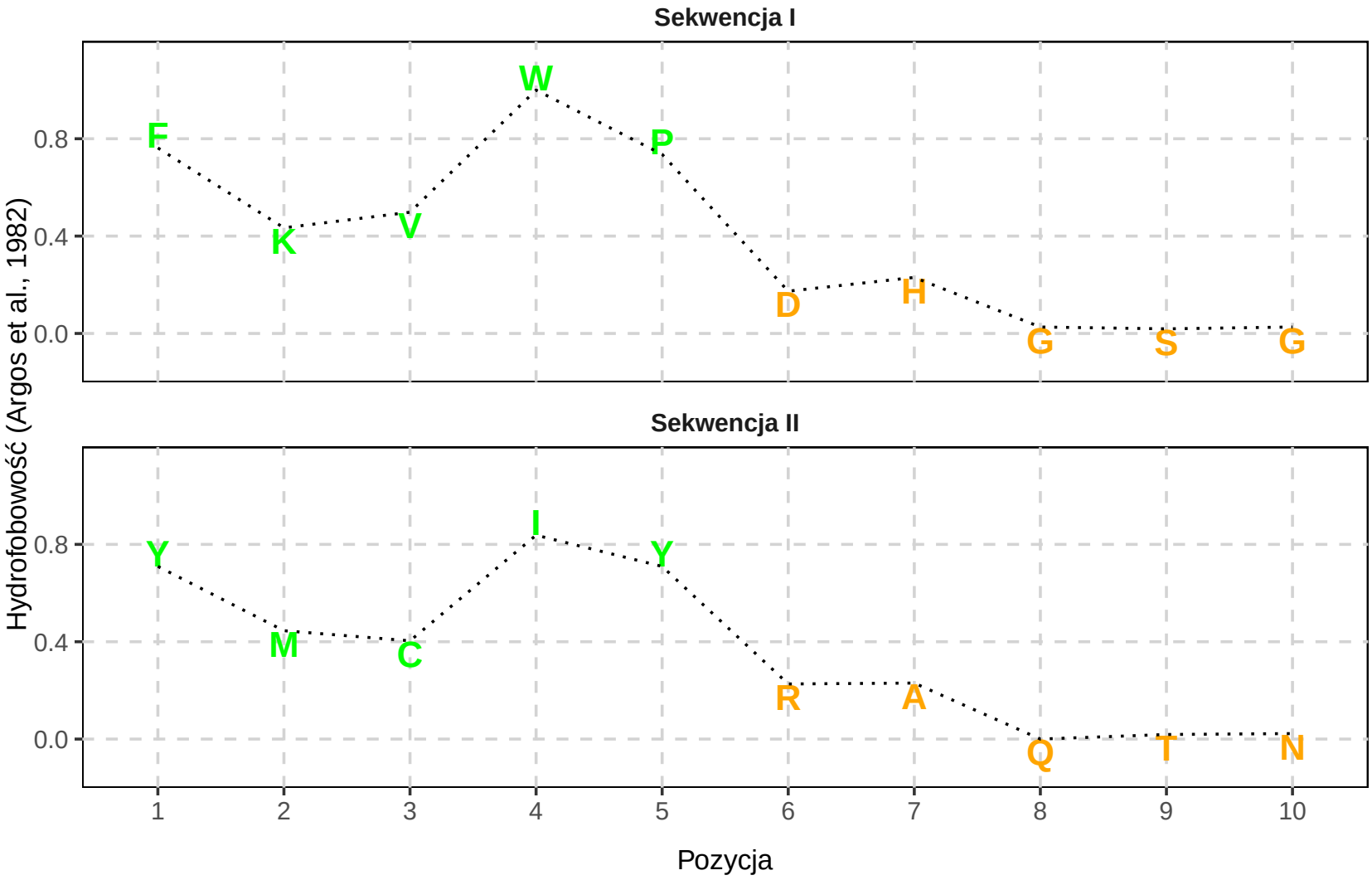
Poniższe peptydy wydają się być całkowicie różne pod względem składu aminokwasowego.

Peptyd I:

`FKVWPDHGSG`

Peptyd II:

`YMCIYRAQTN`


Sekwencja I
F
K
V
W
P
D
H
G
S
G
Sekwencja II
Y
M
C
I
Y
R
A
Q
T
N
Hydrofobowość (Argos et al., 1982)
0.0
0.4
0.8
1
2
3
4
5
6
7
8
9
10
Pozycja

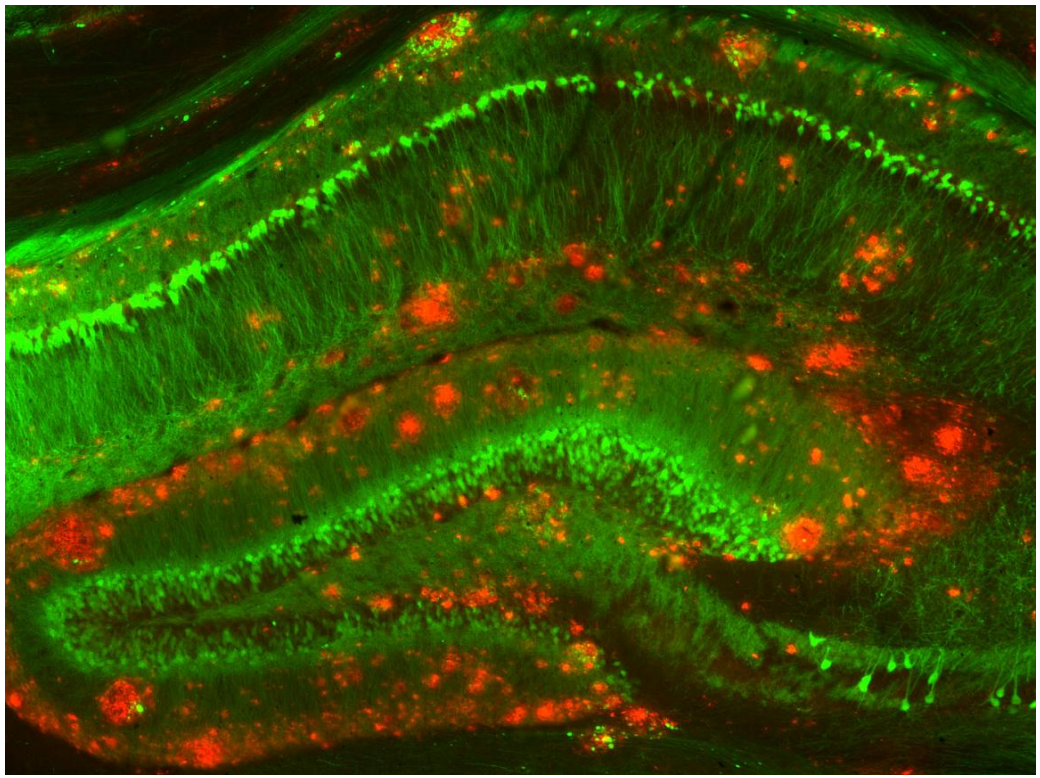
Grupa	Aminokwasy
1	C, I, L, K, M, F, P, W, Y, V
2	A, D, E, G, H, N, Q, R, S, T

Peptyd I: FKVWPDHGSG $\rightarrow$ 1111122222

Peptyd II: YMCIYRAQTN $\rightarrow$ 1111122222

Białka amyloidowe

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

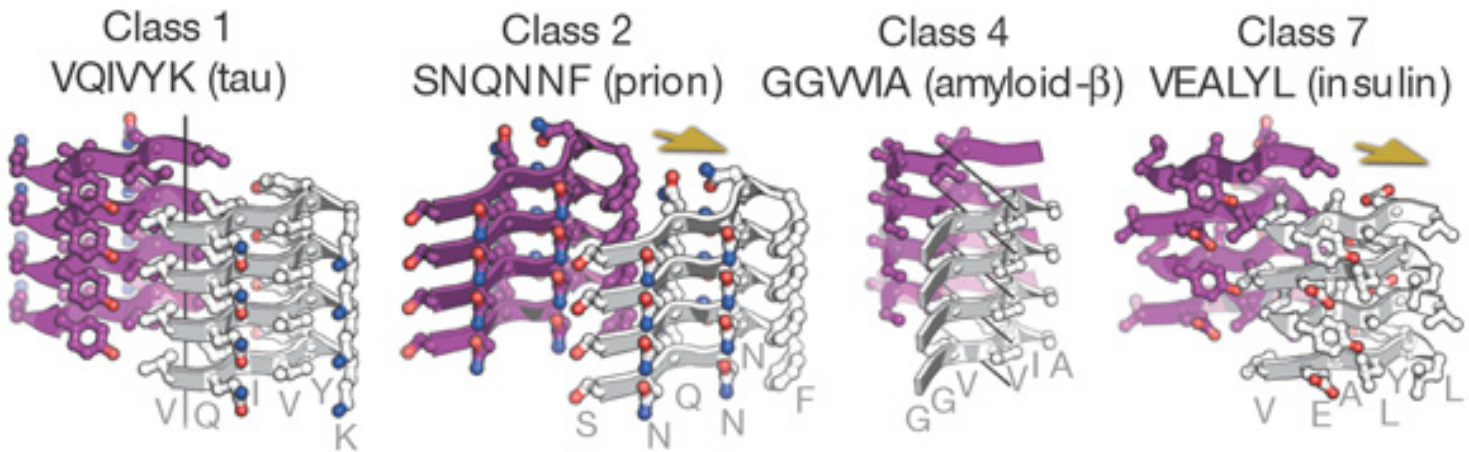


Agregaty amyloidowe (czerwone) wokół neuronów (zielone). Strittmatter Laboratory, Yale University.

Białka amyloidowe

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

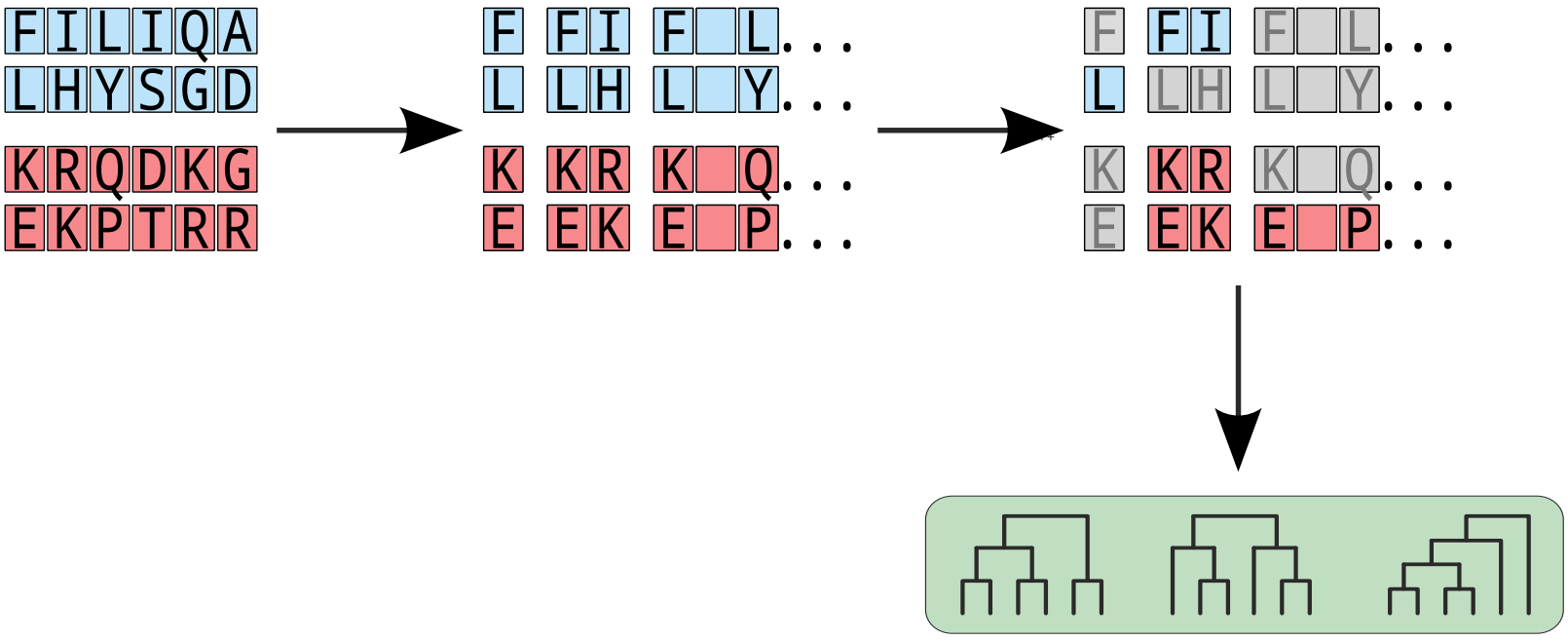
- krótkie (6-15 aminokwasów),
- bardzo zmienny, zazwyczaj hydrofobowy skład aminokwasowy,
- tworzą unikalne β-struktury.



Sawaya et al. (2007)

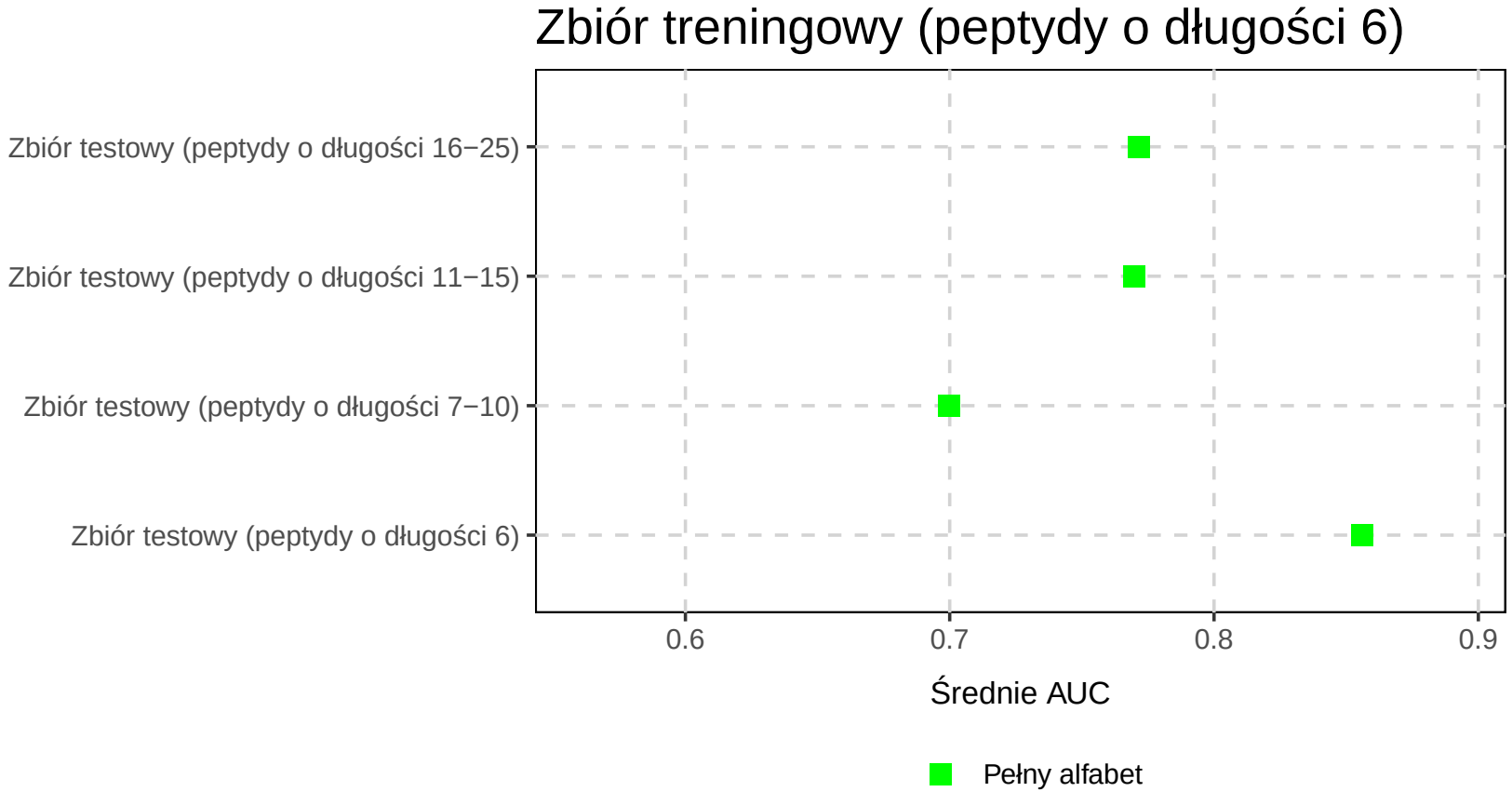
AmyloGram

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

Walidacja krzyżowa

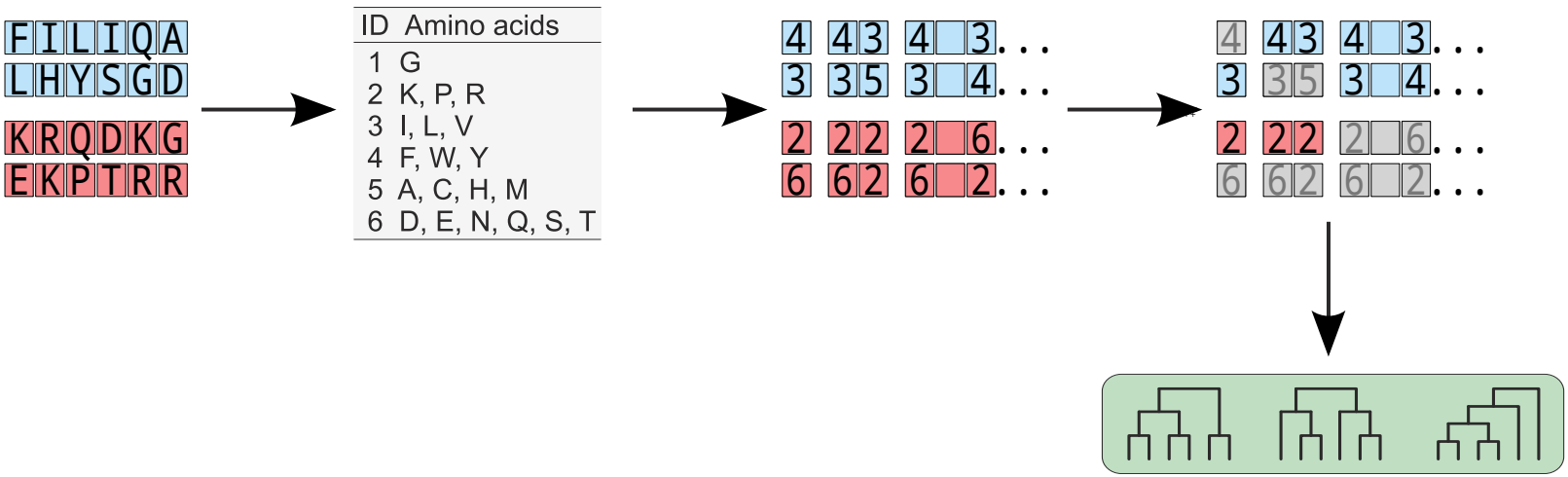


Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

Standardowe uproszczone alfabety

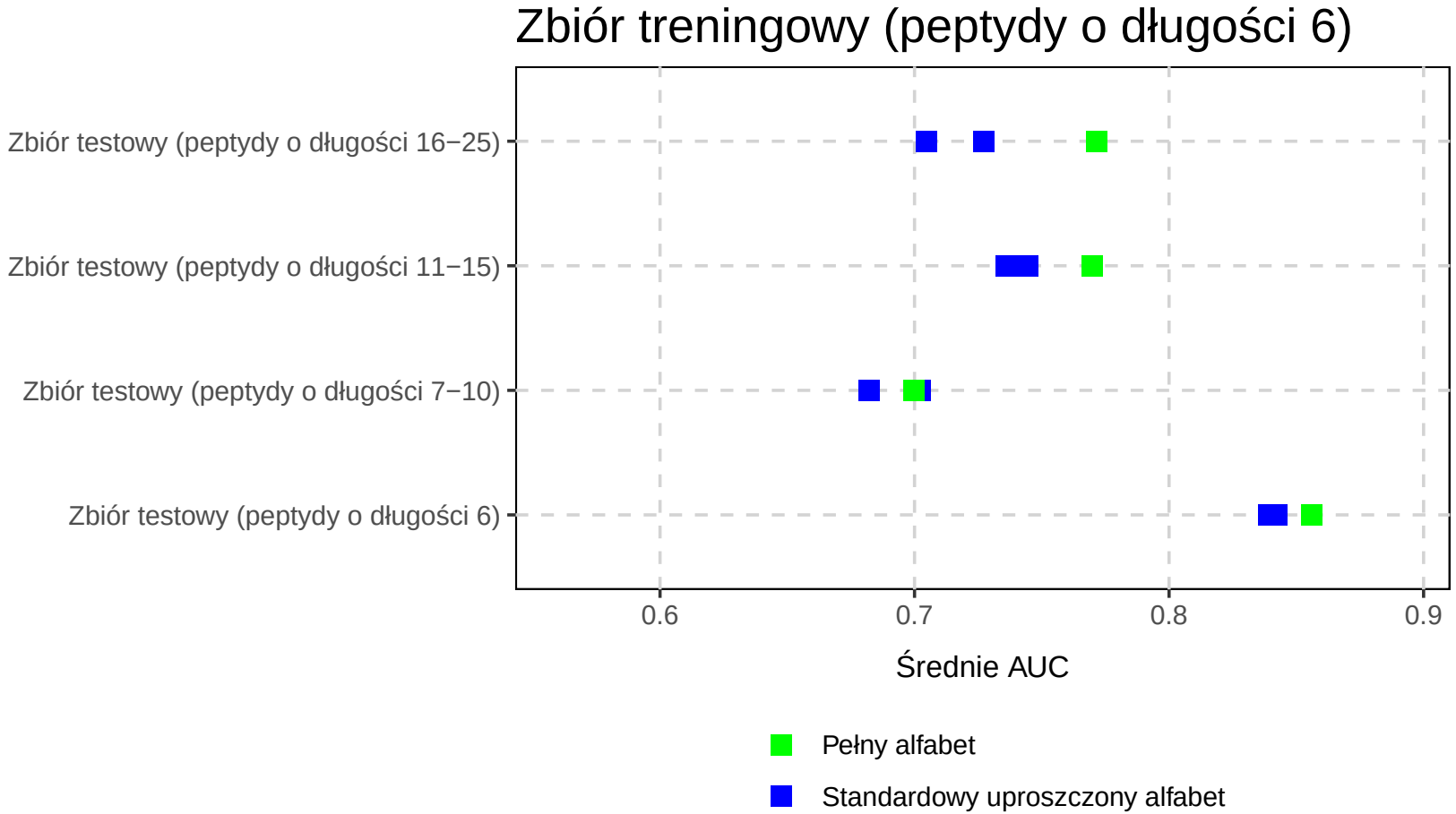
Opublikowano kilka uproszczonych alfabetów, które w założeniu miały służyć do opisywania struktur drugo- i trzeciorzędowych białek (Kosiol et al., 2004; Melo and Marti-Renom, 2006).

Standardowe 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

Walidacja krzyżowa



Standardowe alfabety aminokwasowe nie poprawiają jakości predykcji amyloidów.

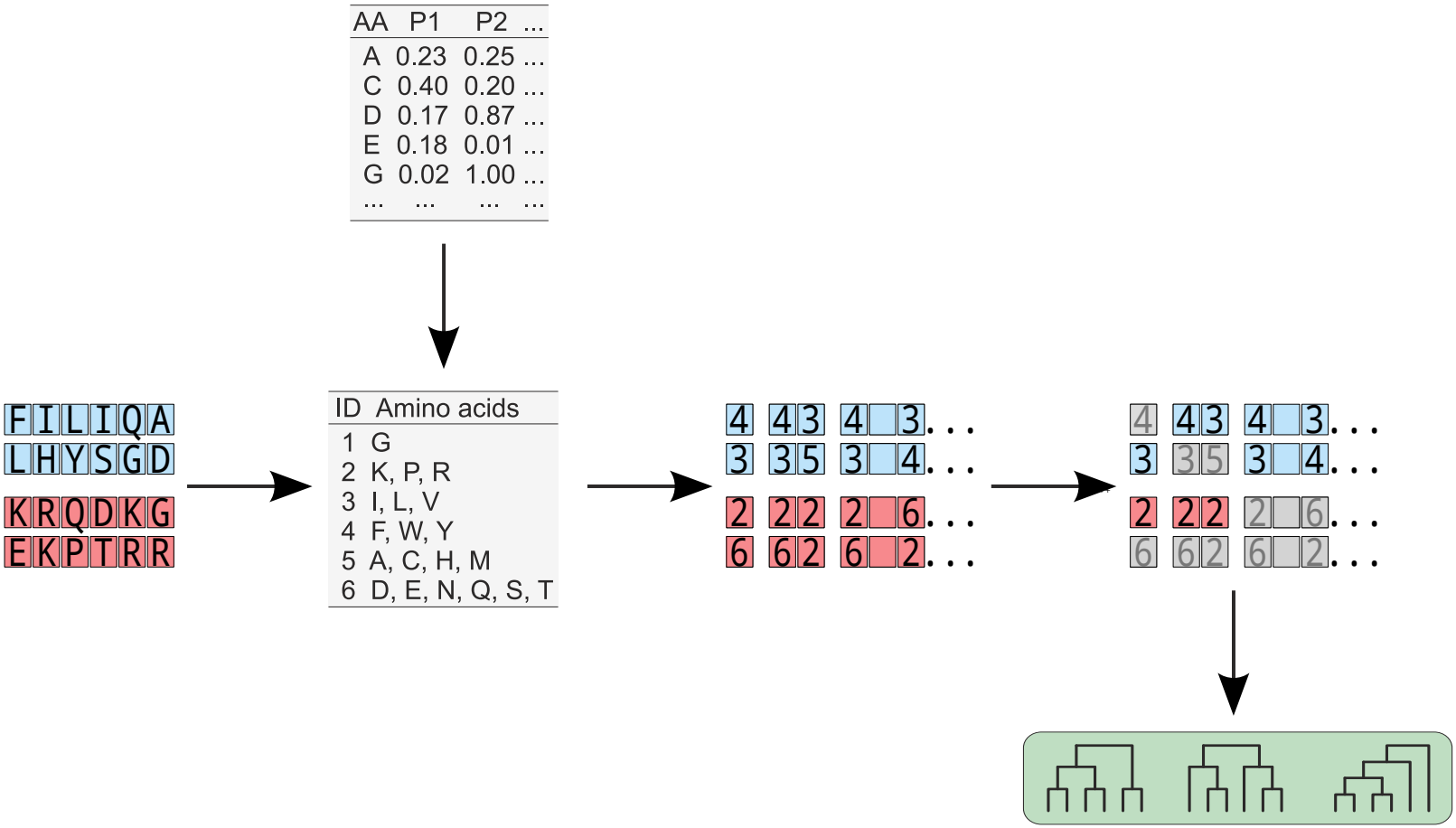
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

Nowe uproszczone alfabety

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

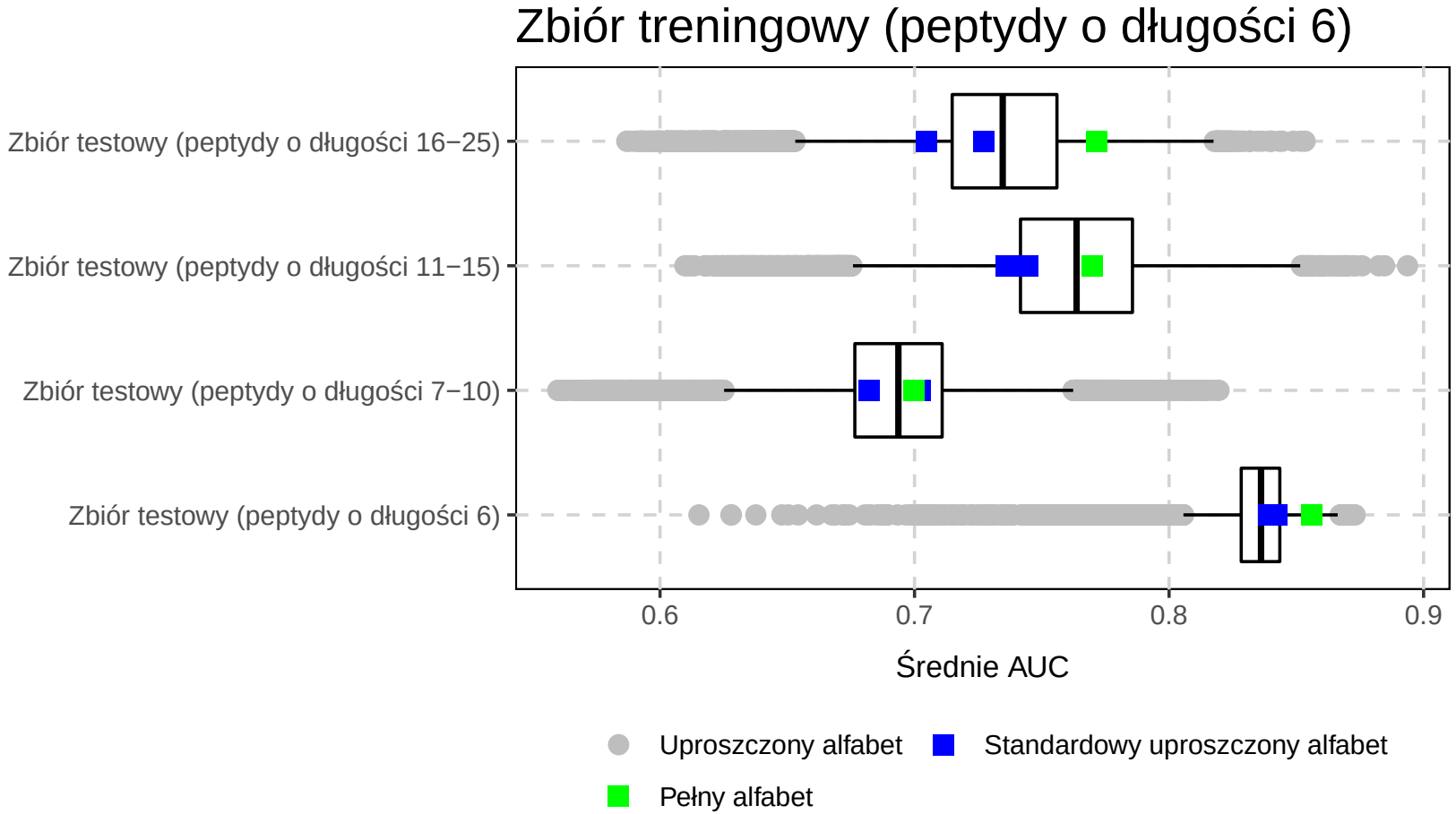
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

Nowe uproszczone alfabety



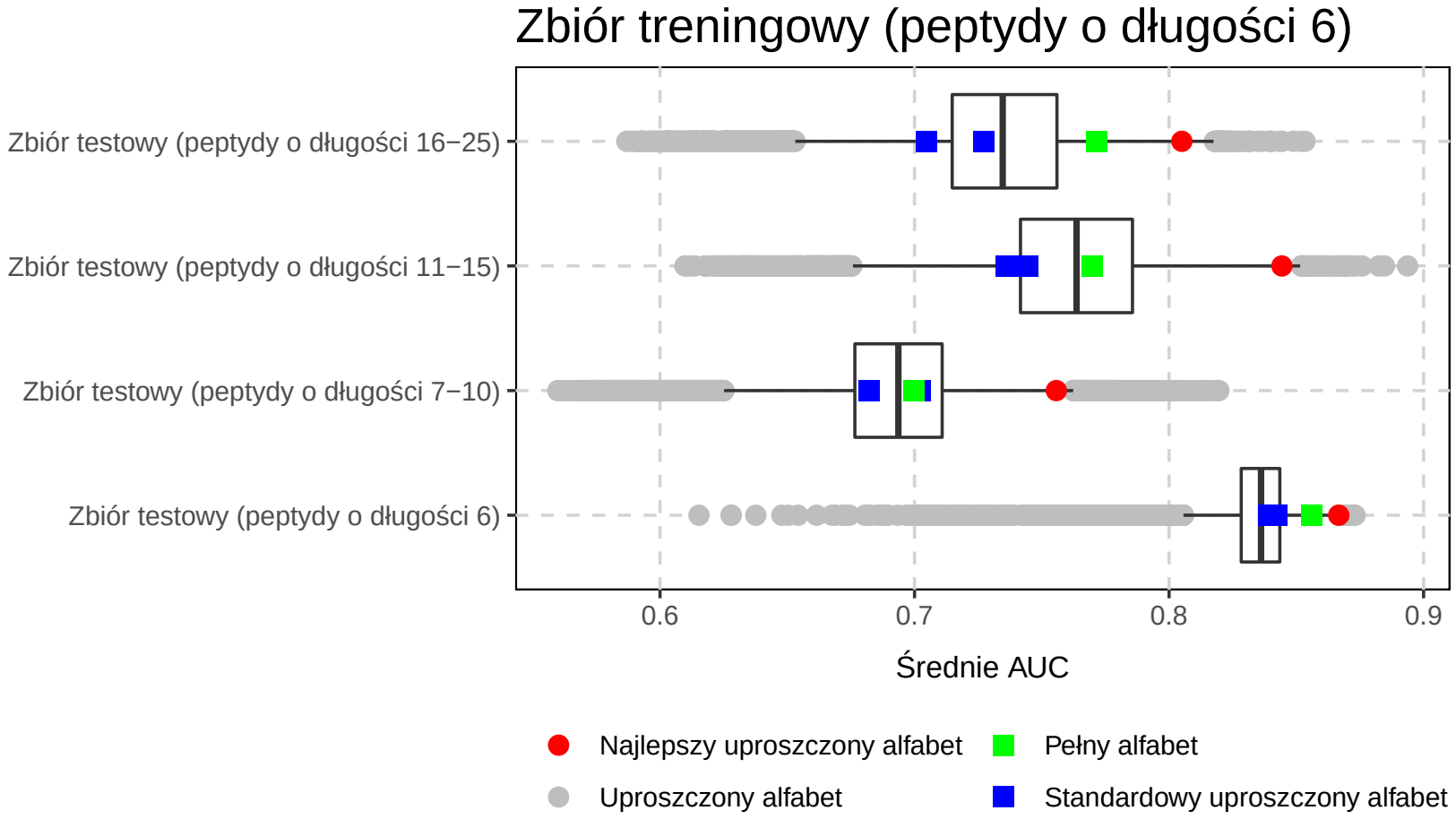
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

Walidacja krzyżowa



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



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.

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

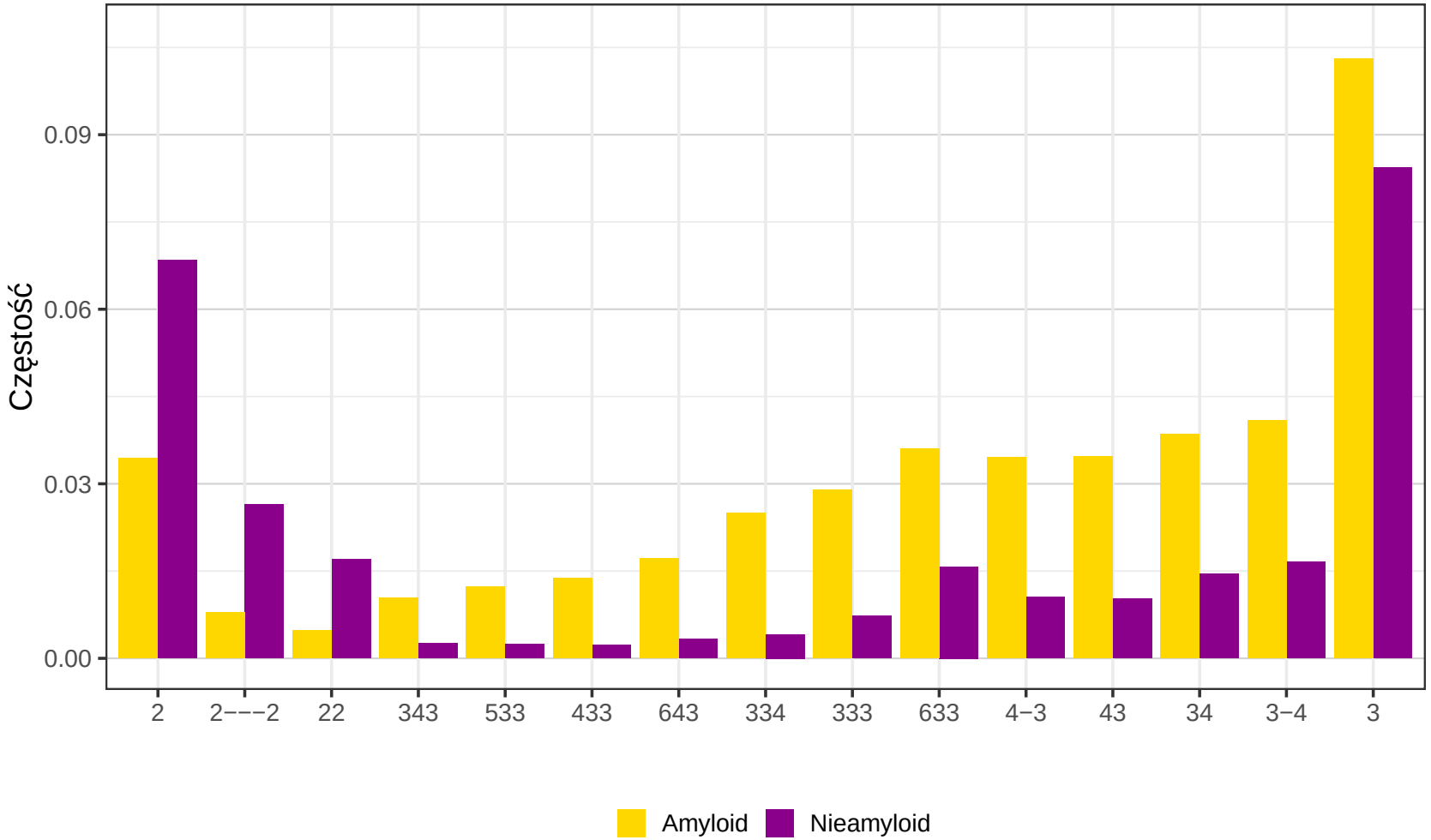
Grupa	Aminokwasy
1	G
2	K, P, R
3	I, L, V
4	F, W, Y
5	A, C, H, M
6	D, E, N, Q, S, T

Grupa 2 - reszty aminokwasowe zakłócające β-struktury.

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 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

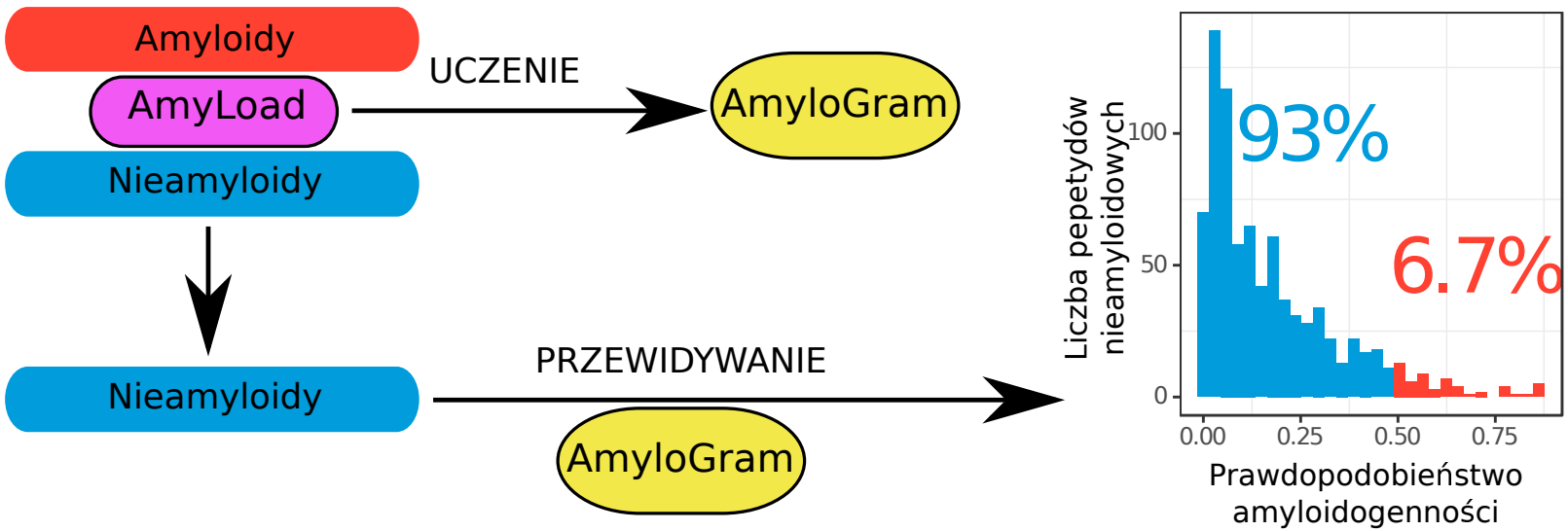
Porównanie z innymi narzędziami

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

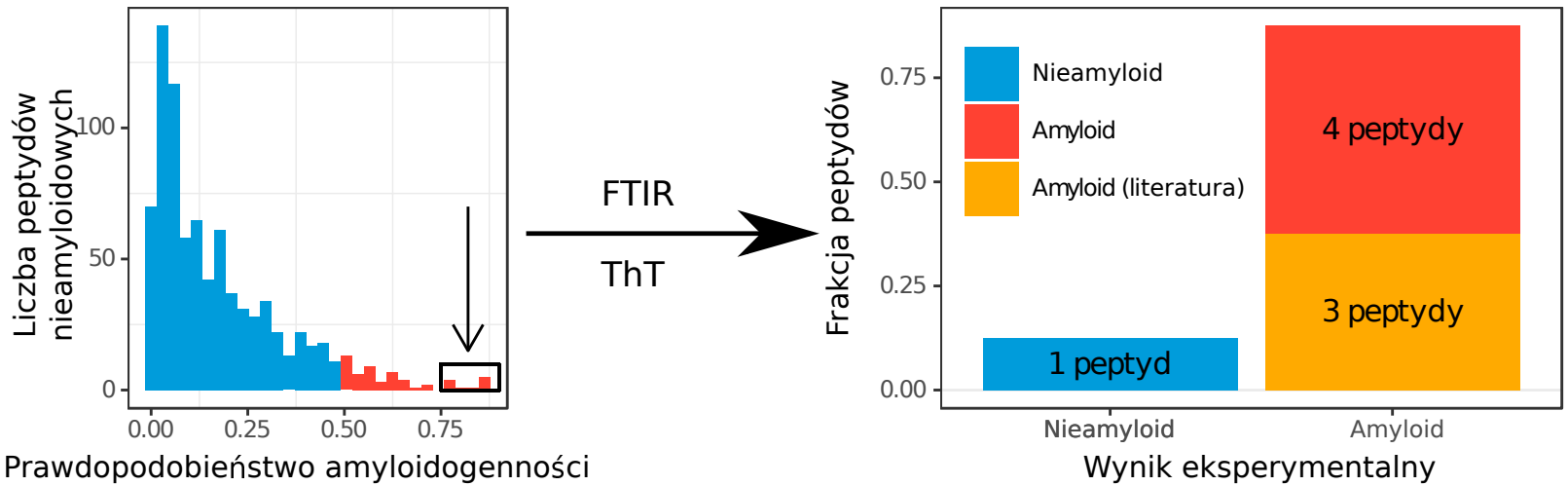
Klasyfikator wytrenowany z wykorzystaniem najlepszego uproszczonego alfabetu, AmyloGram, został porównany z innymi narzędziami do przewidywania amyloidów z użyciem zewnętrznego zbioru danych *pep424*.

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

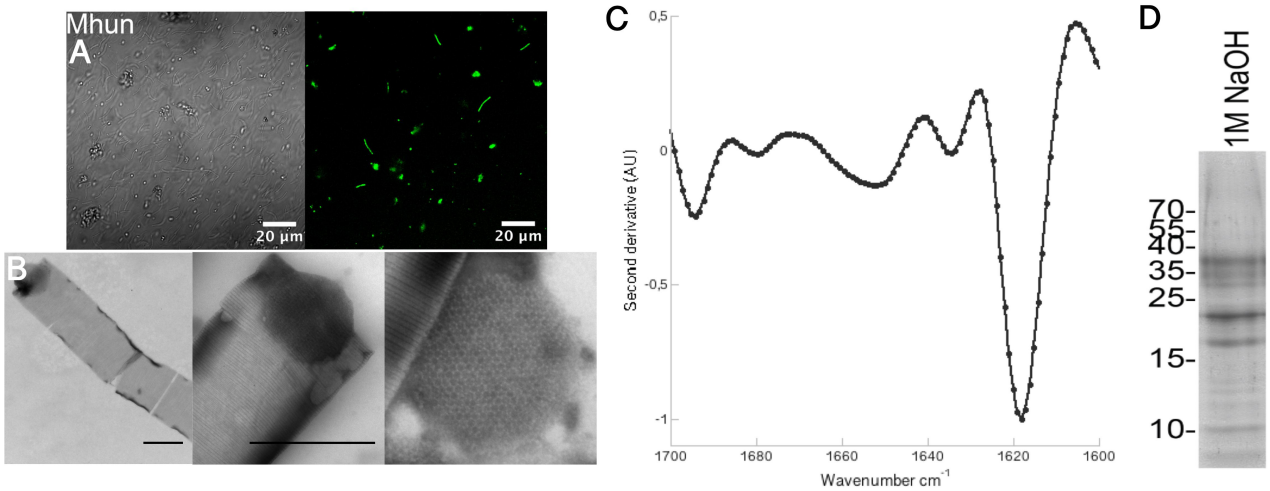
Walidacja eksperymentalna



Walidacja eksperymentalna

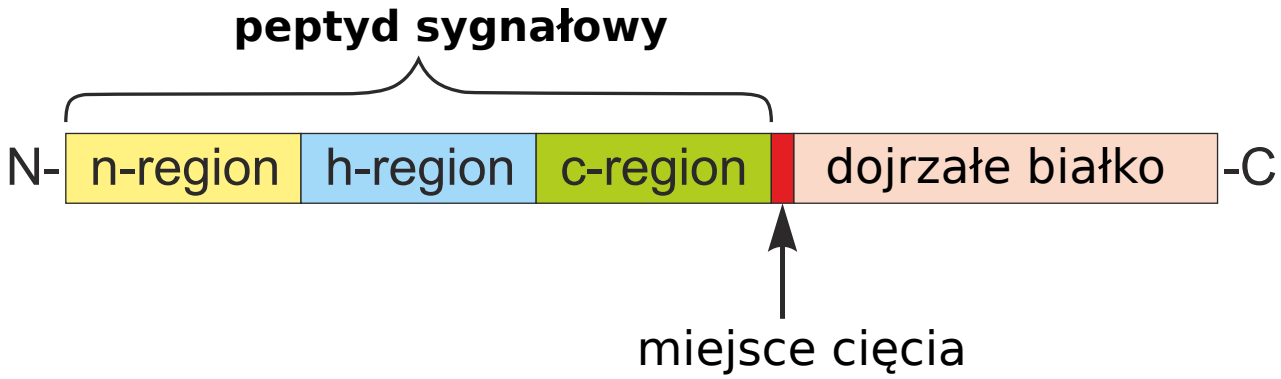


Nowe białko amyloidowe



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

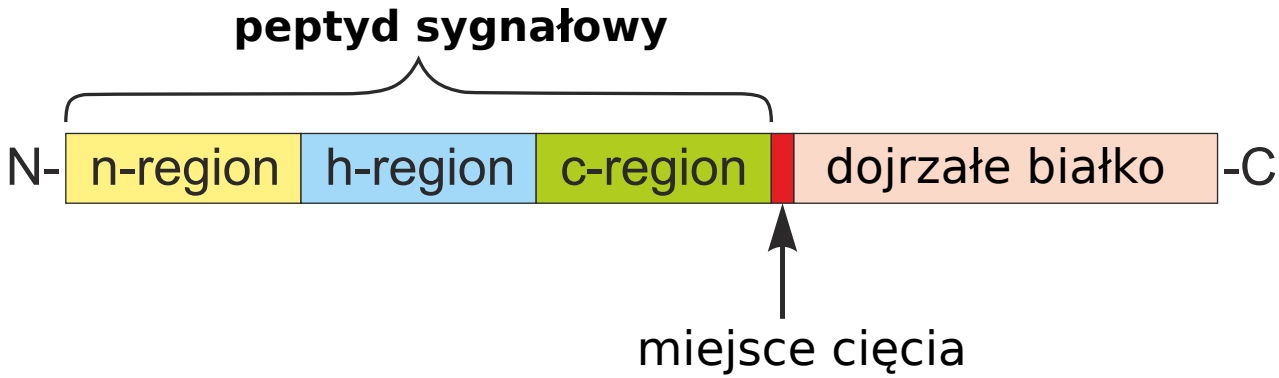
Peptydy sygnałowe



Peptydy sygnałowe:

- krótkie (20-30 reszt) N-końcowe fragmenty białek tworzące α-helisy,

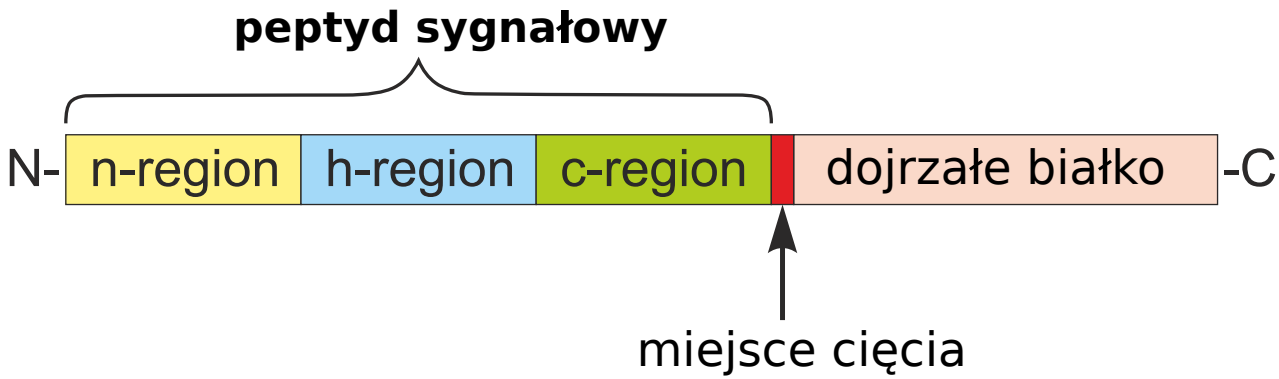
Peptydy sygnałowe



Peptydy sygnałowe:

- krótkie (20-30 reszt) N-końcowe fragmenty białek tworzące α-helisy,
- kierują białka do układu wewnątrzbłonowego a następne do sekrecji lub kompartymentów wewnątrzkomórkowch.

Peptydy sygnałowe

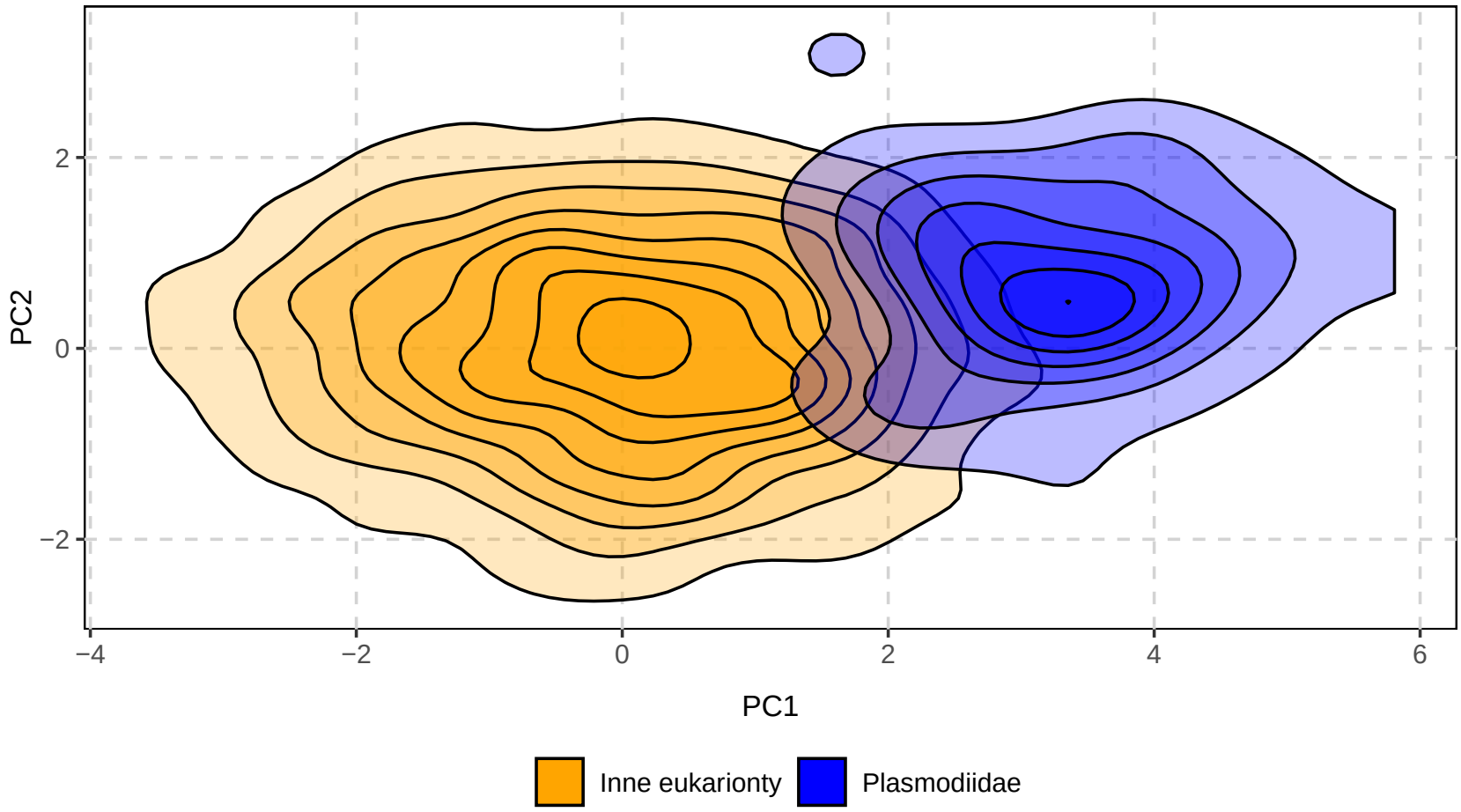


Peptydy sygnałowe są bardzo zmienne, ale zawsze zawierają trzy charakterystyczne domeny (Hegde and Bernstein, 2006):

- n-region: 5-8 dodatnio naładowanych reszt aminokwasowych (Nielsen and Krogh, 1998),
- h-region: bardzo hydrofobowe reszty (Nielsen and Krogh, 1998),
- c-region: kilka (3-5) polarnych reszt.

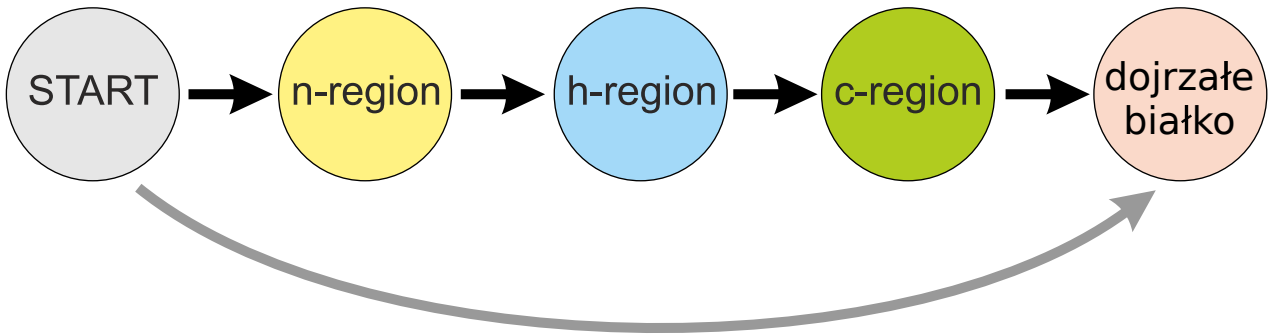
Peptydy sygnałowe

Skład aminokwasowy peptydów sygnałowych u *Plasmodium* sp. (do których należy m. in. zarodziec malarii) jest inny od składu peptydów sygnałowych innych eukariontów. Dlatego też narzędzia do przewidywania peptydów sygnałowych źle radzą sobie z przewidywaniem peptydów sygnałowych u *Plasmodium* sp.



signalHsmm

signalHsmm (Burdukiewicz et al., 2018): zastosowanie ukrytych modeli semi-Markowa i uproszczonych alfabetów aminokwasowych w celu przewidywania peptydów sygnałowych w białkach *Plasmodium* sp.



Porównanie z innymi predyktorami

signalHsmm efektywnie uczy się rozpoznawać peptydy sygnałowe na bardzo małych zbiorach danych.
Przewidywania signalHsmm są na tyle uniersalne, że pozwalają również przewiywać nietypowe peptydy sygnałowe spotykane w białkach *Plasmodium* sp.
W celu porównania się z innymi klasyfikatorami, powstały dwie iteracje signalHsmm: signalHsmm-2010, oparty na peptydach sygnałowych użytych do uczenia signalP
4.1 citeppetersen$_s$*ignalp*$_2$011*orazsignalHsmm* − 1987, *opartynadanychdostpnychw*1987, *kiedyopublikowanopierwszenarzdz*

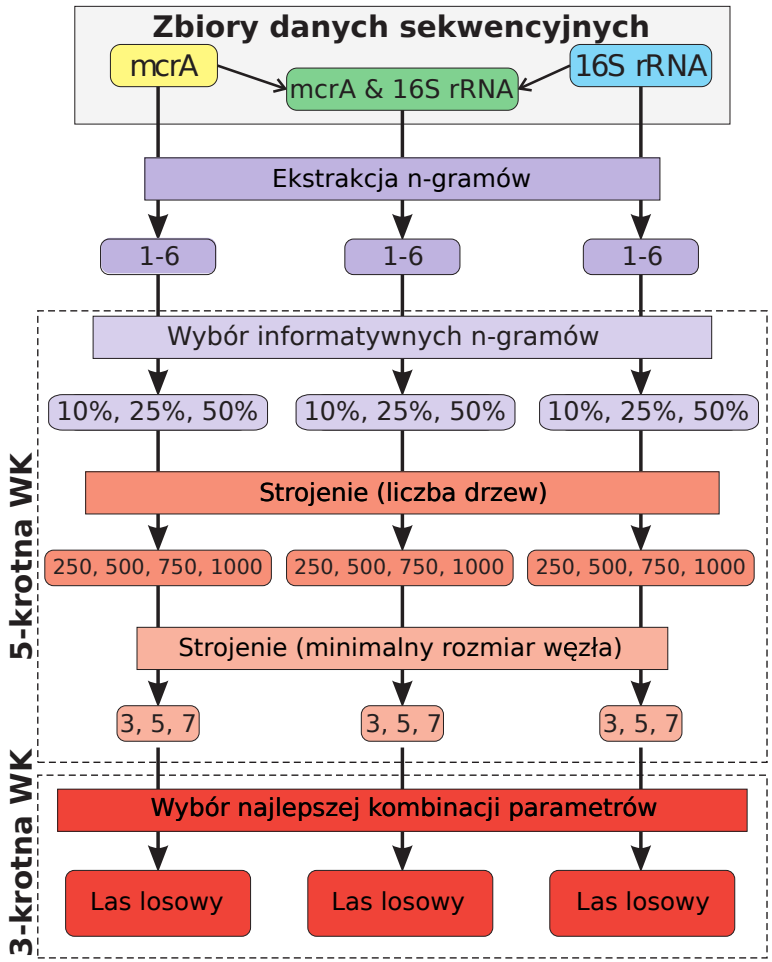
Burdukiewicz, M., Sobczyk, P., Chilimoniuk, J., Gagat, P., and Mackiewicz, P. (2018). Prediction of Signal Peptides in Proteins from Malaria Parasites. International Journal of Molecular Sciences 19, 3709.

Porównanie z innymi predyktorami

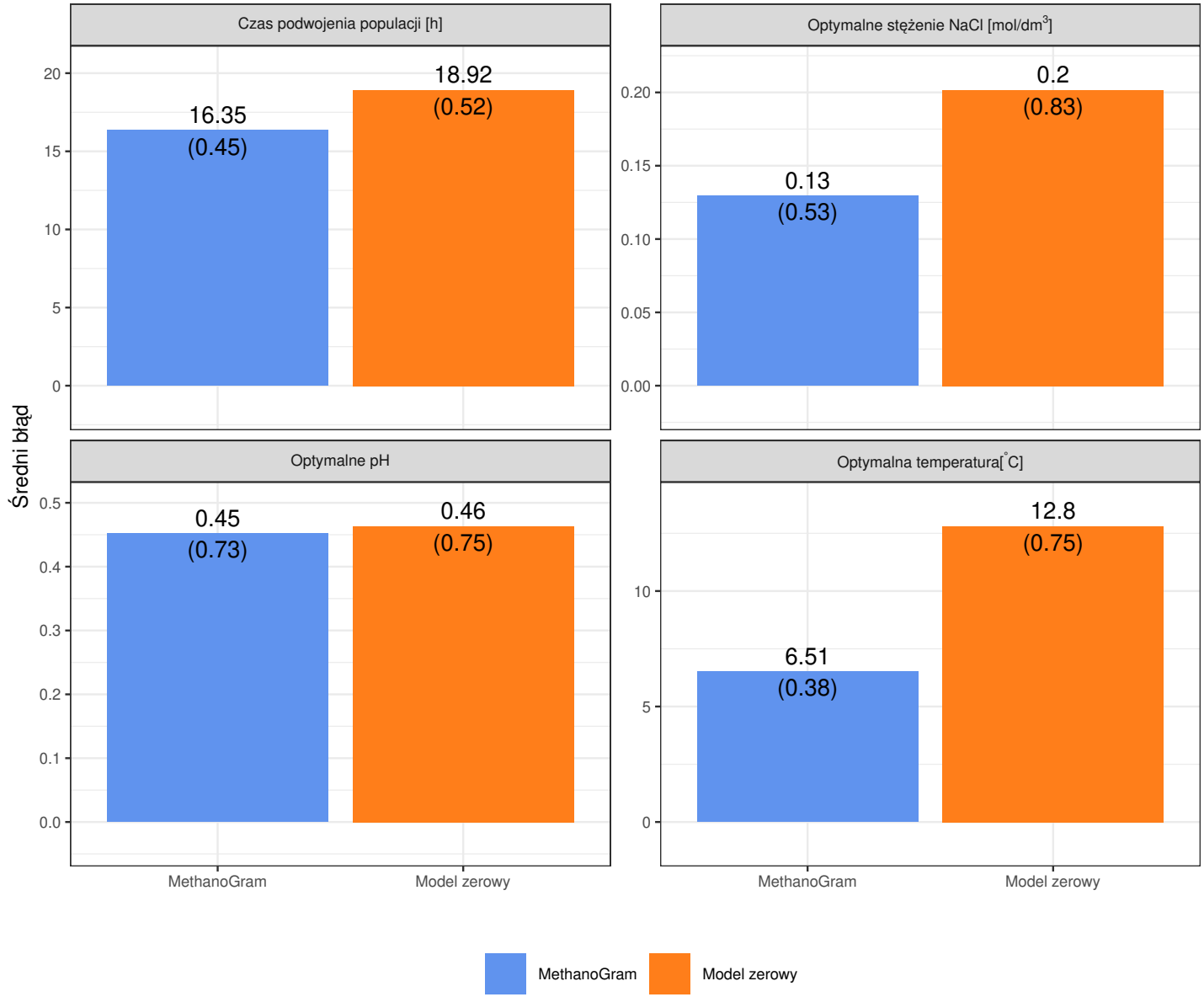
Algorytm	Czułość	Swoistość	MCC	AUC
signalP 4.1 (no tm)	0.8235	0.9100	0.6872	0.8667
signalP 4.1 (tm)	0.6471	0.9431	0.6196	0.7951
signalP 3.0 (NN)	0.8824	0.9052	0.7220	0.8938
signalP 3.0 (HMM)	0.6275	0.9194	0.5553	0.7734
PrediSi	0.3333	**0.9573**	0.3849	0.6453
Philius	0.6078	0.9336	0.5684	0.7707
Phobius	0.6471	0.9289	0.5895	0.7880
signalHsmm-2010	0.9804	0.8720	0.7409	0.9262
signalHsmm-2010 (ident. 50%)	**1.0000**	0.8768	**0.7621**	**0.9384**
signalHsmm-2010 (pełny alfabet)	0.8431	0.9005	0.6853	0.8718
signalHsmm-1987	0.9216	0.8910	0.7271	0.9063
signalHsmm-1987 (ident. 50%)	0.9412	0.8768	0.7194	0.9090
signalHsmm-1987 (pełny alfabet)	0.7647	0.9052	0.6350	0.8350

Burdukiewicz, M., Sobczyk, P., Chilimoniuk, J., Gagat, P., and Mackiewicz, P. (2018). Prediction of Signal Peptides in Proteins from Malaria Parasites. International Journal of Molecular Sciences 19, 3709.

Methanogram: narzędzie predykcyjne bazy PhyMet^2 (phymet2.biotech.uni.wroc.pl)



Burdukiewicz, M., Gagat, P., Jabłoński, S., Chilimoniuk, J., Gaworski, M., Mackiewicz, P., and Łukaszewicz, M. (2018). PhyMet2: a database and toolkit for phylogenetic and metabolic analyses of methanogens. Environmental Microbiology Reports 10, 378–382.



Burdukiewicz, M., Gagat, P., Jabłoński, S., Chilimoniuk, J., Gaworski, M., Mackiewicz, P., and Łukaszewicz, M. (2018). PhyMet2: a database and toolkit for phylogenetic and metabolic analyses of methanogens. Environmental Microbiology Reports 10, 378–382.

Podsumowanie

- Zastosowanie n-gramów (k-merów), czyli ciągów aminokwasów lub nukleotydów o długości n okazało się skuteczne do charakterystyki sekwencji peptydów sygnałowych i amyloidogennych oraz markerów molekularnych metanogenów, a następnie do przewidywania tych peptydów oraz optymalnych warunków hodowli metanogenów w oparciu o te markery.
- Zastosowanie uproszczonych alfabetów aminokwasowych, w których aminokwasy są grupowane ze względu na swoje podobieństwo fizykochemiczne lub funkcjonalne, okazało się bardzo dobrym podejściem do opisu sekwencji aminokwasowych o zróżnicowanym składzie i wyeksponowania ich cech istotnych do przewidywania peptydów sygnałowych i amyloidogennych.
- Opracowana procedura QuiPT (**Qui**ck **P**ermutation **T**est) pozwala na szybki wybór najbardziej informatywnych n-gramów i uzyskiwanie dokładnych p-wartości.
- Zastosowana metodologia tworzenia uproszczonych alfabetów umożliwia identyfikowania kluczowych cech związanych z charakterystycznymi cechami peptydów.

Podsumowanie

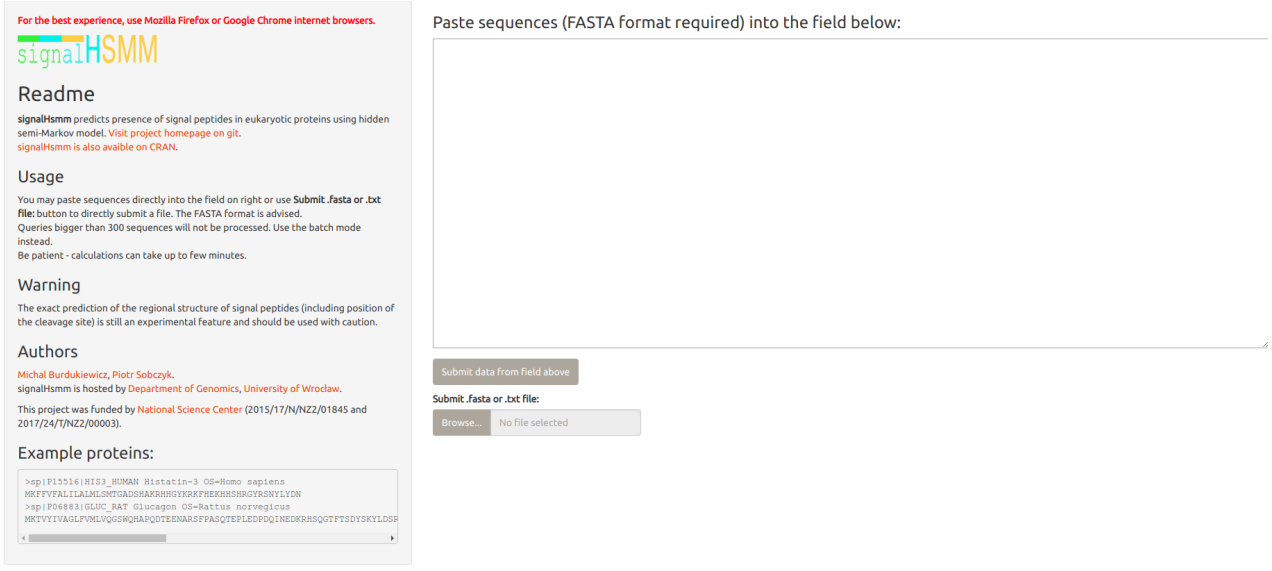
Web serwery:

- **AmyloGram**: `http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/`.
- **MethanoGram**: `http://www.smorfland.uni.wroc.pl/shiny/MethanoGram/`.
- **signalHsmm**: `http://www.smorfland.uni.wroc.pl/shiny/signalHsmm/`.

Pakiety **R**:

- **biogram**: `https://cran.r-project.org/package=biogram`.
- **AmyloGram**: `https://cran.r-project.org/package=AmyloGram`.
- **signalHsmm**: `https://cran.r-project.org/package=signalHsmm`.

Podsumowanie



Podziękowania

Mentorzy:

- **Paweł Mackiewicz (Uniwersytet Wrocławski)**.
- Małgorzata Kotulska (Politechnika Wrocławska).
- Marcin Łukaszewicz (Uniwersytet Wrocławski).
- Andrzej Dąbrowski (Uniwersytet Wrocławski).
- Stefan Rödiger (Brandenburg University of Technology Cottbus-Senftenberg).
- Henrik Nielsen (Technical University of Denmark).
- Lars Kaderali (University of Greifswald).
- Andreas Weinhäusel (Austrian Institute of Technology).

Podziękowania

Współpracownicy:

- Przemysław Gagat (Uniwersytet Wrocławski).
- Jarosław Chilimoniuk (Uniwersytet Wrocławski).
- Rafał Kolenda (Uniwersytet Przyrodniczy we Wrocławiu).
- Piotr Sobczyk (Politechnika Wrocławska).
- Sławomir Jabłoński (Uniwersytet Wrocławski).
- Marlena Gąsior-Głogowska (Politechnika Wrocławska).
- Chris Lauber (Technical University Dresden).
- Anna Duda-Madej (Uniwersytet Medyczny im. Piastów Śląskich we Wrocławiu).

Podziękowania

Finansowanie:

- Narodowe Centrum Nauki (Preludium i Etiuda).
- COST ACTION CA15110 (Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research).
- KNOW Wrocław Center for Biotechnology.
- InnoProfile-Transfer-Projekt 03IPT611X przyznanym przez Ministerstwo Edukacji i Badań Naukowych Niemiec.

Publikacje I

1. Burdukiewicz, M., Sobczyk, P., Chilimoniuk, J., Gagat, P., and Mackiewicz, P., *Prediction of Signal Peptides in Proteins from Malaria Parasites*. **International Journal of Molecular Sciences**, 2018.

2. Kolenda R., Burdukiewicz M., Schiebel J., Rödiger S, Sauer L., Szabo I., Orłowska A., Weinreich J., Nitschke J., Böhm, A., Gerber U., Roggenbuck D., Schierack P., *Adhesion of Salmonella to pancreatic secretory granule membrane major glycoprotein GP2 of human and porcine origin depends on FimH sequence variation*, **Frontiers in microbiology**, 2018.

3. Mackiewicz D., Posacki P., Burdukiewicz M., Błażej P. *Role of recombination and faithfulness to partner in sex chromosome degeneration*. **Scientific Reports**, 2018.

4. Burdukiewicz M., Gagat P. Jabłoński S., Chilimoniuk J., Gaworski M., Mackiewicz P., Łukaszewicz M. *PhyMet2: a database and toolkit for phylogenetic and metabolic analyses of methanogens*. **Environmental Microbiology Reports**, 2018.

Publikacje II

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

6. Schiebel J., Böhm A., Nitschke J., Burdukiewicz M., Weinreich J., Ali A., Roggenbuck D., Rödiger S., Schierack P., *Genotypic and phenotypic characteristics in association with biofilm formation in different pathotypes of human clinical Escherichia coli isolates*, **Applied and Environmental Microbiology**, 2017.

7. Rödiger S., Burdukiewicz M., Spiess A.-N., Blagodatskikh K., *Enabling reproducible real-time quantitative PCR research: the RDML package*. **Bioinformatics**, 2017.

8. Burdukiewicz M., Rödiger S., Sobczyk P., Menschikowski M., Schierack P., Mackiewicz P., *Methods for comparing multiple digital PCR experiments*, **Biomolecular Detection and Quantification**, 2016.

Publikacje III

9. Spiess A.-N., Rödiger S., Burdukiewicz M., Volksdorf T., Tellinghuisen J., *System- specific periodicity in quantitative real-time polymerase chain reaction data questions threshold-based quantitation*, **Scientific Reports**, 2016.

10. Kolenda R., Burdukiewicz M., Schierack P., *A systematic review and meta-analysis of the epidemiology of pathogenic escherichia coli of calves and the role of calves as reservoirs for human pathogenic E. coli*. **Frontiers in Cellular and Infection Microbiology**, 2015.

11. Rödiger S., Burdukiewicz M., Schierack P., *chipPCR: an R Package to Pre-Process Raw Data of Amplification Curves*. **Bioinformatics**, 2015.

12. Rödiger S., Burdukiewicz M., Blagodatskikh K., Jahn M., Schierack P., *R as an Environment for the Reproducible Analysis of DNA Amplification Experiments*, **R Journal**, 2015.

Publikacje IV

13. Spiess A.-N., Deutschmann C., Burdukiewicz M., Himmelreich R., Klat K., Schierack P., Rödiger S., *Impact of smoothing on parameter estimation in quantitative dna amplification experiments*. **Clinical Chemistry**, 2014.

Podsumowanie

- Zastosowanie n-gramów (k-merów), czyli ciągów aminokwasów lub nukleotydów o długości n okazało się skuteczne do charakterystyki sekwencji peptydów sygnałowych i amyloidogennych oraz markerów molekularnych metanogenów, a następnie do przewidywania tych peptydów oraz optymalnych warunków hodowli metanogenów w oparciu o te markery.
- Zastosowanie uproszczonych alfabetów aminokwasowych, w których aminokwasy są grupowane ze względu na swoje podobieństwo fizykochemiczne lub funkcjonalne, okazało się bardzo dobrym podejściem do opisu sekwencji aminokwasowych o zróżnicowanym składzie i wyeksponowania ich cech istotnych do przewidywania peptydów sygnałowych i amyloidogennych.
- Opracowana procedura QuiPT (**Qui**ck **P**ermutation **T**est) pozwala na szybki wybór najbardziej informatywnych n-gramów i uzyskiwanie dokładnych p-wartości.
- Zastosowana metodologia tworzenia uproszczonych alfabetów umożliwia identyfikowania kluczowych cech związanych z charakterystycznymi cechami peptydów.

References I

Burdukiewicz, M., Sobczyk, P., Chilimoniuk, J., Gagat, P., and Mackiewicz, P. (2018). Prediction of Signal Peptides in Proteins from Malaria Parasites. *International Journal of Molecular Sciences*, 19(12):3709.

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.

References II

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.

Hegde, R. S. and Bernstein, H. D. (2006). The surprising complexity of signal sequences. *Trends in Biochemical Sciences*, 31(10):563–571.

Kosiol, C., Goldman, N., and Buttimore, N. H. (2004). A new criterion and method for amino acid classification. *Journal of Theoretical Biology*, 228(1):97–106.

Melo, F. and Marti-Renom, M. A. (2006). Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. *Proteins*, 63(4):986–995.

References III

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.

Nielsen, H. and Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 6:122–130.

Paz, M. L. d. l. 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 IV

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.

Upraszczanie alfabetów i QuiPT - wyjaśnienia

- QuiPT został zaprojektowany do zadań typowych w przewidywaniu właściwości sekwencji, gdzie często stosuje się dane binarne.
- Zwykłe metody upraszczania alfabetów zalecają duże (np. powyżej 1000 białek zawierających ok. 26000 reszt aminokwasowych) zbiory danych. Bacardit, J., Stout, M., Hirst, J.D., Valencia, A., Smith, R.E., and Krasnogor, N. (2009). Automated Alphabet Reduction for Protein Datasets. BMC Bioinformatics 10, 6.

signalHsmm - wyjaśnienia I

- Trudno stworzyć program lepszy dla ogólu peptydów sygnałowych niż signalP 4.1, ponieważ większość z nich jest często najpierw idenyfikowana *in silico* przez signalP 4.1 (komunikacja ustna: H. Nielsen)
- Nie ma dużego zbioru danych peptydów sygnałowych w białkach *Plasmodium* sp. potwierdzonych eksperymentalnie. Inne narzędzia np. ApicoAP też wykorzystują zbiór danych użyty do walidacji signalHsmm. Właściwości tych potencjalnych peptydów sygnalowych są ewidentne, a same peptydy występują przy białkach ulegających sekrecji.

AmyloGram - wyjaśnienia I

Zdefiniowane długości n-gramów pochodzą z prac eksperymentalnych definiujących motywy amyloidogenne:

- de Groot, N.S., Parella, T., Aviles, F.X., Vendrell, J., and Ventura, S. (2007). *Ile-Phe Dipeptide Self-Assembly: Clues to Amyloid Formation*. Biophysical Journal 92, 1732–1741.
- Ivanova, M.I., Thompson, M.J., and Eisenberg, D. (2006). *A systematic screen of 2-microglobulin and insulin for amyloid-like segments*. Proc Natl Acad Sci U S A 103, 4079–4082.
- Sant'Anna, R., Fernández, M.R., Batlle, C., Navarro, S., de Groot, N.S., Serpell, L., and Ventura, S. (2016). Characterization of Amyloid Cores in Prion Domains. Scientific Reports 6, 34274.

AmyloGram - wyjaśnienia

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*.

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

MethanoGram - wyjaśnienia

- MethanoGram trafnie przewiduje tylko dwa spośród czterech wybranych warunków hodowlanych, ale są to jedne z kluczowych parametrów.
- Narzędzie jest dedykowane dla wąskich grup mikroorganizmów. Jego zasada działania w oparciu o markery molekularne, utrudnia zastosowanie do szerszych grup.

QuiPT - wyjaśnienia I

1. Hipoteza testu: dwustronna.
   H_0: częstość występowania n-gramu z danej sekwencji nie jest związane z przynależnością tej sekwencji do danej grupy.
   H_A: częstość występowania n-gramu z danej sekwencji jest związane z przynależnością tej sekwencji do danej grupy.
2. Statystyka testowa: wzajemna informacja I zdefiniowana następującym wzorem:
   $\mathrm{I}(X;Y) = \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$
   - p(x,y): wspólny rozkład prawdopodobieństwa X i Y.
   - p(x): rozkład brzegowy X.
   - X: zmienna losowa o rozkładzie dwumianowym (obecność n-gramu w sekwencji).
   - Y: zmienna losowa o rozkładzie dwumianowym (przynależność sekwencji do grupy).
3. Moc testu opartego na wzajemnej informacji:

QuiPT - wyjaśnienia II

- François, D., Wertz, V., and Verleysen, M. (2006). *The permutation test for feature selection by mutual information*. Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN 2006).
- Szymczak, S., Nuzzo, A., Fuchsberger, C., Schwarz, D.F., Ziegler, A., Bellazzi, R., and Igl, B.-W. (2007). *Genetic association studies for gene expressions: permutation-based mutual information in a comparison with standard ANOVA and as a novel approach for feature selection*. BMC Proc 1, S9.
- Airola, A., Pahikkala, T., Boberg, J., and Salakoski, T. (2010). *Applying Permutation Tests for Assessing the Statistical Significance of Wrapper Based Feature Selection*. In 2010 Ninth International Conference on Machine Learning and Applications, pp. 989–994.

QuiPT - wyjaśnienia III

4. Test został nazwany permutacyjnym, ponieważ zastępuje testy permutacyjne niezależności rozkładów. W literaturze przedmiotu testy te zwyczajowo nazywane są permutacyjnymi, dlatego przyjęto nazwę w konwencji stosowanej przez typowych użytkowników. Dzięki tej nazwie test będzie lepiej rozpoznawany przez potencjalnych użytkowników.
5. Stosuje się test dwustronny, ponieważ zarówno elementy nadreprezentowane jak i niedoreprezentowane ponieważ te ostatnie wnoszą cenną informację o układach, które są zabronione w danym układzie.
6. p-wartości: dowolnie małe w zależności od zbioru danych.

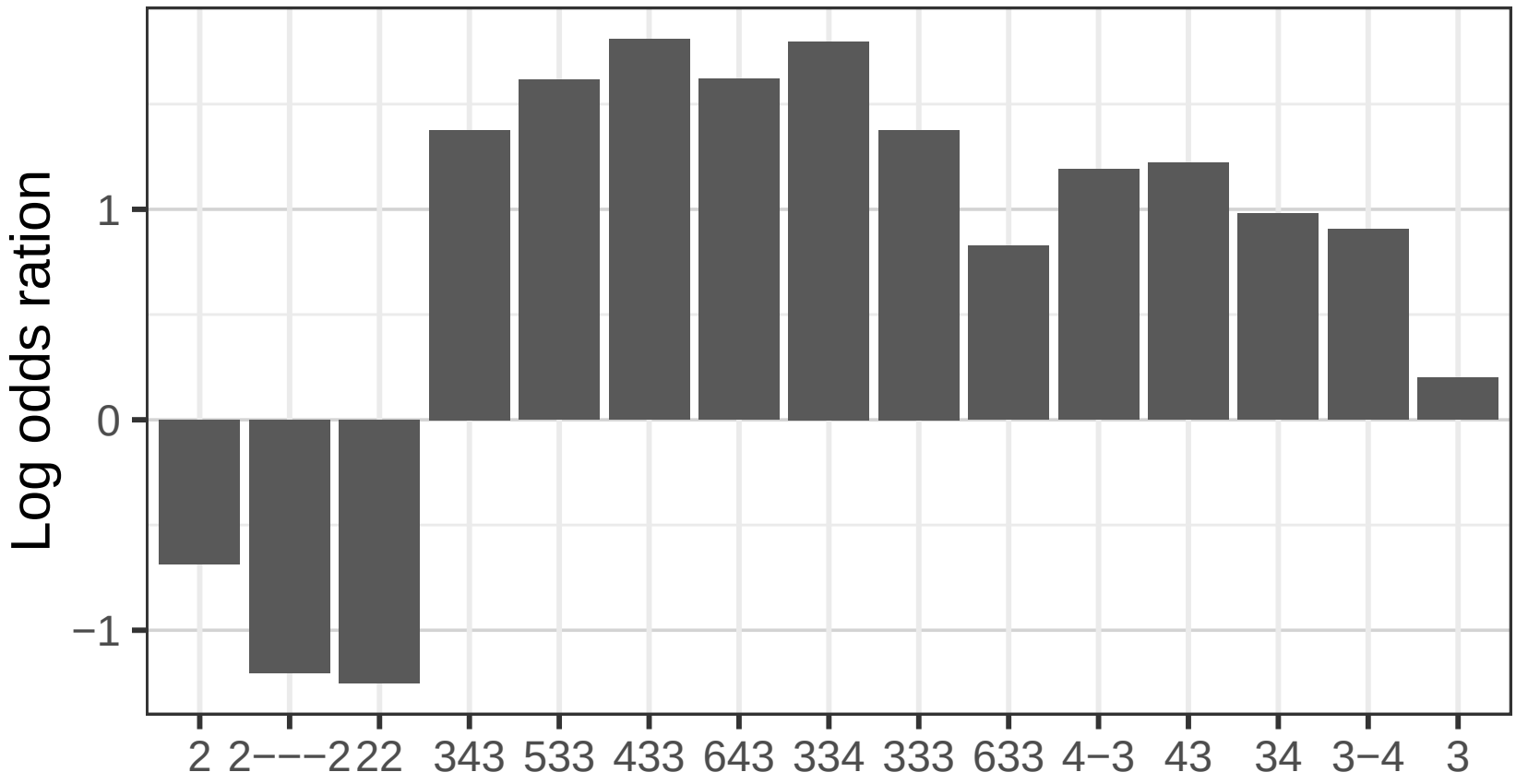
Upraszczanie alfabetów - wyjaśnienia I

- Grupowanie aminokwasów: odleglość euklidesowa, normalizacja 0-1.
- Uzyskanie reprezentacji alfabetu aminokwasowego dla małej liczby sekwencji nie jest możliwe za pomocą metod głębokich.

signalHsmm - wyjaśnienia I

- W literaturze predmiotu preferowane są modele o największej czułości (Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions. Nature Methods 8, 785). Wybrany alfabet zapewnia zarówno największą czułość jak i AUC.
- Odchylenia standardowe uzyskanych alfabetów zostały obliczone w drodze 10-krotnej walidacji krzyżowej. Ich niewielka zmienność wynika z dużej stabilności uzyskanego klasyfikatora, którą można zaobserwować porównując iterację uczoną na danych z 1987 roku i 2010 roku.
- Wnioski na stronie 40 wynikają z wielu tabel i rysunków, przykładowo rys. 9.
- PCA: przedstawone na rysunku dwa pierwsze komponenty składowe wyjaśniają 40% wariancji.

AmyloGram - wyjaśnienia



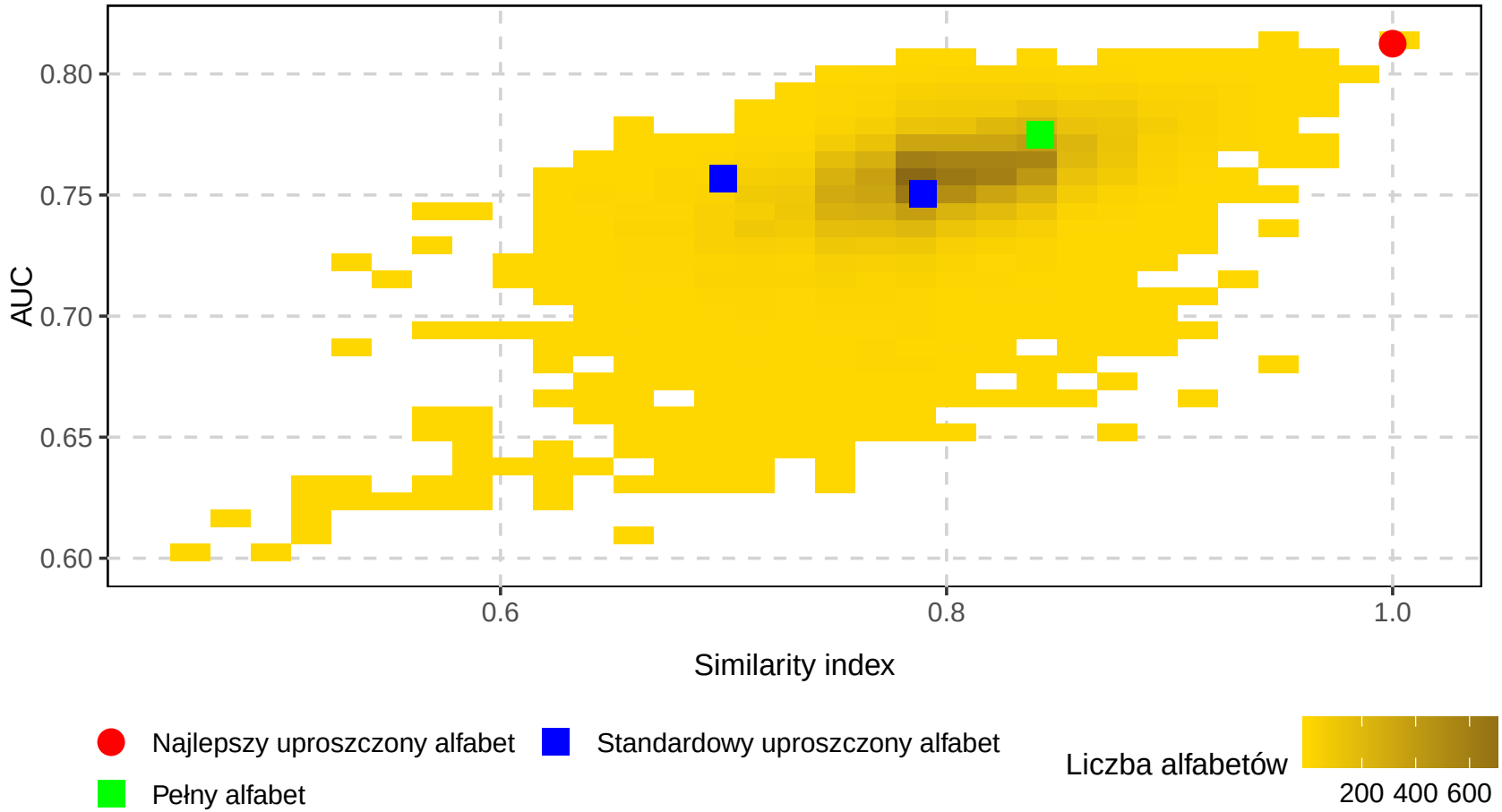
MethanoGram - wyjaśnienia

Mały zbiór danych (61 gatunków) nie pozwala na tworzenie modelu uwzględniającego zbyt wiele cech.

Podobieństwo alfabetów i jakość predykcji

Czy alfabety podobne do najlepszego uproszczonego alfabetu również dobrze przewidują amyloidy?

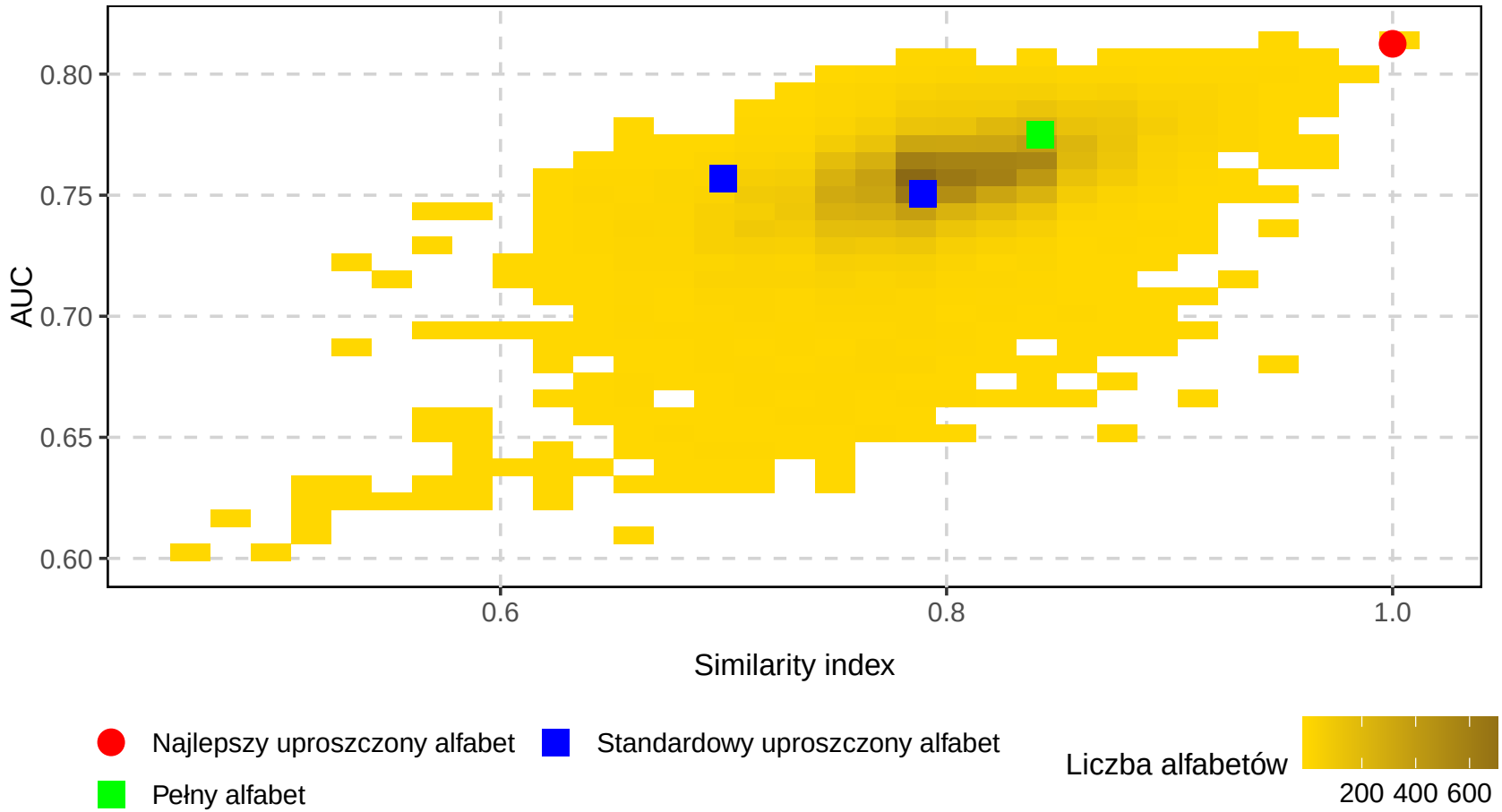
Similarity index



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

Similarity index



Korelacja między similarity index i średnim AUC jest istotna ($\text{p-value} \leqslant 2.2^{-16}$; $\rho = 0.51$). 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