Przewidywanie właściwości sekwencji biologicznych w oparciu o analize n-gramów

Michał Burdukiewicz

Zakład Bioinformatyki i Genomiki, Uniwersytet Wrocławski

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

Plan prezentacji

Badania *In silico* pozwalaja efektywniej planować prace eksperymentalne.
Przykłady:

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

Cel

Opracowanie metodologii analizy sekwencji biologicznych opierajacej sie na zrozumiałych dla człowieka regułach decyzyjnych.

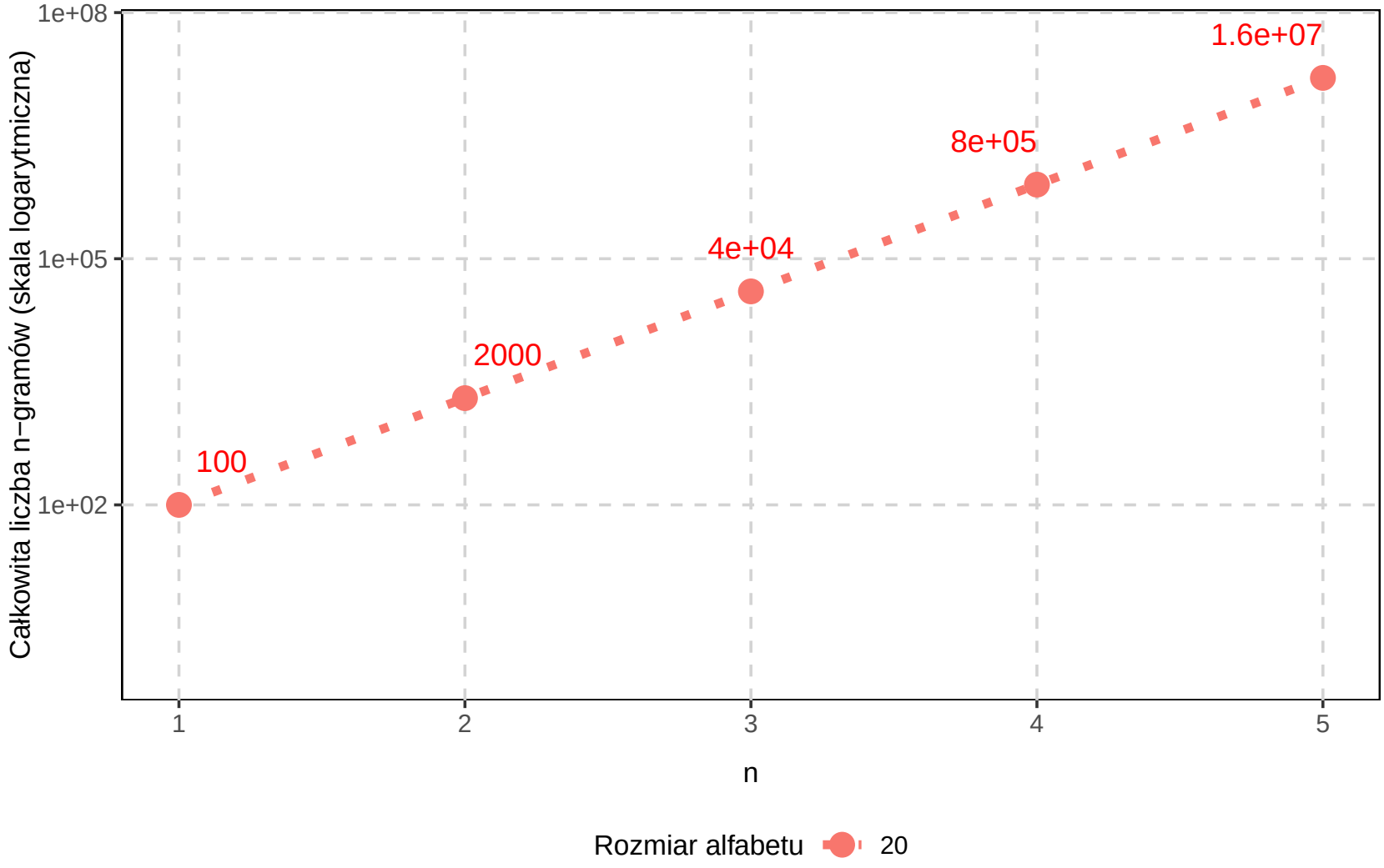
n-gramy (k-tuple, k-mery):

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

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
S1	H	T	E	S	Q	R	C	W	Y	M
S2	A	Q	R	G	N	D	K	I	P	V

Przykłady n-gramów:

1. H, A, T, Q
2. HT, AQ, TE, QR
3. H-E, A-R, T-S, Q-G
4. H–SQ, A–GN, T–QR, Q–ND



Dłuższe n-gramy sa bardziej informatywne, ale tworza wieksze przestrzenie atrybutów.

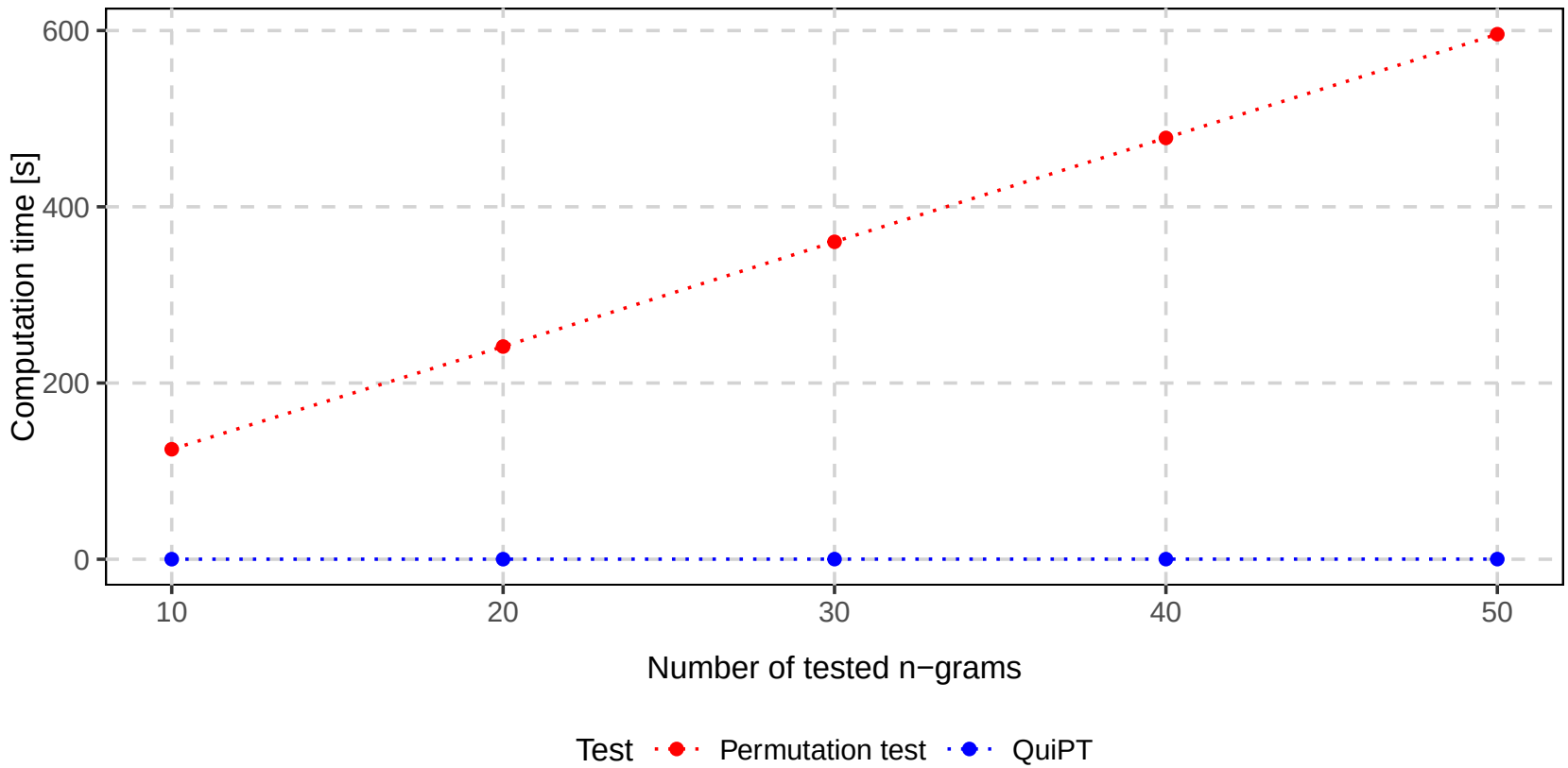
Testy permutacyjne

Informatywne n-gramy sa zazwyczaj wybierane za pomoca testów permutacyjnych.
During a permutation test we shuffle randomly class labels and compute a defined statistic (e.g. information gain). Values of statistic for permuted data are compared with the value of statistic for original data.

$$\text{p-value} = \frac{N_{T_P > T_R}}{N}$$

$N_{T_P > T_R}$: number of cases, where T_P (permuted test statistic) has more extreme values than T_R (test statistic for original data).
N: number of permutations.

QuiPT



QuiPT (available as part of the **biogram** R package) is faster than classical permutation tests and returns exact p-values.

Uproszczone alfabety:

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

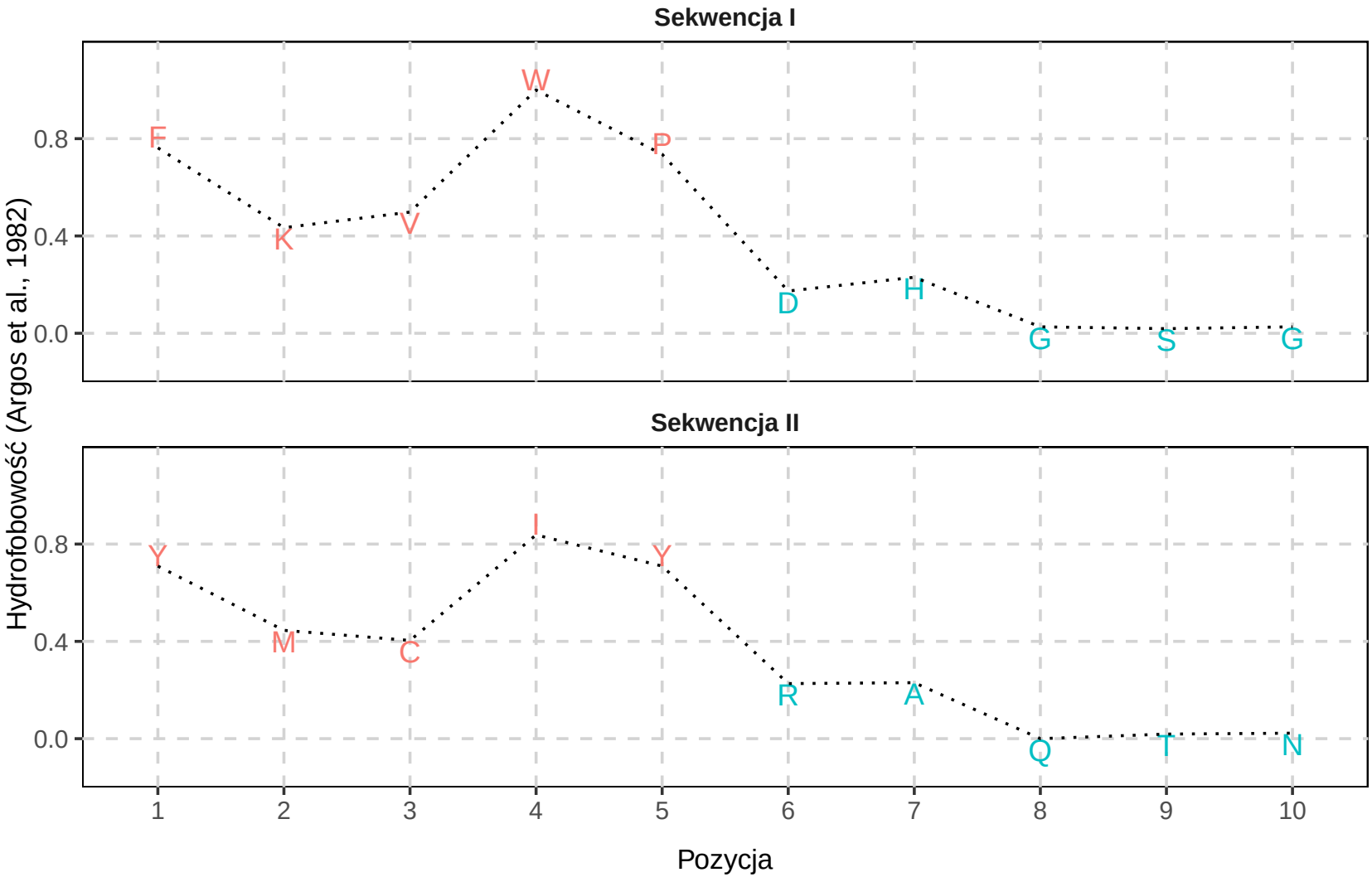
Poniższe peptydy wydaja sie być całkowicie inne pod wzgledem składu aminokwasowego.

Sekwencja I:

`FKVWPDHGSG`

Sekwencja II:

`YMCIYRAQTN`

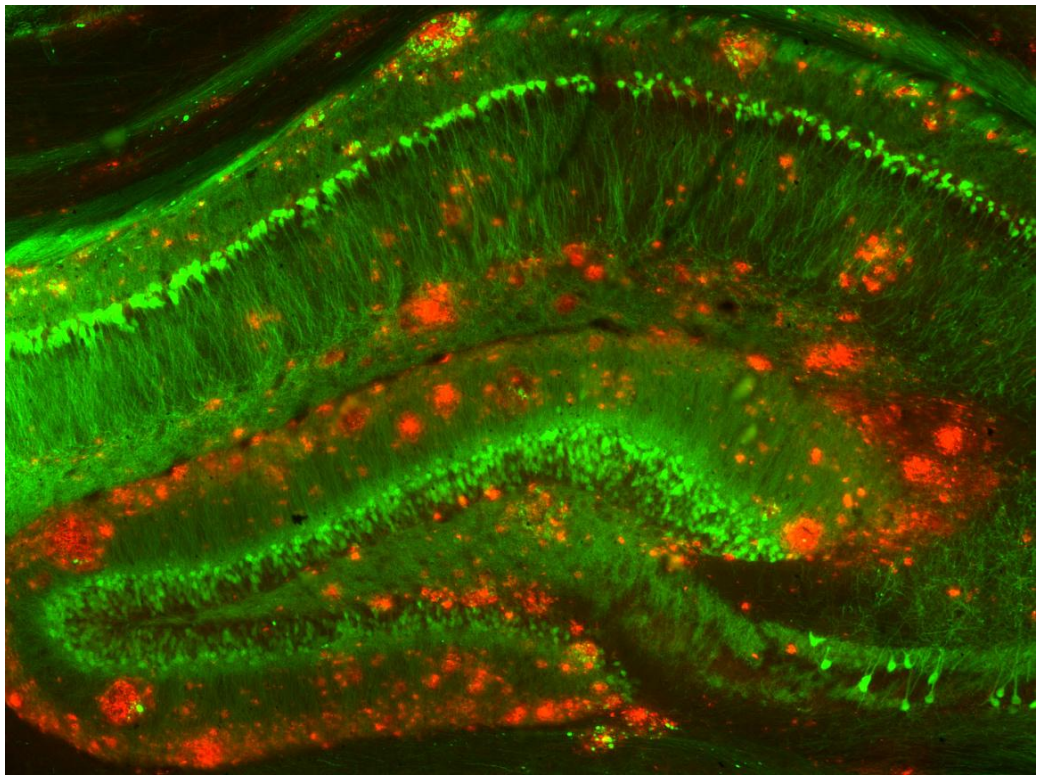

Sekwencja I
F
K
V
W
P
D
H
G
S
G
Sekwencja II
Y
M
C
I
Y
R
A
Q
T
N
Hydrofobowość (Argos et al., 1982)
0.0
0.4
0.8
1
2
3
4
5
6
7
8
9
10
Pozycja

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

Sekwencja I: `FKVWPDHGSG` `1111122222`
Sekwencja II: `YMCIYRAQTN` `1111122222`

Białka amyloidowe

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

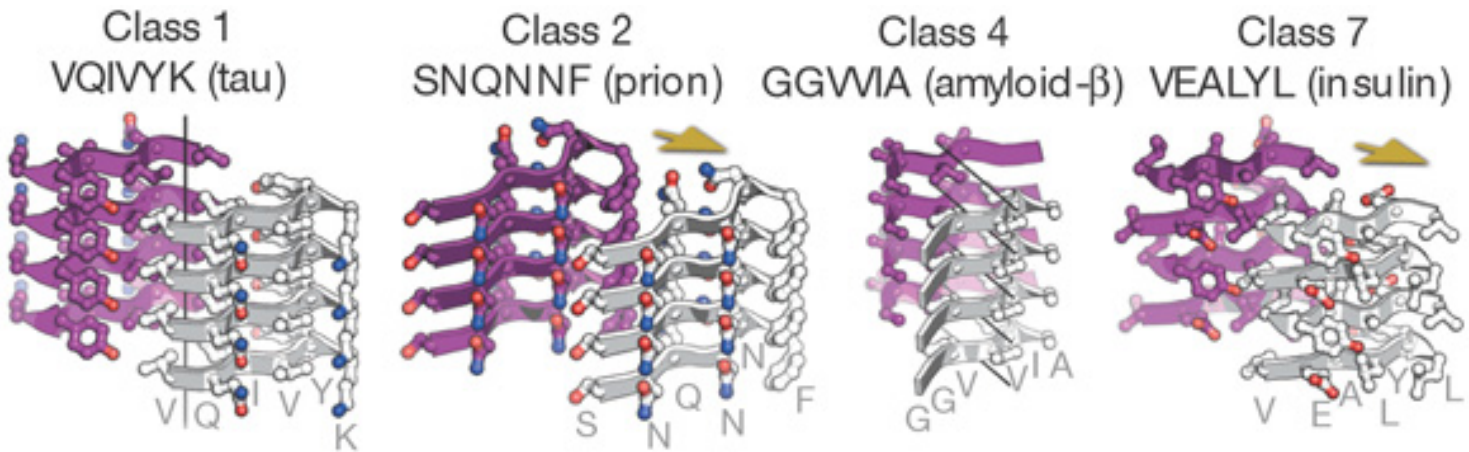


Agregaty amyloidowe (czerwone) wokół neuronów (zielone). Strittmatter Laboratory, Yale University.

Białka amyloidowe

Za agregacje białek amyloidogennych odpowiedzialne sa sekwencje peptydowe o właściwościach amyloidogennych (hot spots):

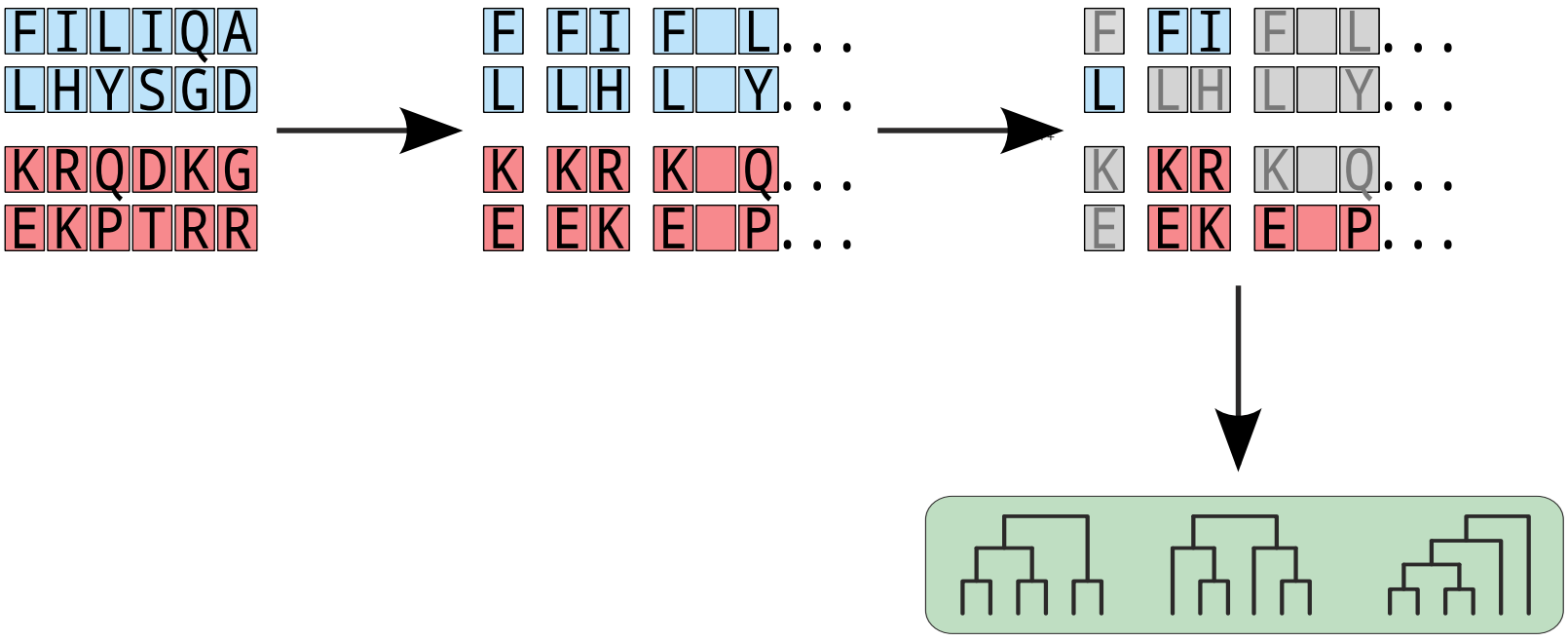
- krótkie (6-15 aminokwasów),
- bardzo zmienny skład aminokwasowy,
- tworza unikalne β-struktury.



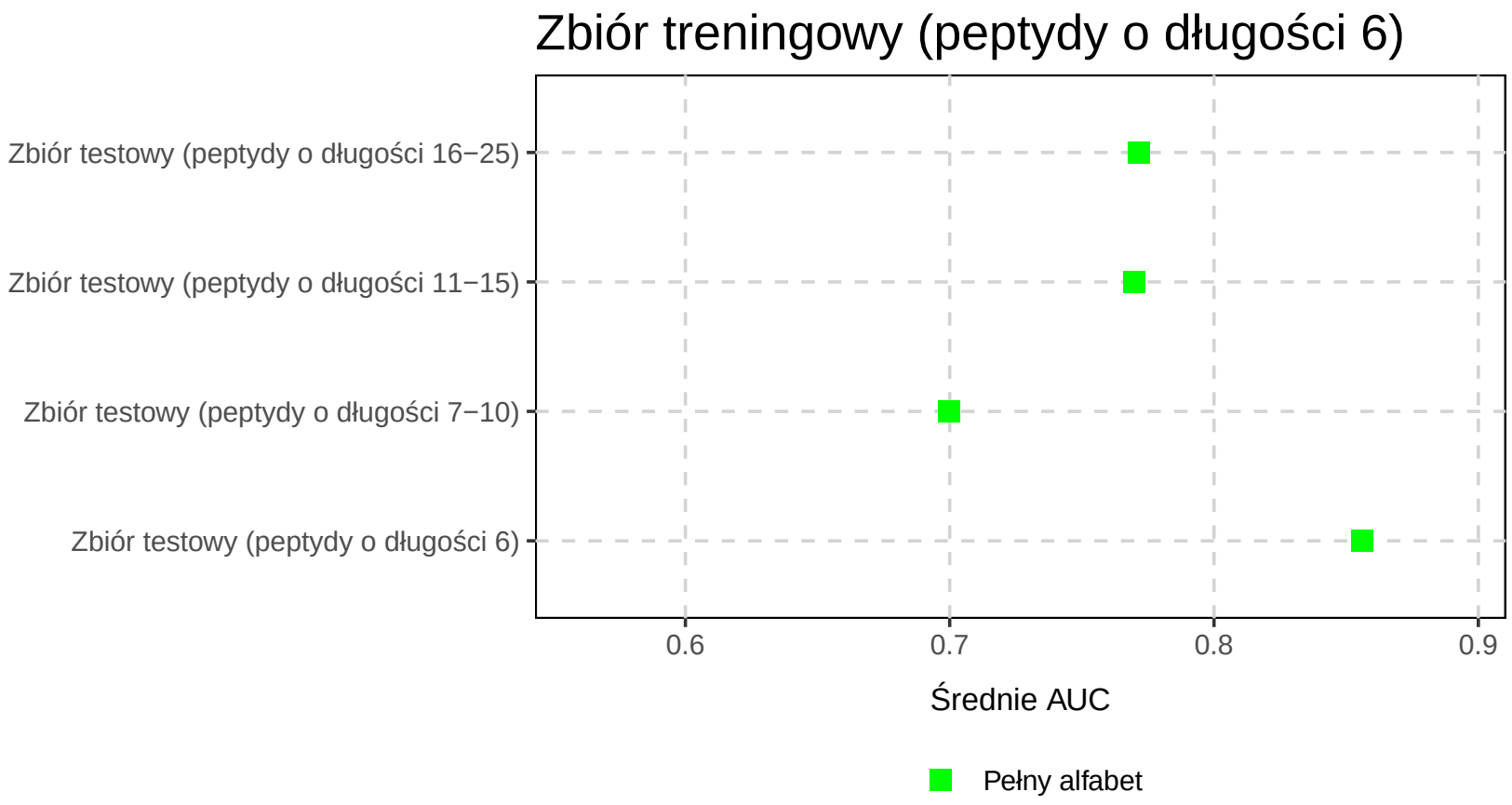
Sawaya et al. (2007)

AmyloGram

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


FILIQA
LHYSGD
KRQDKG
EKPTRR
F FI F L...
L LH L Y...
K KR K Q...
E EK E P...
F FI F L...
L LH L Y...
K KR K Q...
E EK E P...

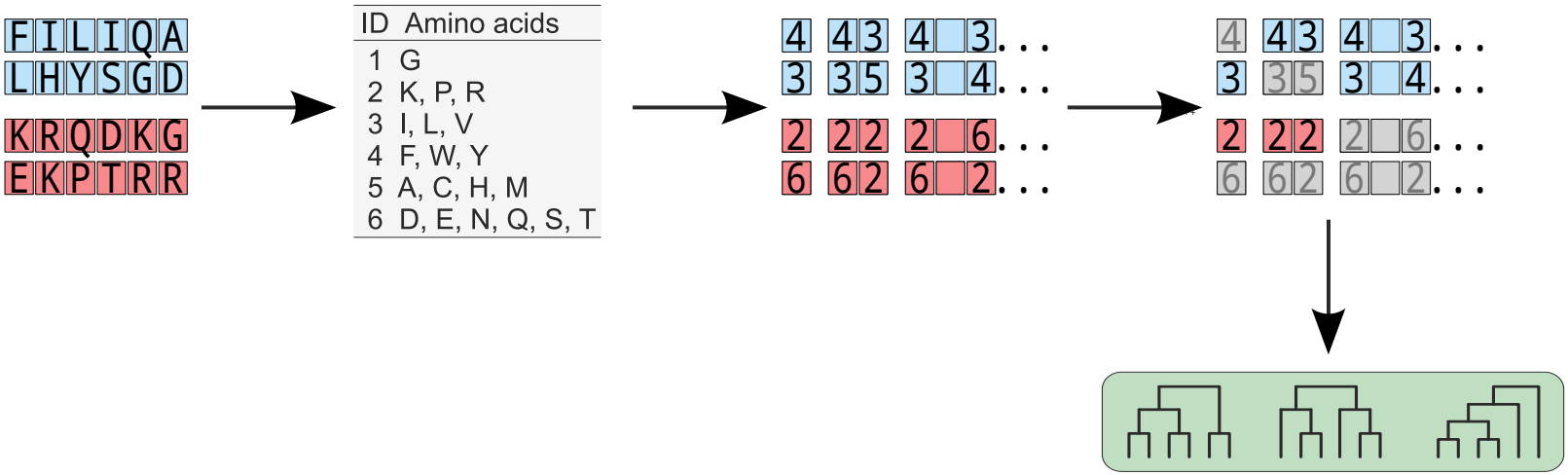
Walidacja krzyżowa



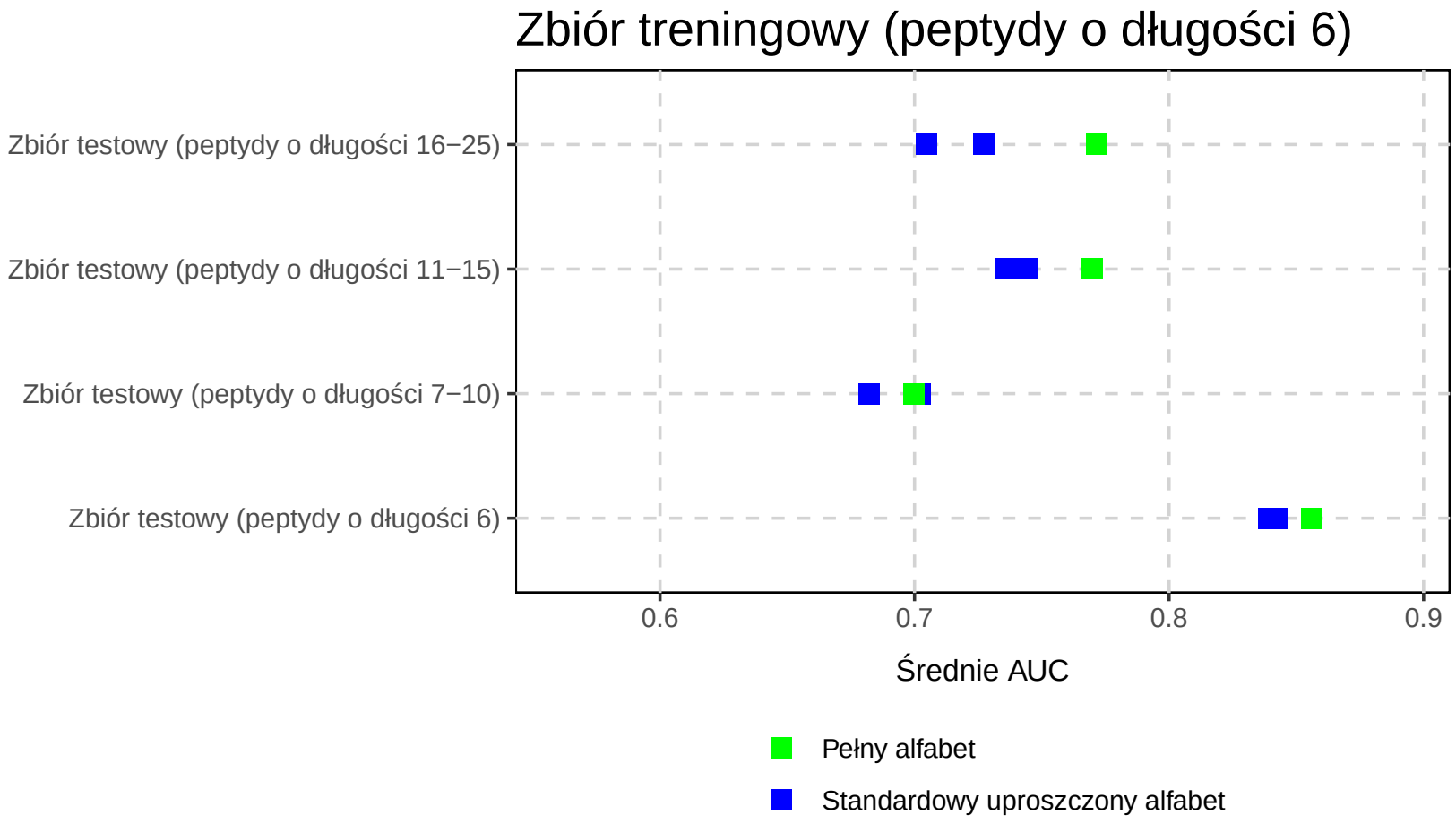
Standardowe uproszczone alfabety

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

Standardowe uproszczone alfabety



Walidacja krzyżowa

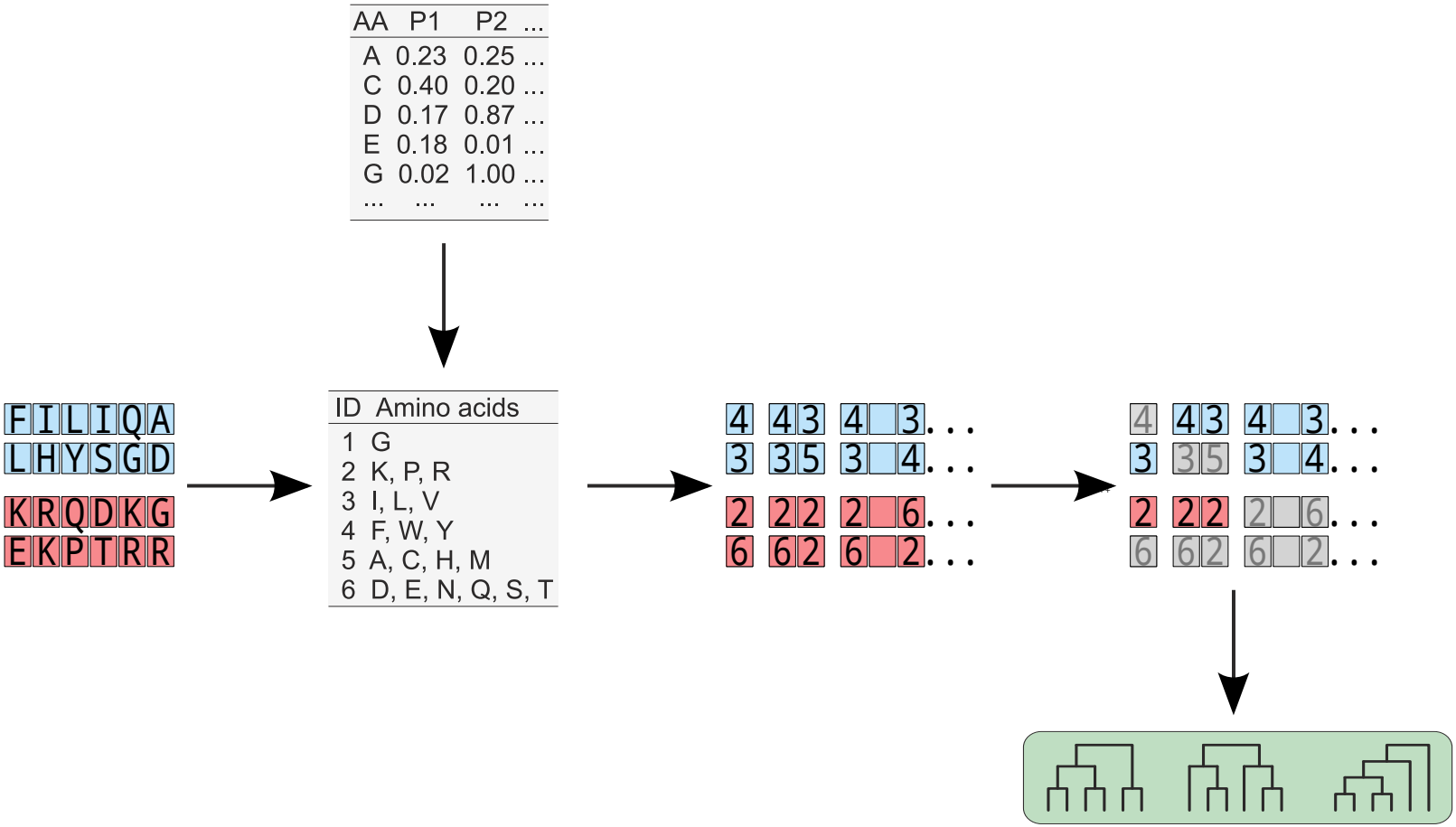


Standardowe alfabety aminokwasowe nie poprawiaja jakości predykcji amyloidów.

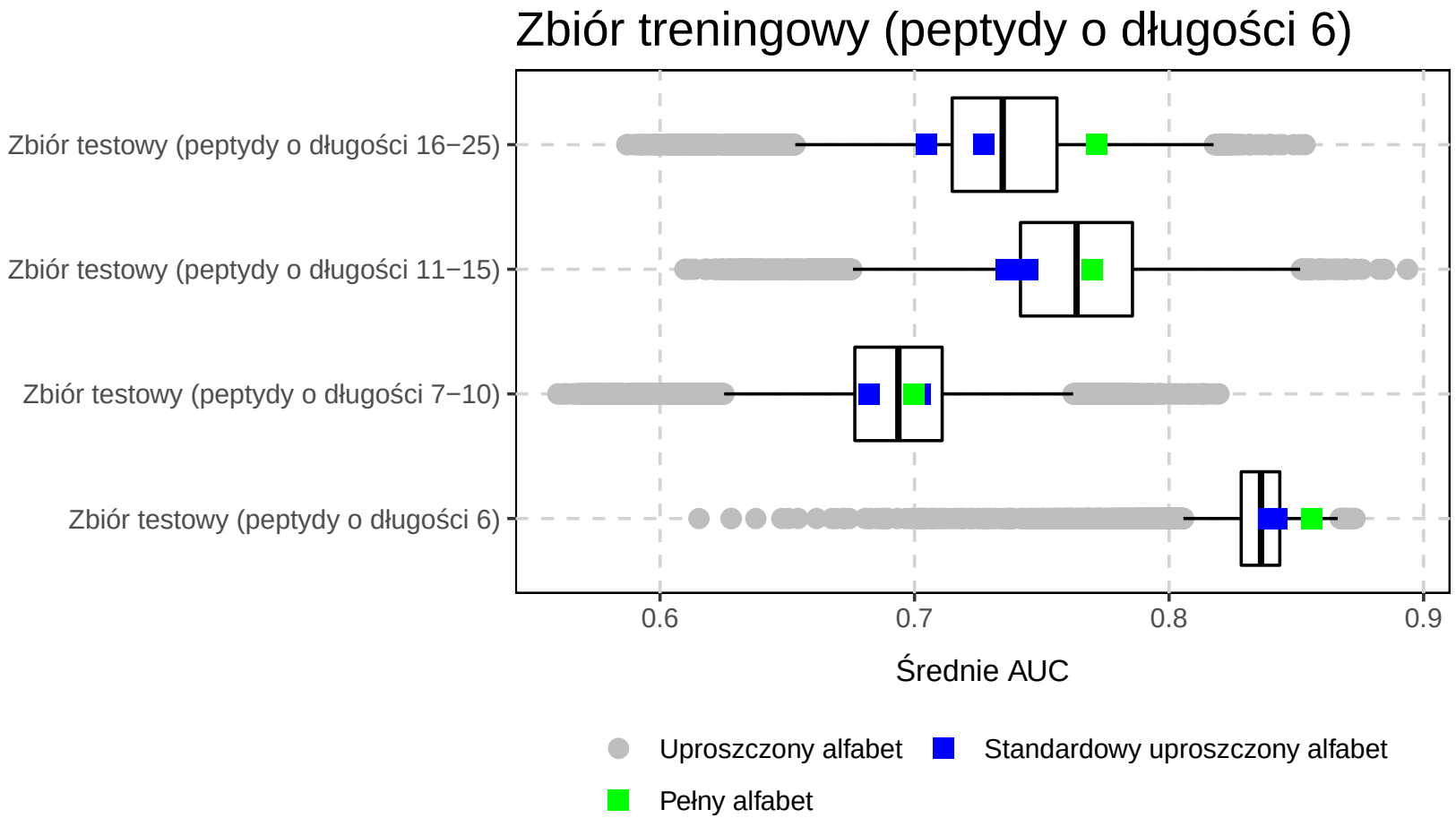
Nowe uproszczone alfabety

- 17 miar wybranych z bazy AAIndex:
  - size of residues,
  - hydrophobicity,
  - solvent surface area,
  - frequency in β-sheets,
  - contactivity.
- 524 284 amino acid simplified alphabets with different level of amino acid alphabet reduction (three to six amino acid groups).

Novel simplified amino acid alphabets

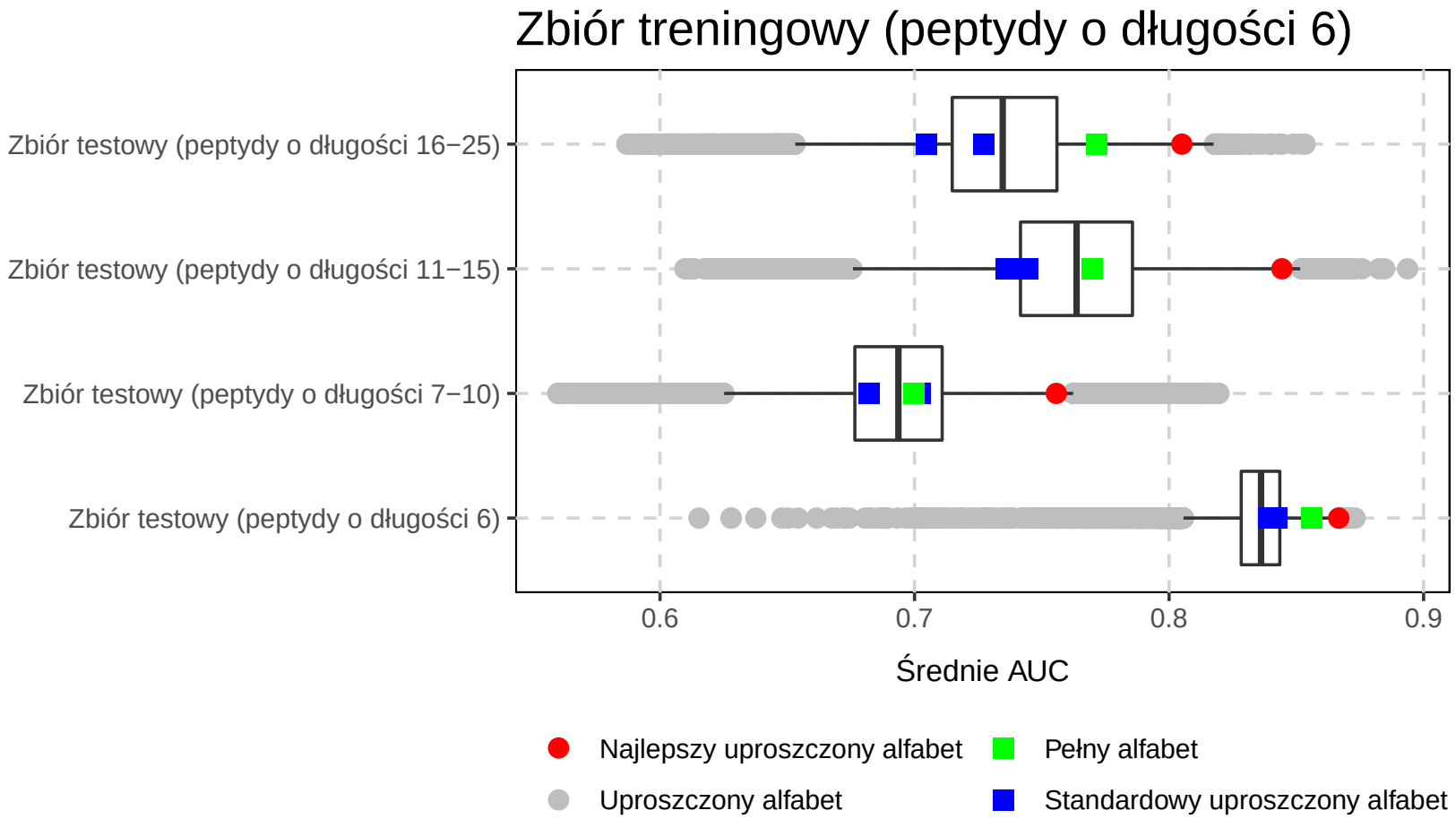


Walidacja krzyżowa



Hinges of boxes correspond to the 0.25 and 0.75 quartiles. The bar inside the box represents the median. The gray circles correspond to the simplified alphabets with the AUC outside the 0.95 confidence interval.

The best-performing simplified alphabet



The best-performing simplified alphabet

Subgroup 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

The best-performing simplified alphabet

Subgroup 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

Group 3 and 4 - hydrophobic amino acids.

The best-performing simplified alphabet

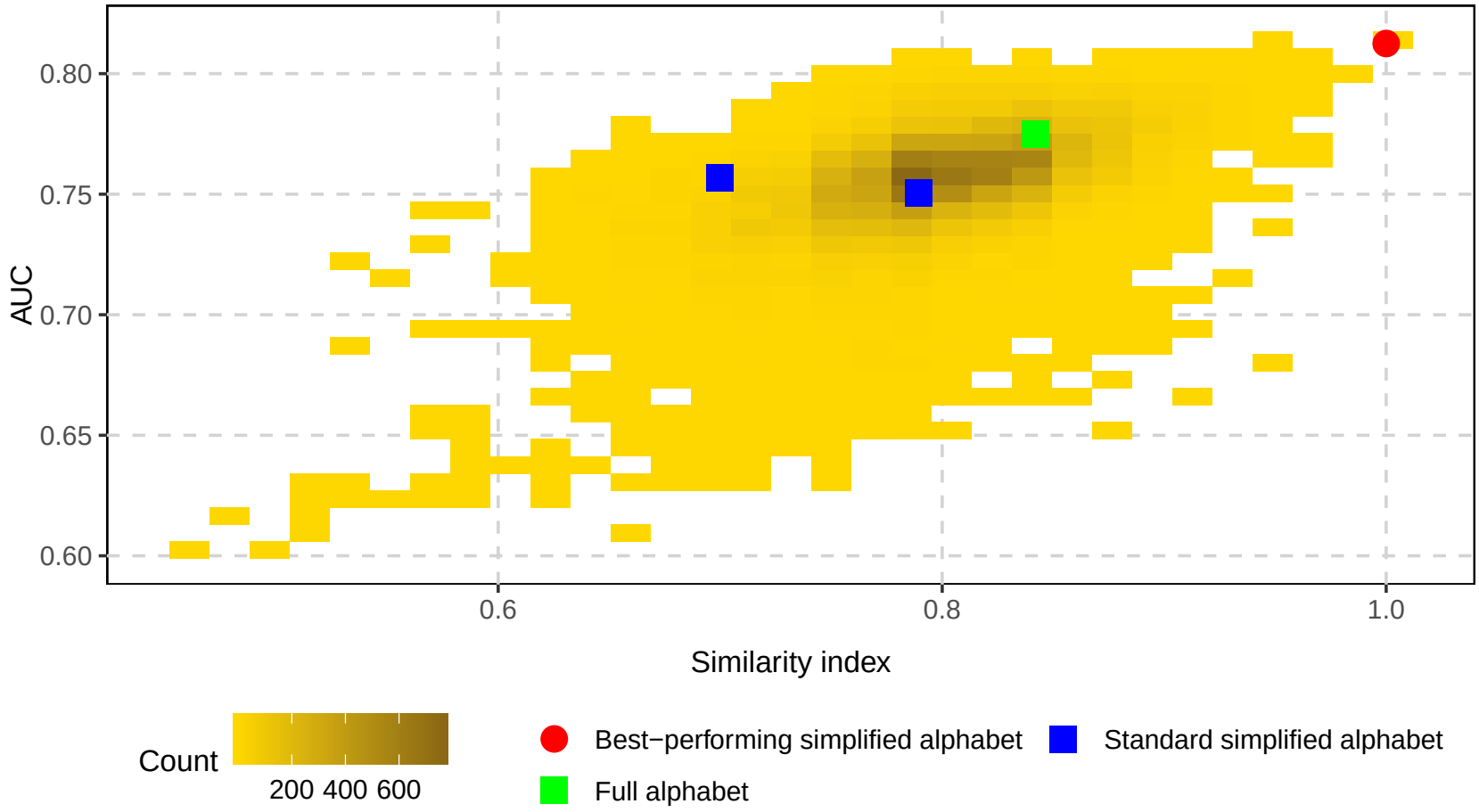
Subgroup ID	Amino acids
1	G
2	K, P, R
3	I, L, V
4	F, W, Y
5	A, C, H, M
6	D, E, N, Q, S, T

Group 2 - charged breakers of β-structures.

Alphabet similarity and performance

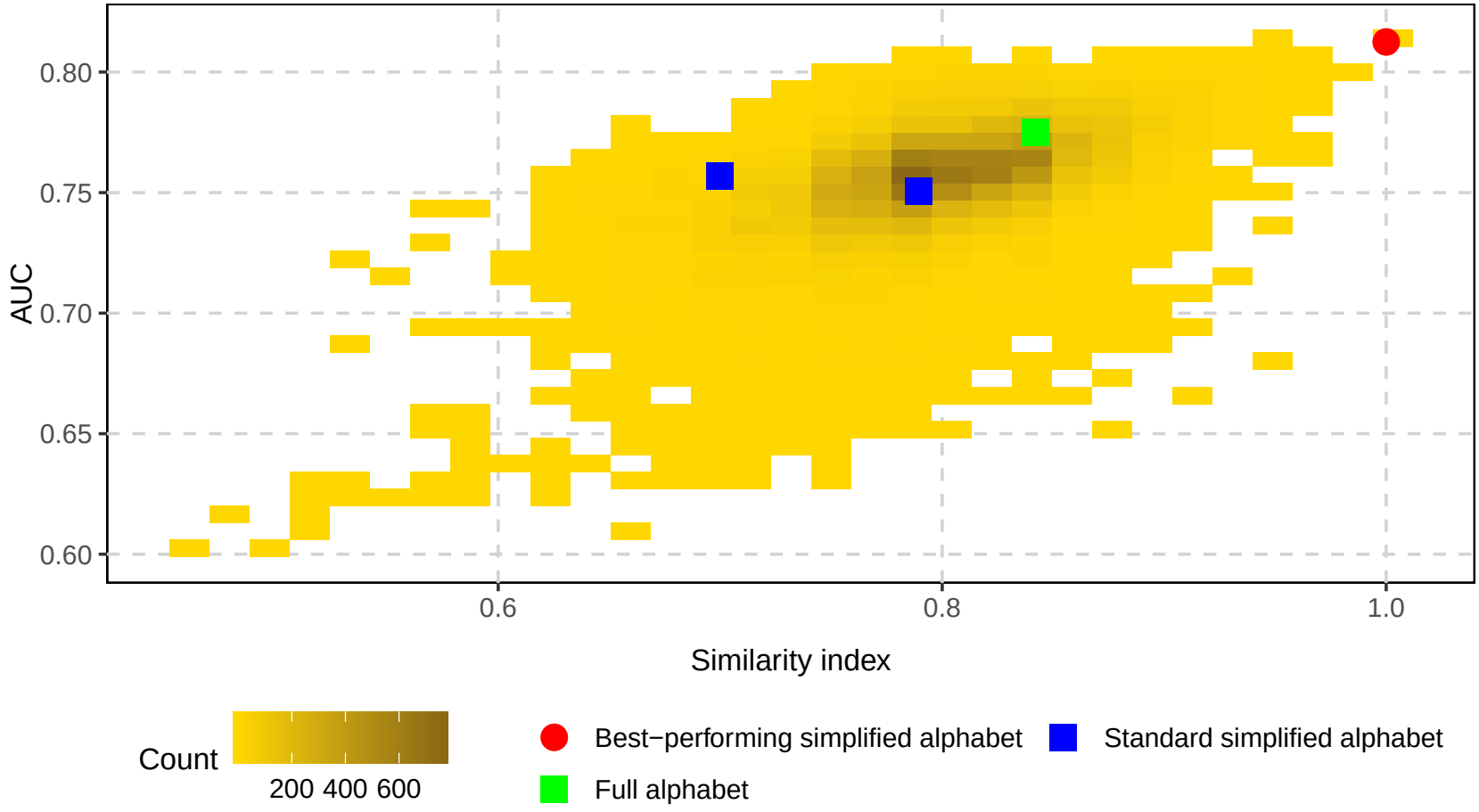
Is the best-performing simplified amino alphabet associated with amyloidogenicity?

Similarity index



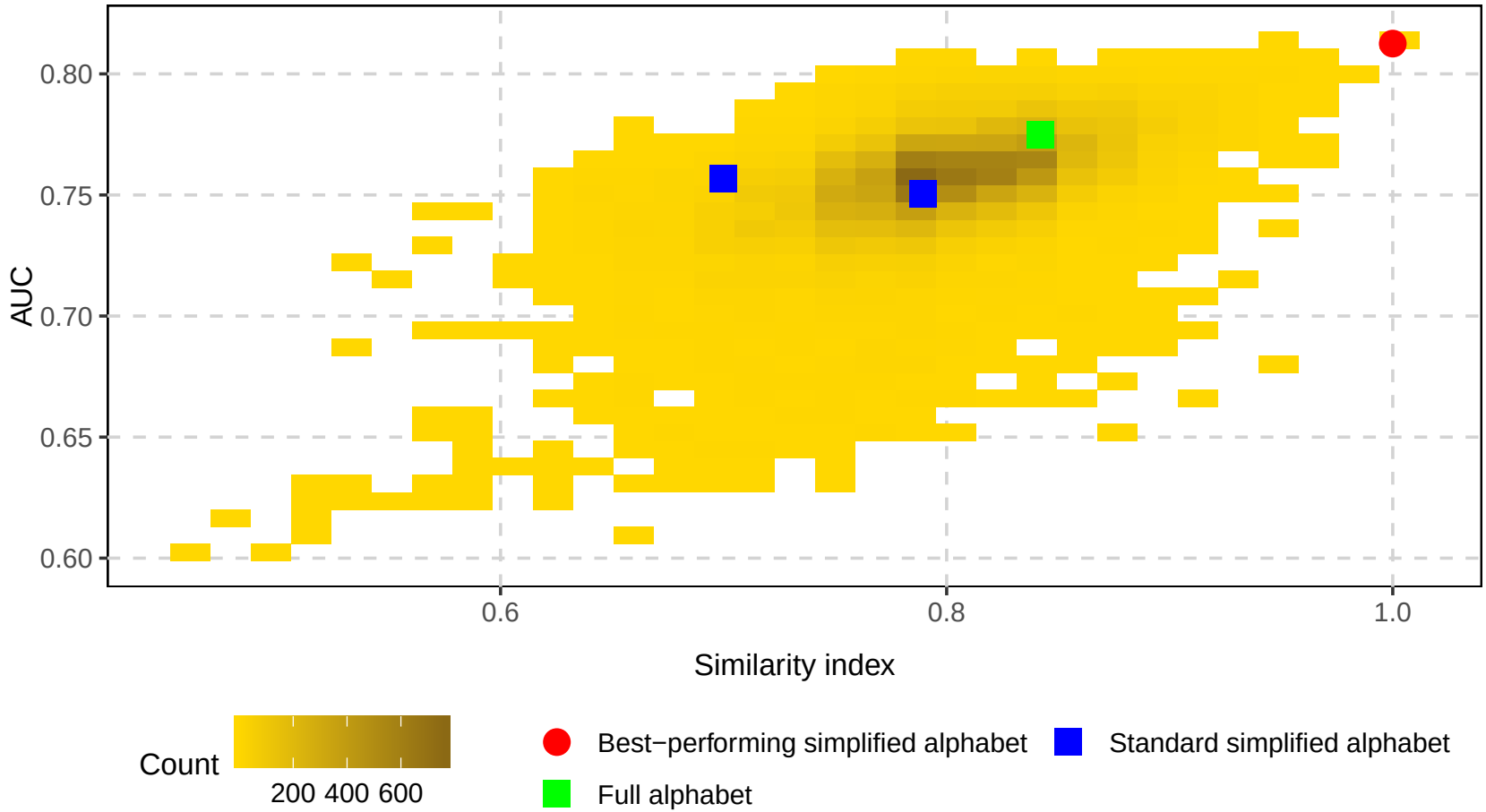
Similarity index (Stephenson and Freeland, 2013) measures the similarity between two simplified alphabets (1 - identical, 0, totally dissimilar).

Similarity index



The color of a square is proportional to the number of simplified alphabets in its area.

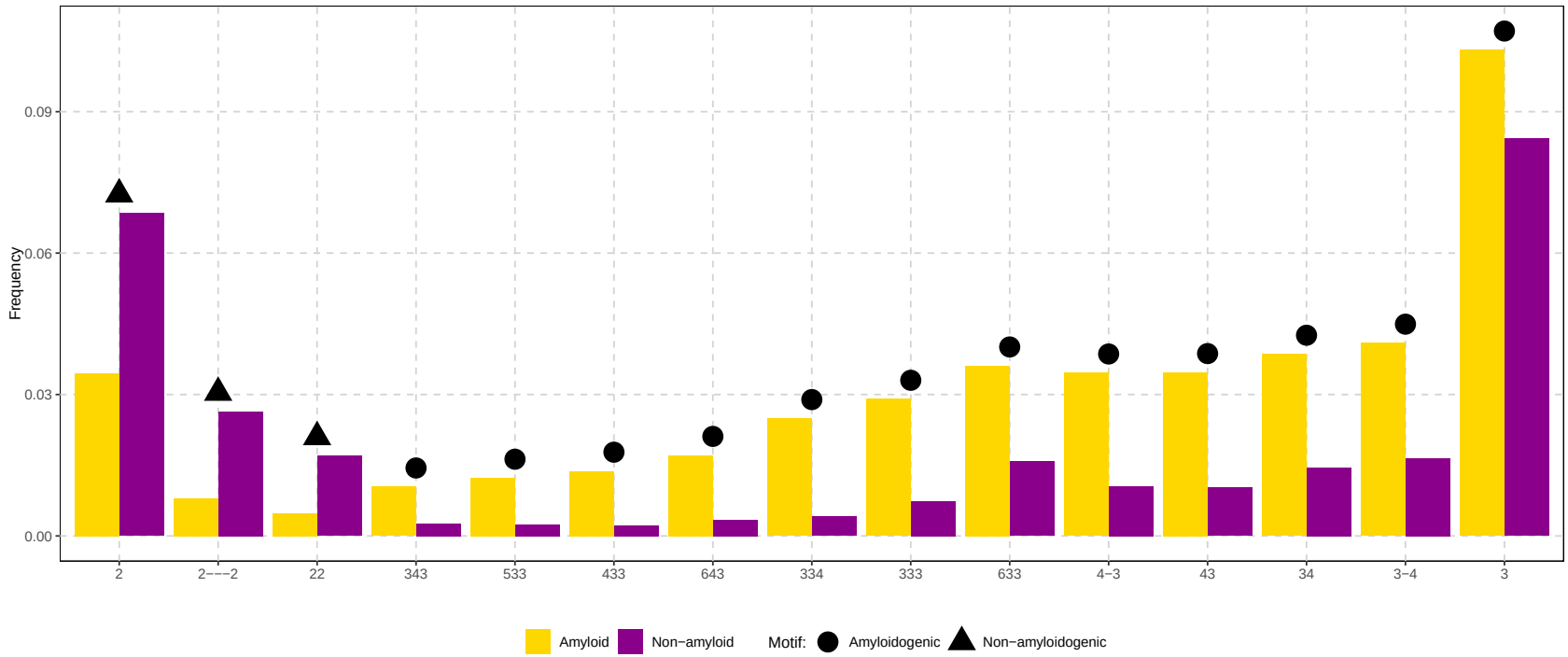
Similarity index



The correlation between mean AUC an similarity index is significant ($\text{p-value} \leq 2.2^{-16}$; $\rho = 0.51$).

Are informative n-grams found by QuiPT associated with amyloidogenicity?

Informative n-grams



Out of 65 the most informative n-grams, 15 (23%) were also found in the motifs validated experimentally (Paz and Serrano, 2004).

Benchmark results

Classifier	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

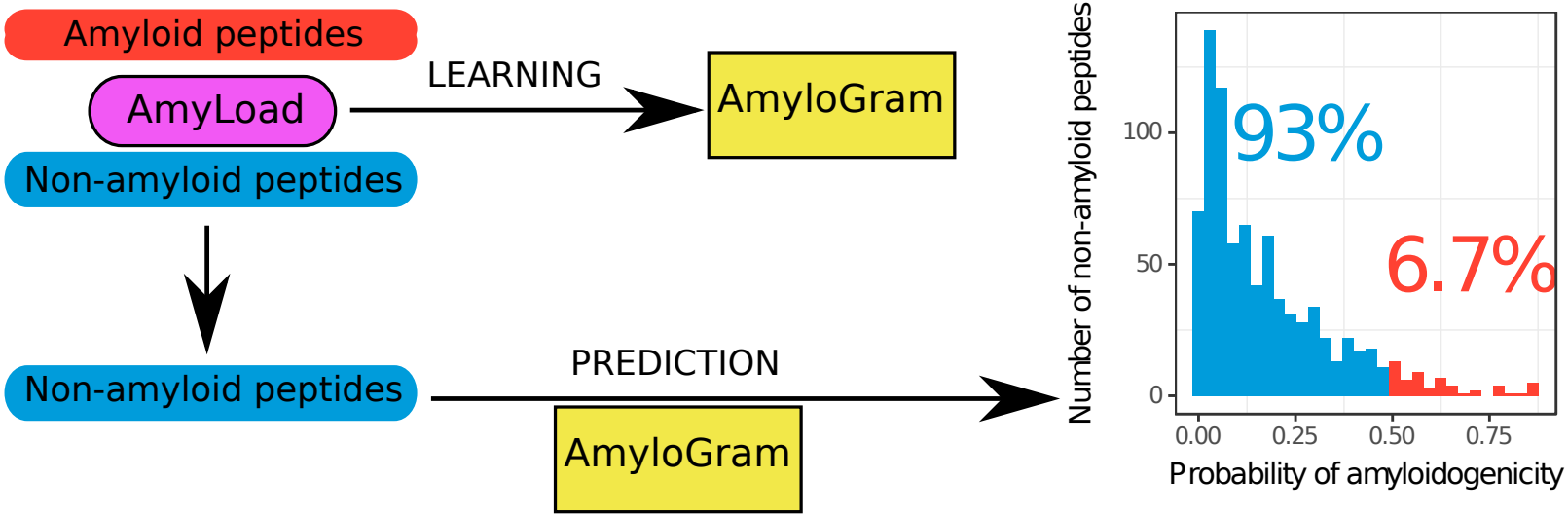
The predictor based on the best-performing alphabet, called AmyloGram, was benchmarked against the most popular tools for the detection of amyloid peptides using an external data set *pep424*.

Benchmark results

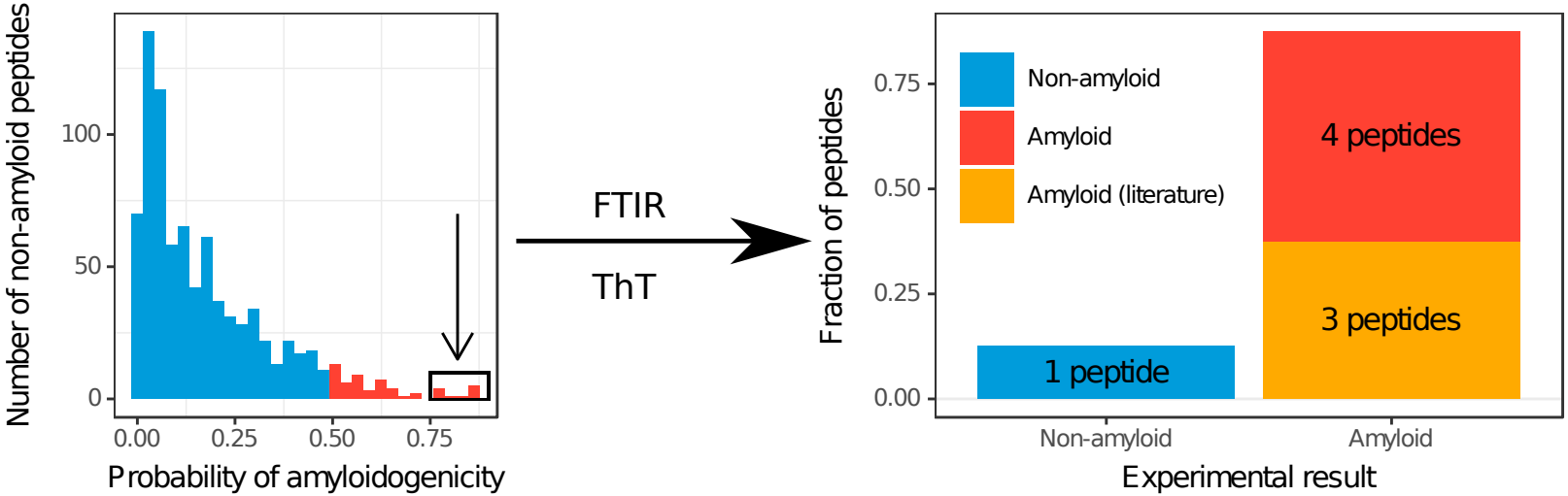
Classifier	AUC	MCC
AmyloGram	**0.8972**	**0.6307**
PASTA 2.0 (Walsh et al., 2014)	0.8550	0.4291
FoldAmyloid (Garbuzynskiy et al., 2010)	0.7351	0.4526
APPNN (Família et al., 2015)	0.8343	0.5823

MCC (Matthew's Correlation Coefficient) measures the performance of a classifier (1 - classifier always properly recognizes amyloid proteins, -1 - classifier never properly recognizes amyloid proteins).

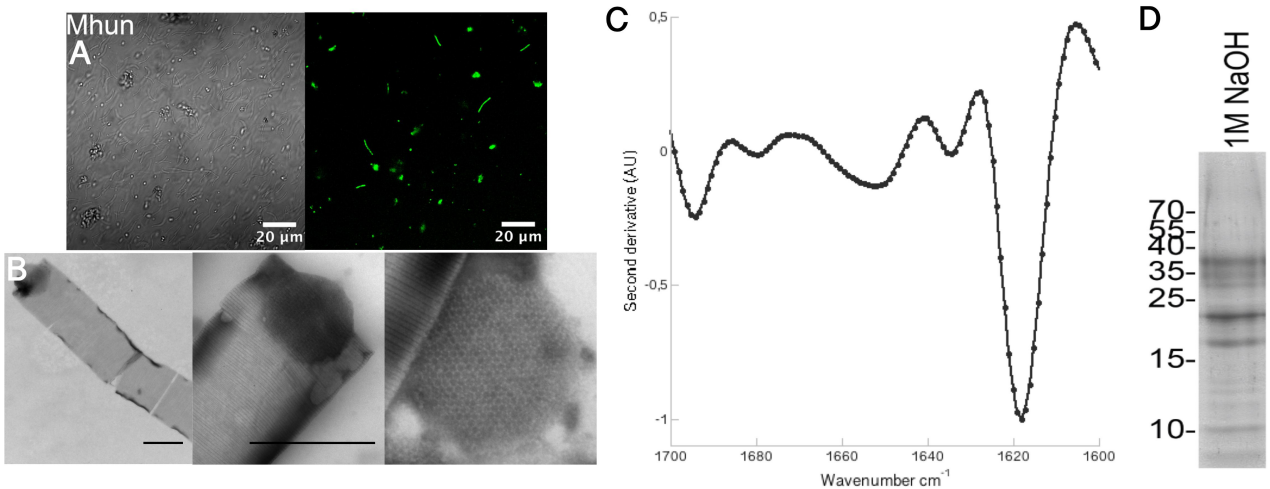
Experimental validation



Experimental validation

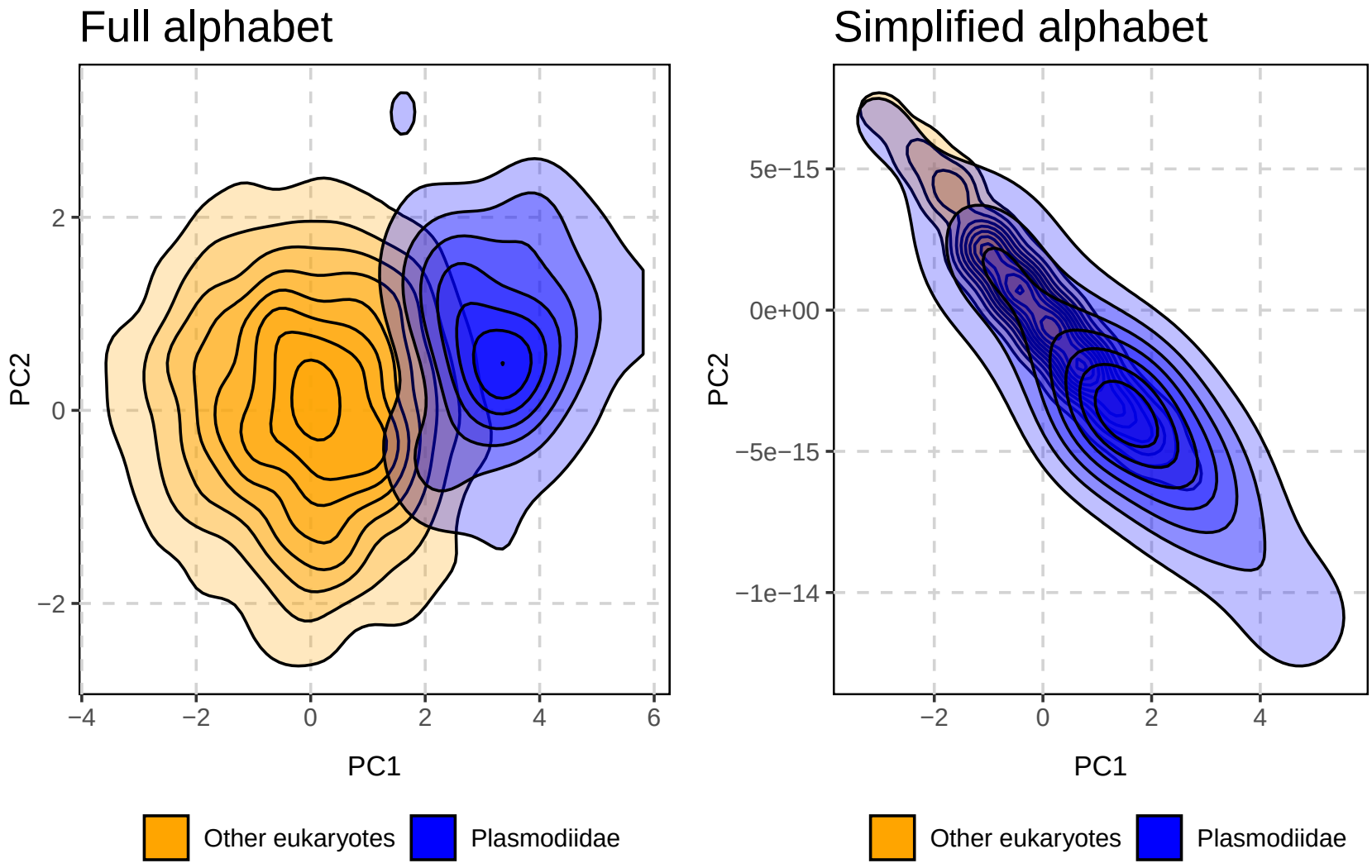


Novel amyloid protein



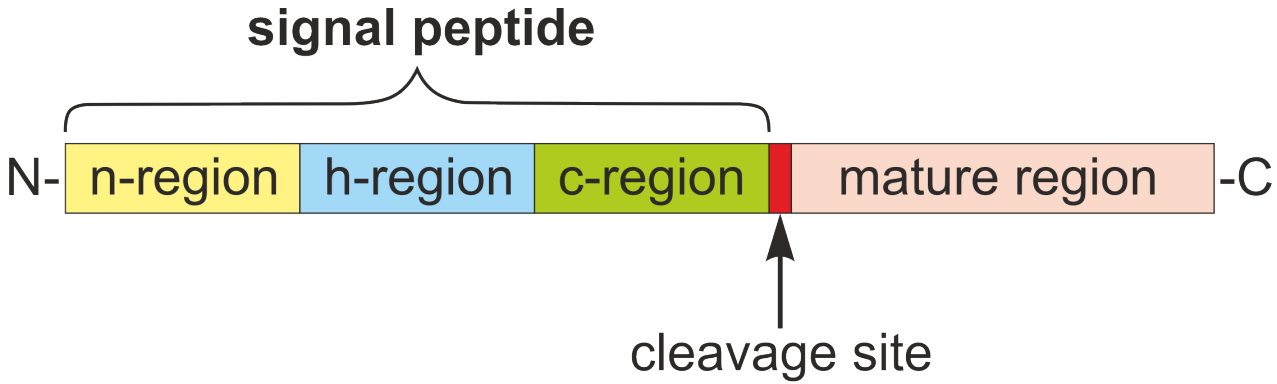
Methanospirillum sp. (Christensen et al., 2018)

Signal peptide prediction



PCA of amino acid frequency in signal peptides.

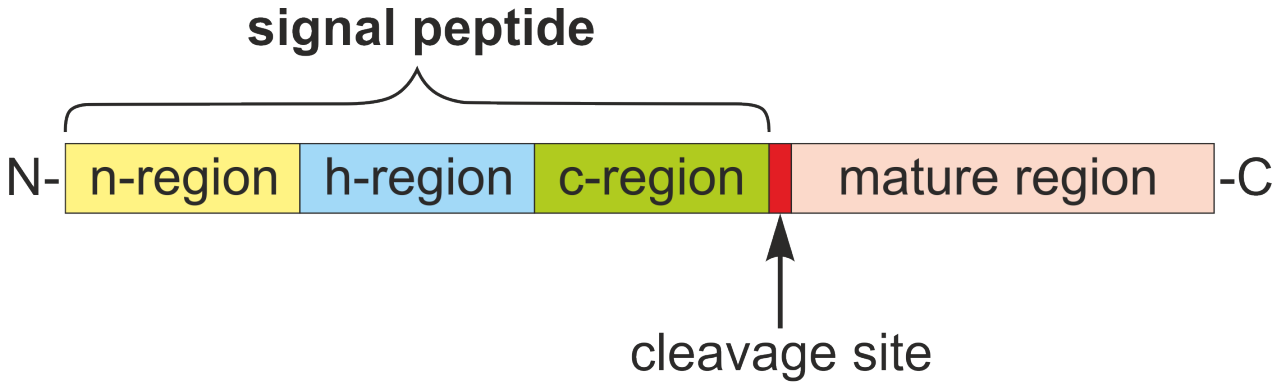
Signal peptides



Signal peptides:

- are short (20-30 residues) N-terminal amino acid sequences forming α-helices,

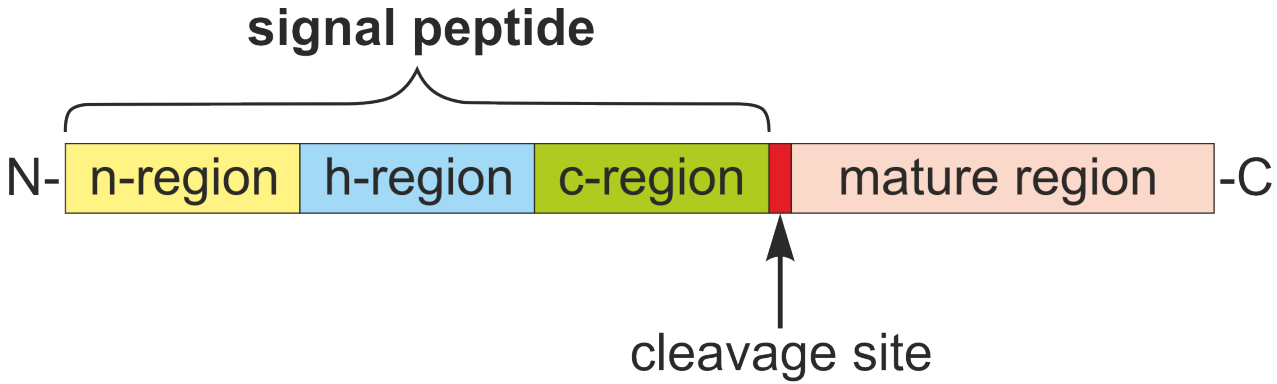
Signal peptides



Signal peptides:

- are short (20-30 residues) N-terminal amino acid sequences forming α-helices,
- direct proteins to the endomembrane system and next to extra- or intracellular localizations,

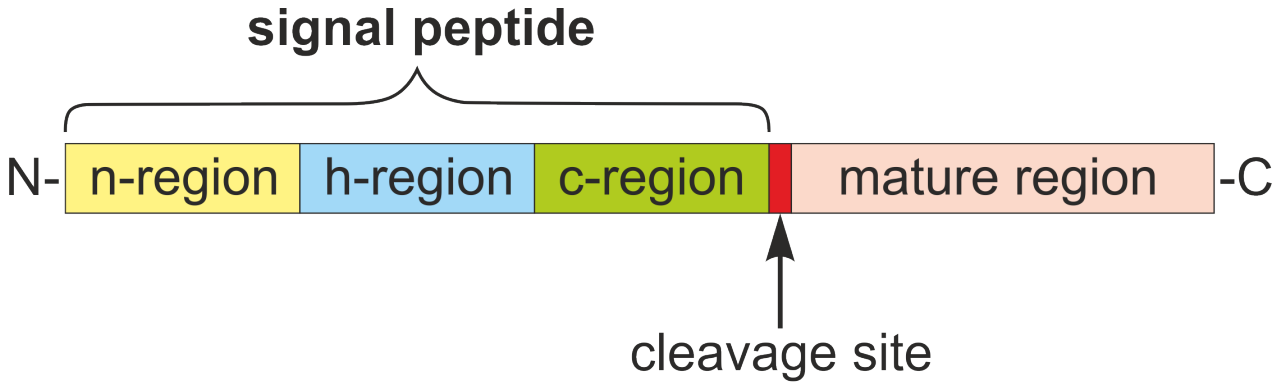
Signal peptides



Signal peptides:

- are short (20-30 residues) N-terminal amino acid sequences forming α-helices,
- direct proteins to the endomembrane system and next to extra- or intracellular localizations,
- are universal enough to direct properly proteins in different secretory systems.

Signal peptides

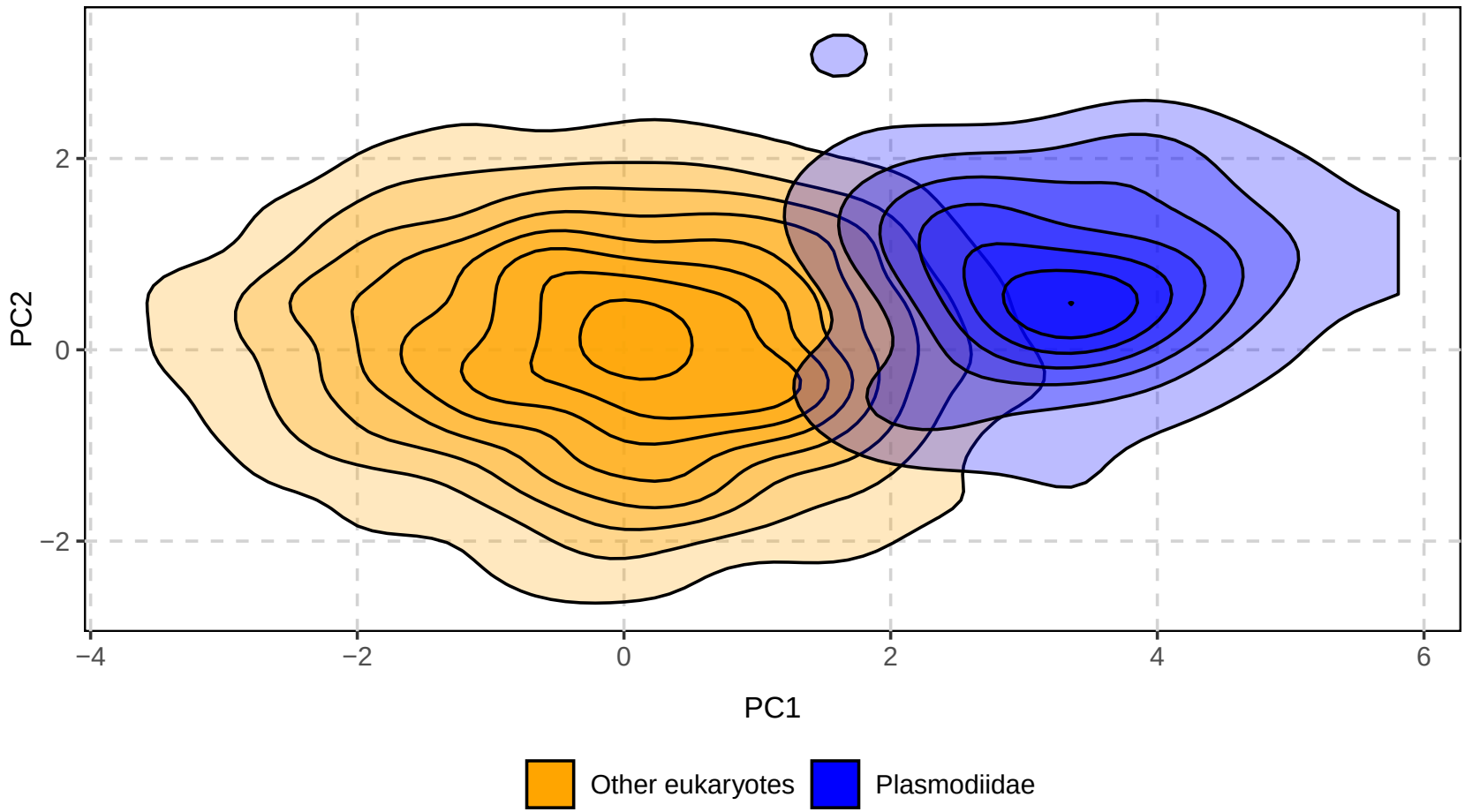


Signal peptides possess three distinct domains with variable length and characteristic amino acid composition (Hegde and Bernstein, 2006):

- n-region: mostly basic residues (Nielsen and Krogh, 1998),
- h-region: strongly hydrophobic residues (Nielsen and Krogh, 1998),
- c-region: a few polar, uncharged residues.

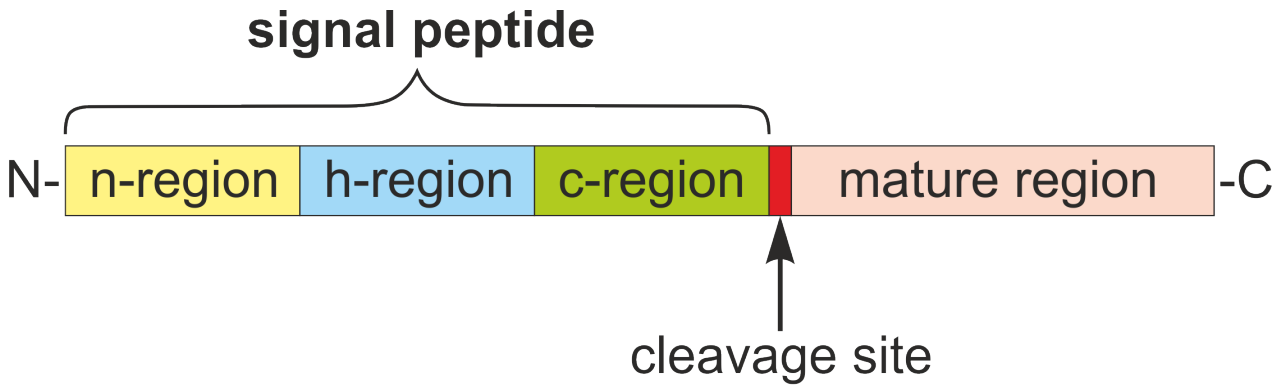
Signal peptides

Amino acid composition of signal peptides differ between Plasmodium sp. and other eukaryotes. Therefore, predictors of signal peptides do not detect malarial signal peptides accurately.



PCA of amino acid frequency in signal peptides.

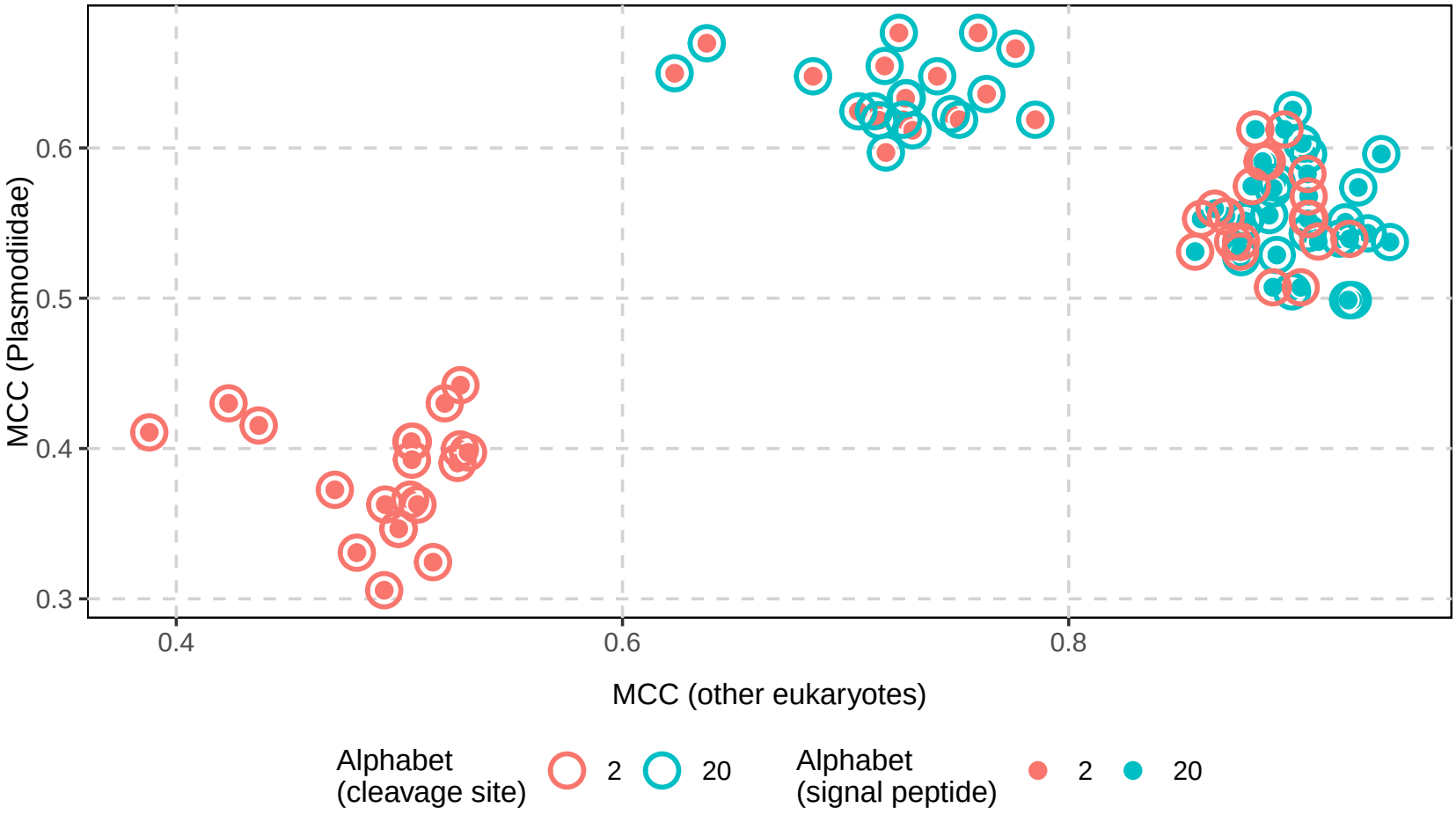
Signal peptide prediction



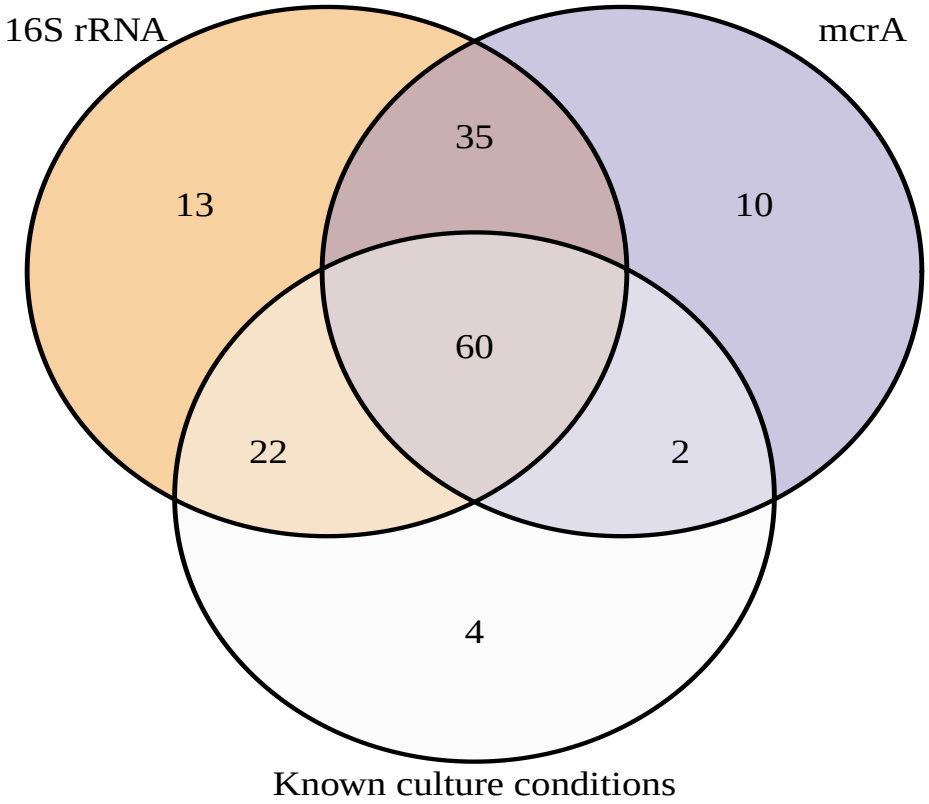
SignalP 4.1 (Petersen et al., 2011) combines output of two separate predictors:

- cleavage site,
- signal peptide.

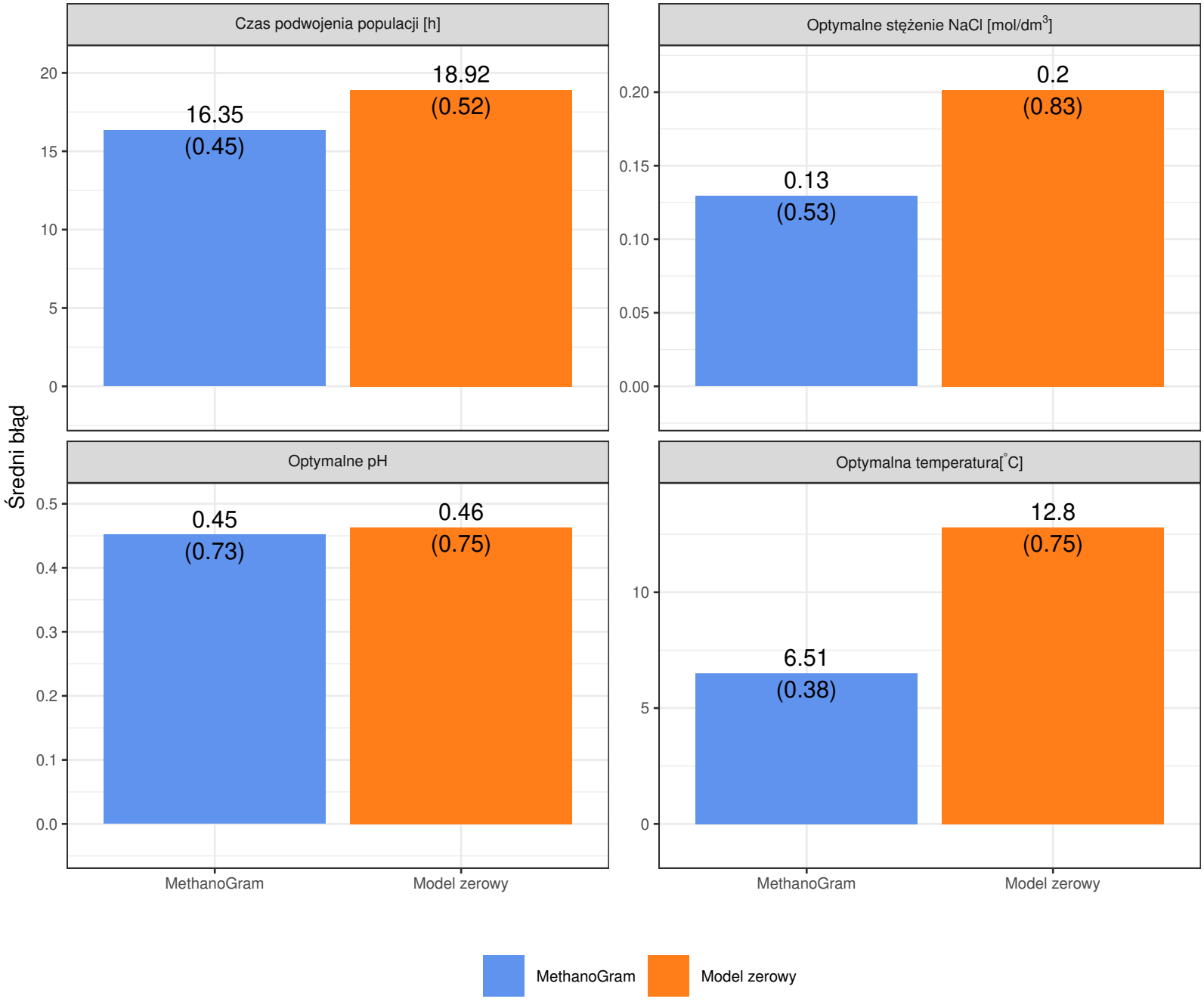
Signal peptide prediction



Prediction of culturing conditions



`metanogen.biotech.uni.wroc.pl` (Jabłoński et al., 2015)


Czas podwojenia populacji [h]
Optymalne stężenie NaCl [mol/dm^3]
Optymalne pH
Optymalna temperatura[$^\circ$C]
Średni błąd
16.35
(0.45)
18.92
(0.52)
0.13
(0.53)
0.2
(0.83)
0.45
(0.73)
0.46
(0.75)
6.51
(0.38)
12.8
(0.75)
MethanoGram
Model zerowy
MethanoGram
Model zerowy

Summary

1. Created algorithms effectively filtering n-grams.
2. Introduced new methods for search of simplified amino acids.
3. Implemented novel algorithms in the **R** package *biogram*.
4. Applied the n-gram analysis framework to:
   - prediction of amyloids (AmyloGram),
   - prediction of atypical signal peptides,
   - prediction of culture conditions of methanogenes (MethanoGram).

Summary

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`.

Podziekowania

Mentorzy:

- **Paweł Mackiewicz (University of Wrocław)**.
- Lars Kaderali (University of Greifswald).
- Małgorzata Kotulska (Wrocław University of Science and Technology).
- Marcin Łukaszewicz (University of Wrocław).
- Henrik Nielsen (Technical University of Denmark).
- Stefan Rödiger (Brandenburg University of Technology Cottbus-Senftenberg).
- Andreas Weinhäusel (Austrian Institute of Technology).

Acknowledgments

Peers:

- Agata Błaszczyńska (Wrocław University of Science and Technology).
- Anna Duda-Madej (Wrocław Medical University).
- Przemysław Gagat (University of Wrocław).
- Marlena Gasior-Głogowska (Wrocław University of Science and Technology).
- Sławomir Jabłoński (University of Wrocław).
- Rafał Kolenda (Sanger Institute).
- Chris Lauber (Technical University Dresden).
- Natalia Niedzielska (Wrocław University of Science and Technology).
- Piotr Sobczyk (Wrocław University of Science and Technology).

Acknowledgments

Funding:

- National Science Center (Preludium and Etiuda).
- COST ACTION CA15110 (Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research).
- KNOW Wrocław Center for Biotechnology.

Publications I

1. 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 [liczba cytacji: 0].
2. Mackiewicz D., Posacki P., Burdukiewicz M., Błażej P. *Role of recombination and faithfulness to partner in sex chromosome degeneration*. **Scientific Reports**, 2018 [liczba cytacji: 0].
3. 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 [liczba cytacji: 0].
4. Burdukiewicz M., Sobczyk P. Rödiger S., Duda-Madej A., Mackiewicz P., Kotulska M., *Amyloidogenic motifs revealed by n-gram analysis*. **Scientific Reports**, 2017 [liczba cytacji: 2].

Publications II

5. 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 [liczba cytacji: 2].

6. Rödiger S., Burdukiewicz M., Spiess A.-N., Blagodatskikh K., *Enabling reproducible real-time quantitative PCR research: the RDML package*. **Bioinformatics**, 2017 [liczba cytacji: 0].

7. Burdukiewicz M., Rödiger S., Sobczyk P., Menschikowski M., Schierack P., Mackiewicz P., *Methods for comparing multiple digital PCR experiments*, **Biomolecular Detection and Quantification**, 2016 [liczba cytacji: 2].

8. 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 [liczba cytacji: 4].

Publications III

9. 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 [liczba cytacji: 34].

10. Rödiger S., Burdukiewicz M., Schierack P., *chipPCR: an R Package to Pre-Process Raw Data of Amplification Curves*. **Bioinformatics**, 2015 [liczba cytacji: 12].

11. 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 [liczba cytacji: 14].

12. 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 [liczba cytacji: 13].

Summary

1. Created a new algorithm for effective filtering of n-grams.
2. Introduced new methods for search of simplified amino acids.
3. Implemented novel algorithms in the **R** package *AmyloGram*.
4. Applied the n-gram analysis framework to:
   - prediction of amyloids (AmyloGram),
   - prediction of atypical signal peptides,
   - prediction of culture conditions of methanogenes (MethanoGram).

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.

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

Jabłoński, S., Rodowicz, P., and Łukaszewicz, M. (2015). Methanogenic archaea database containing physiological and biochemical characteristics. *Int J Syst Evol Microbiol*, 65(4):1360–1368.

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.

References III

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.

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.

References IV

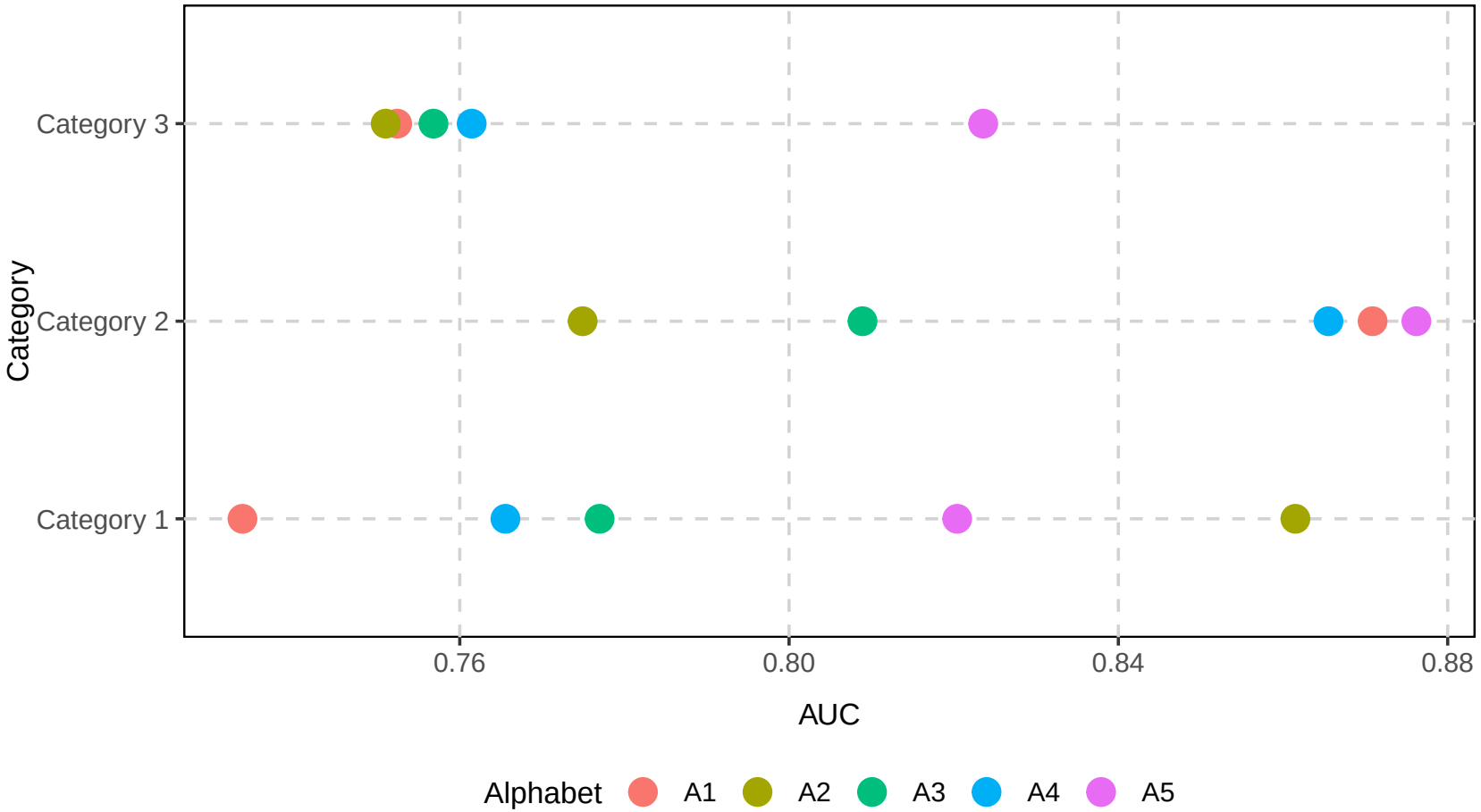
Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8(10):785–786.

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.

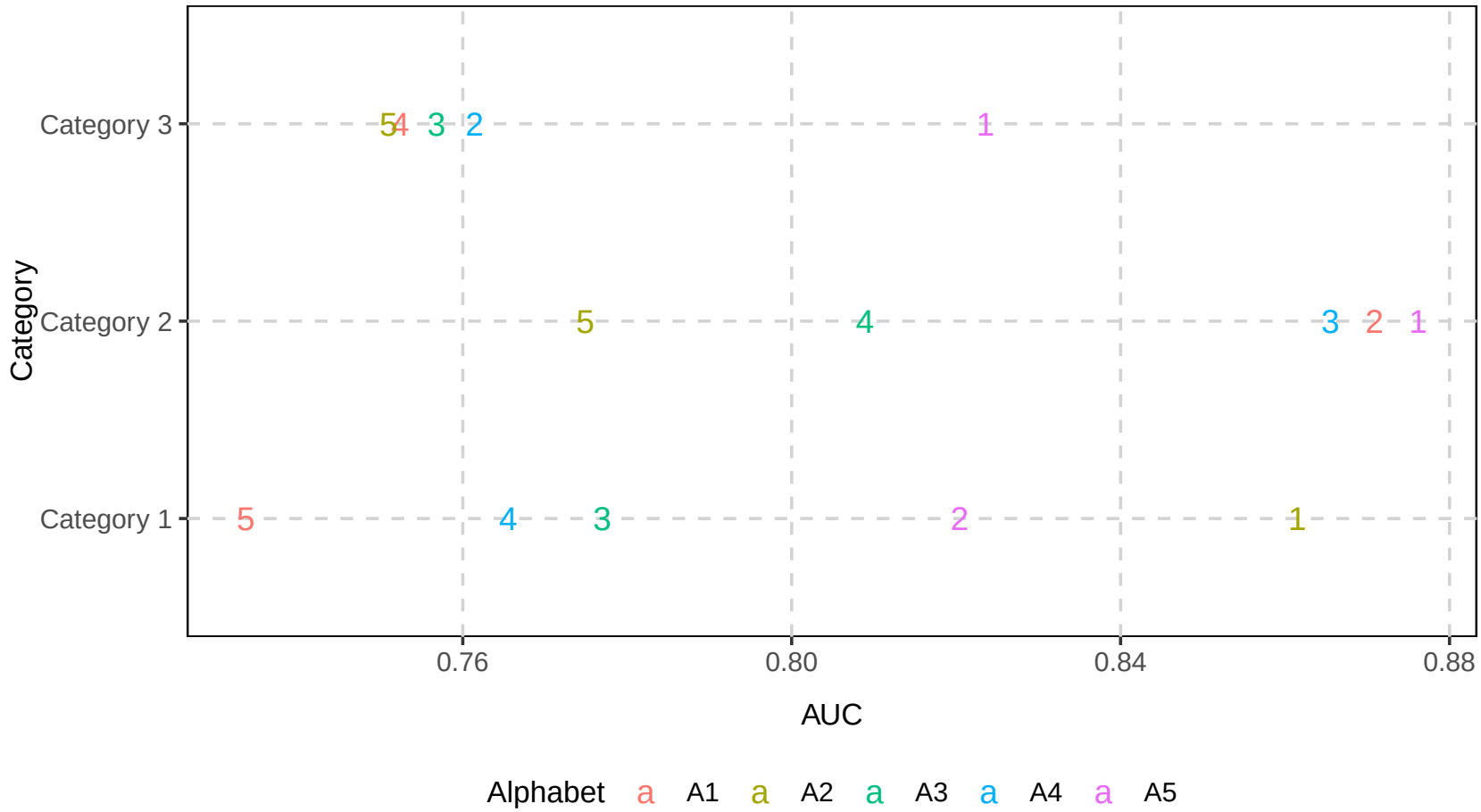
Stephenson, J. D. and Freeland, S. J. (2013). Unearthing the root of amino acid similarity. *Journal of Molecular Evolution*, 77(4):159–169.

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.

Ranking alphabets

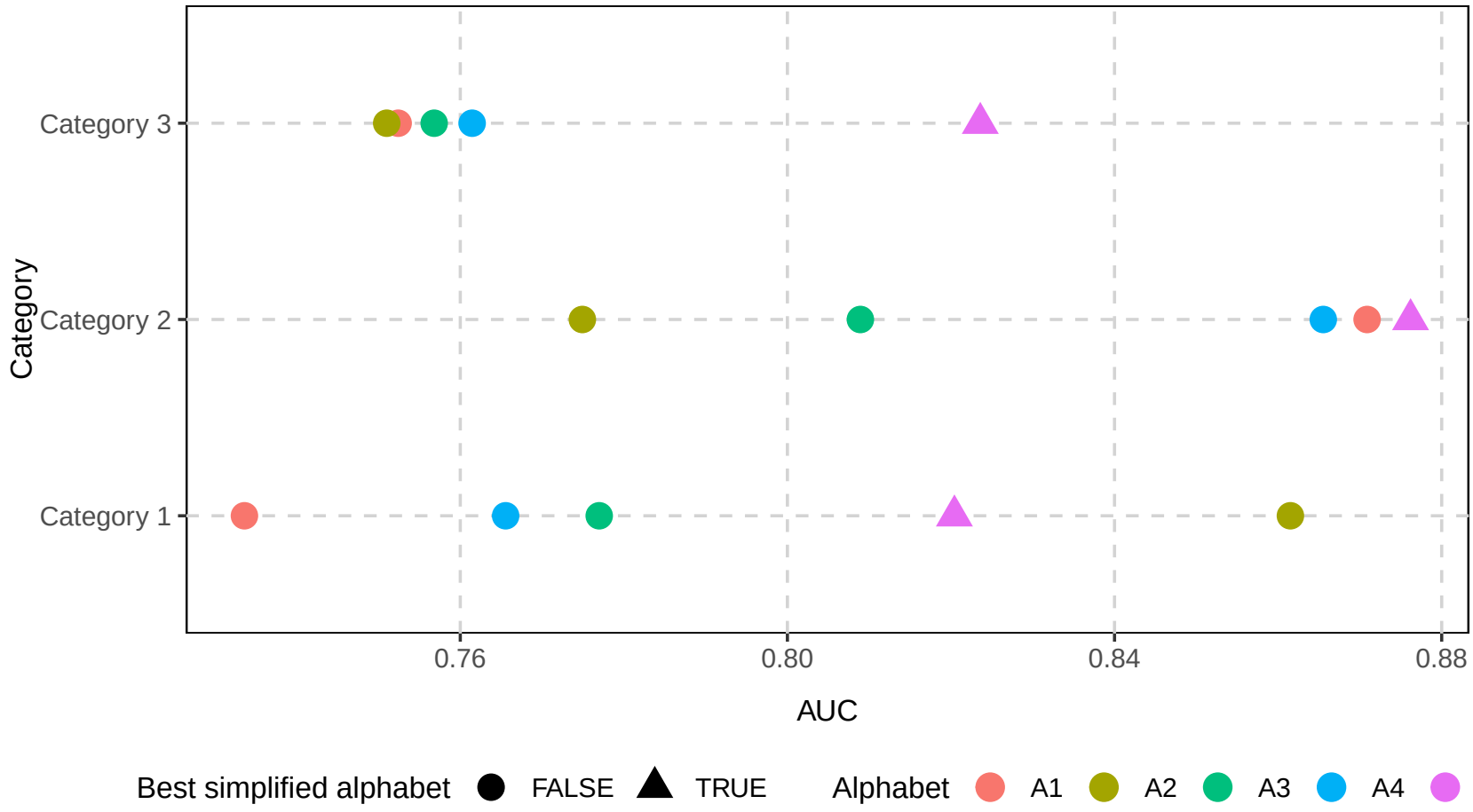


Ranking alphabets



We rank alphabets separately in all length categories assuming the rank 1 for the best AUC, rank 2 for the second best AUC and so on.

Ranking alphabets



The best-performing alphabet has the lowest sum of ranks.