Predicting properties of biological sequences using n-gram analysis

Michał Burdukiewicz

Department of Genomics, University of Wrocław

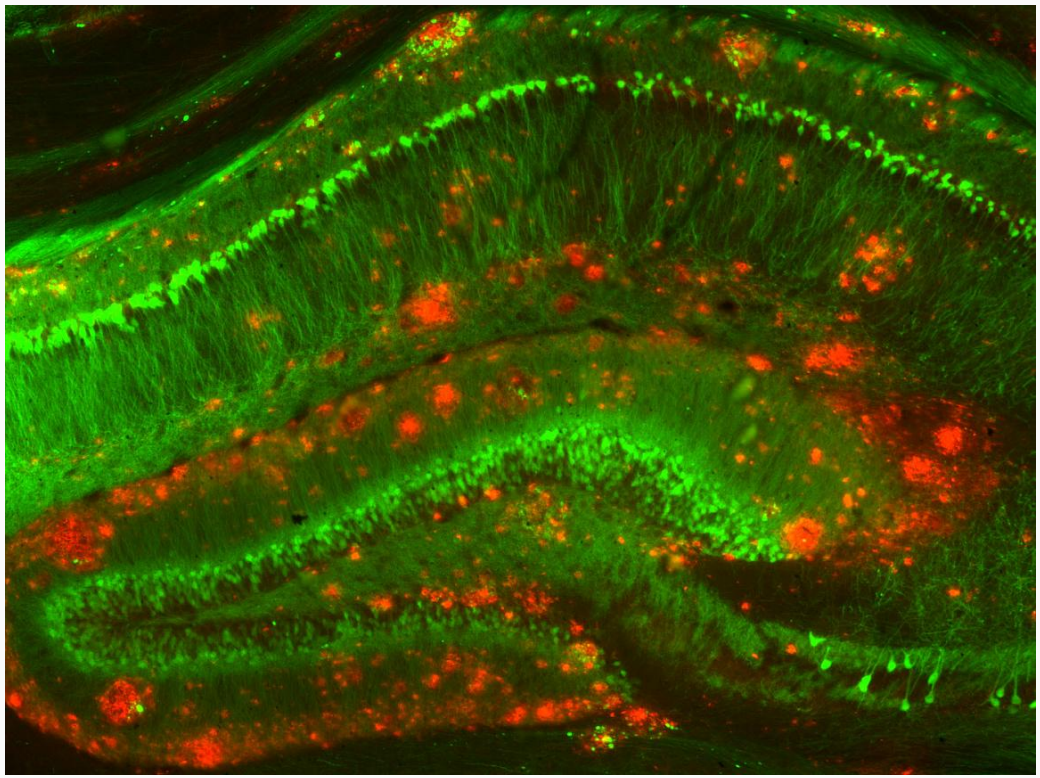
In silico research allows scientists to more efficiently design experimentl studies.

Examples:

- prediction of protein properties (presence of signal peptides, amyloidogenicity),
- predicting culture conditions.

Amyloid proteins

Amyloid are proteins associated with various diseases (e.g., Alzheimer's, Creutzfeldt-Jakob's and Huntington's diseases) which are able to form harmful aggregates.

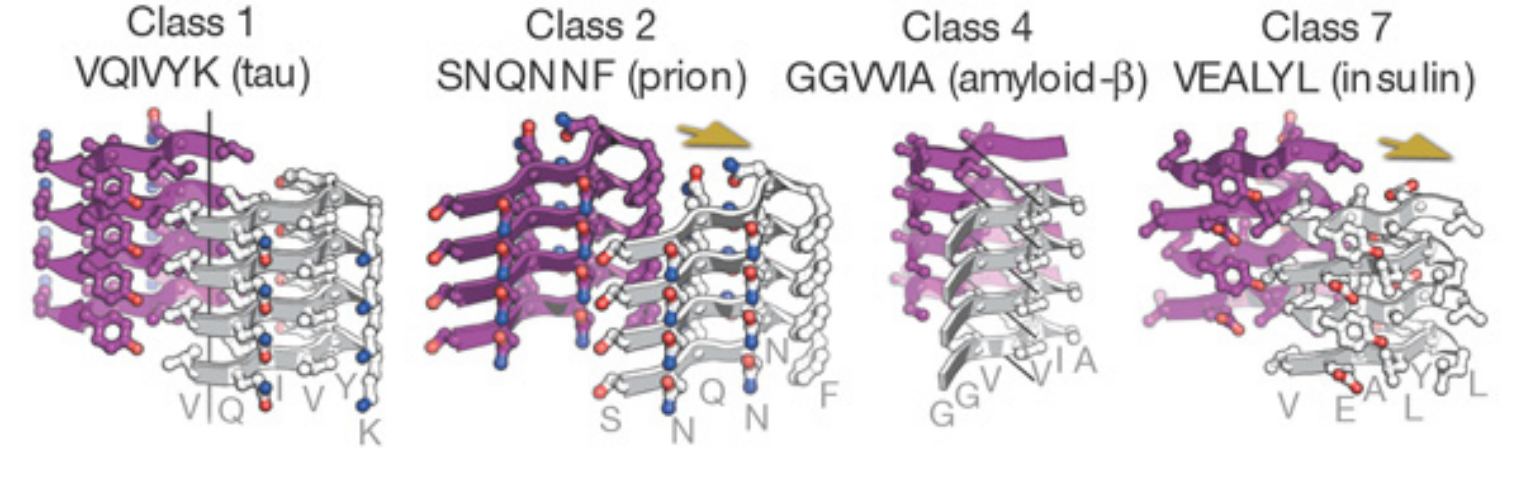


Amyloid aggregates (red) around neurons (green). Strittmatter Laboratory, Yale University.

Amyloid proteins

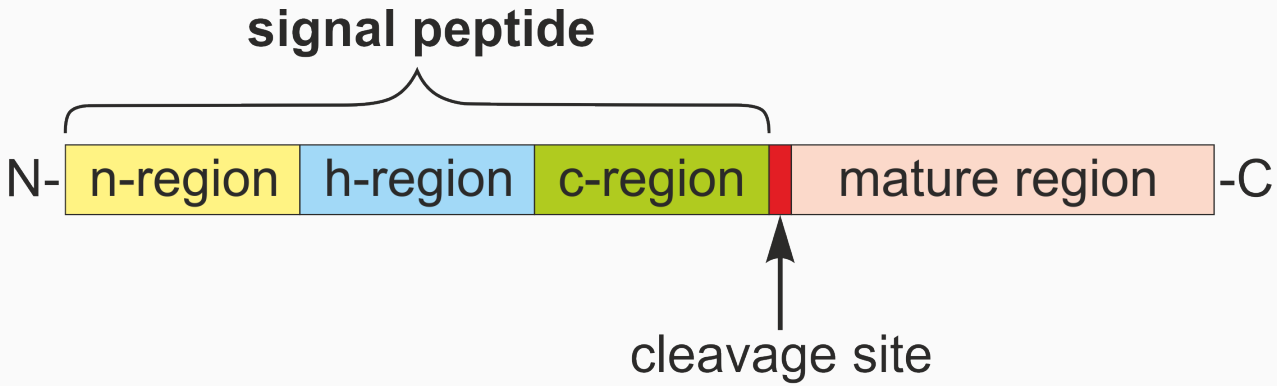
Hot-spots:

- short (6-15 amino acids),
- very high variability of amino acid composition,
- initiate amyloid aggregation,
- create specific "zipper-like" β-structures.



Sawaya et al. (2007)

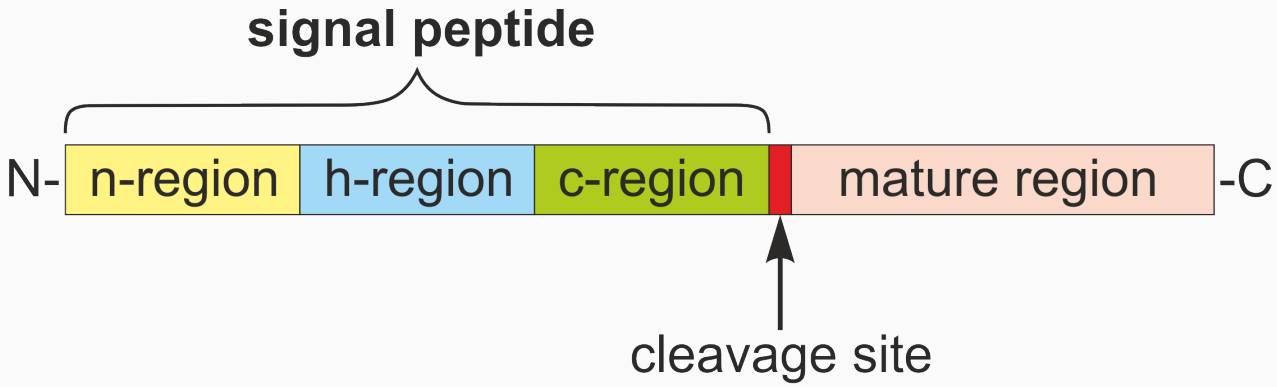
Signal peptides



Signal peptides:

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

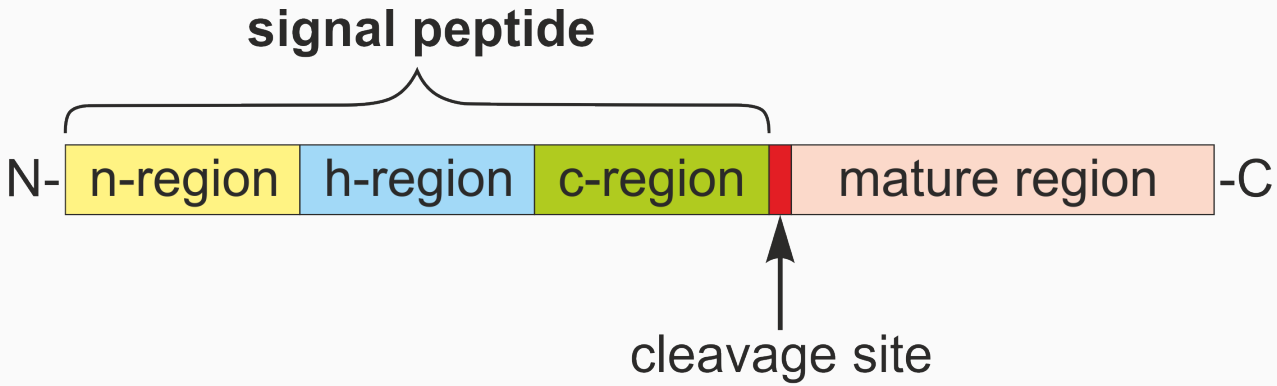
Signal peptides



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,

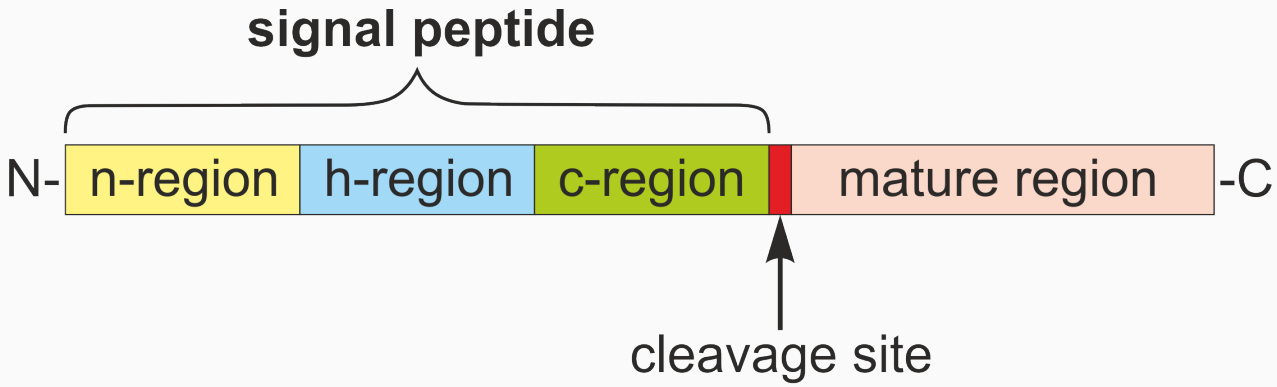
Signal peptides



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

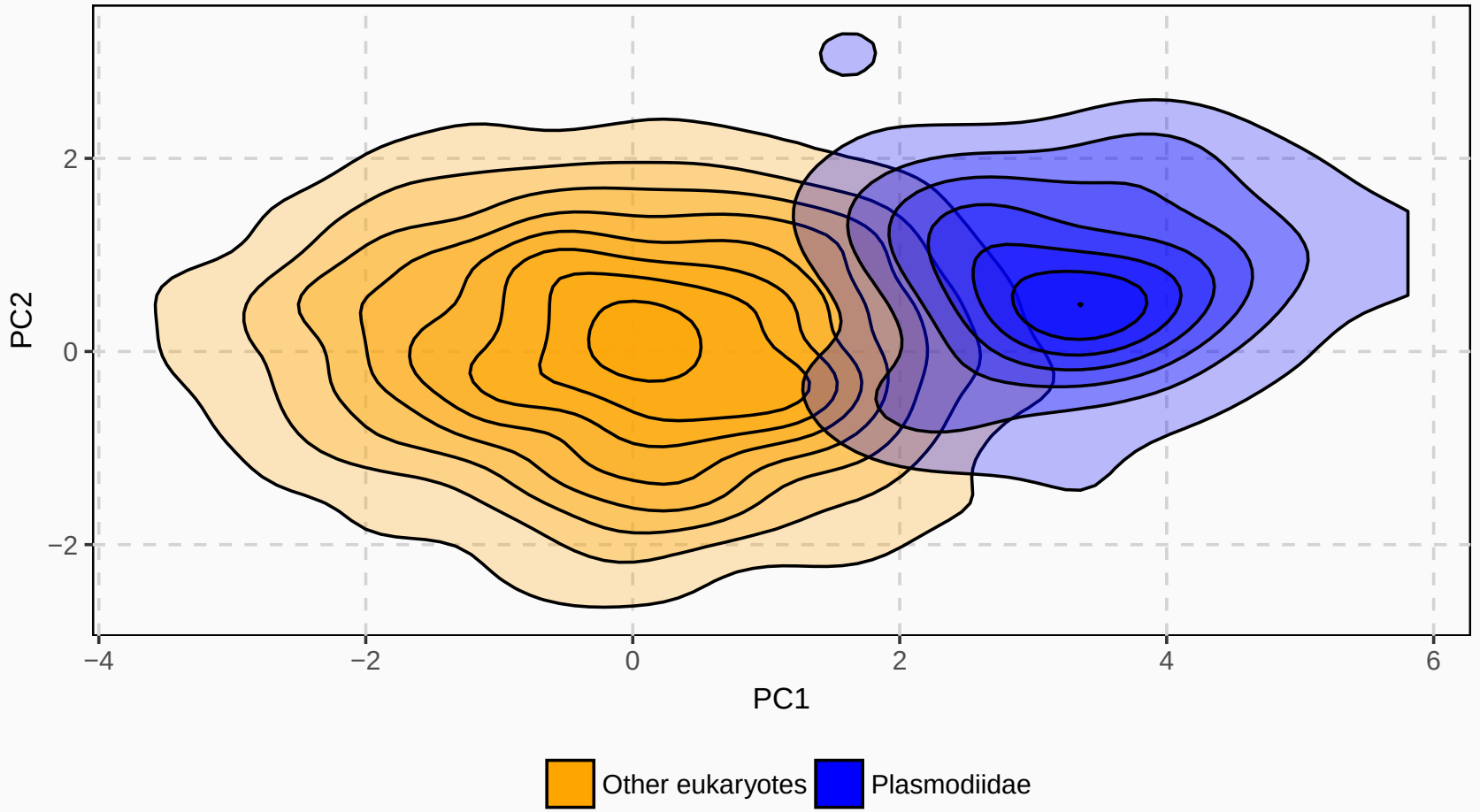


Signal peptides possess three distinct domains with variable length and characteristic amino acid composition (Hegde and Bernstein, 2006):

- 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.



n-grams

n-grams (k-tuples, k-mers):

- subsequences (continuous or gapped) of *n* residues,
- considers the context of a specific residue.

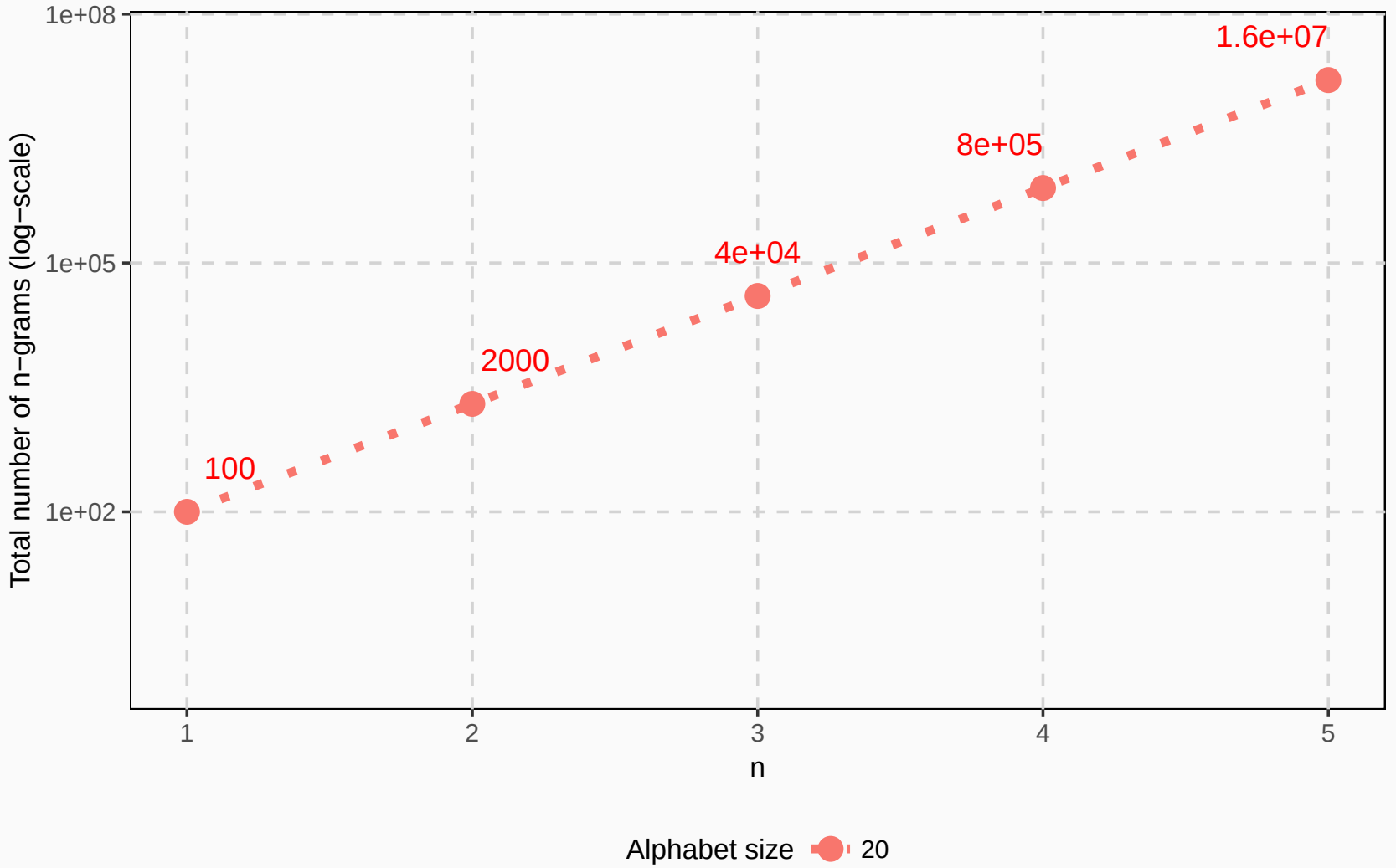
	P1	P2	P3	P4	P5
S1	M	R	K	L	Y

2-grams: MR, RK, KL, LY

2-grams (gap 1): M – K, R – L, K – Y

3-grams: MRK, RKL, KLY

n-grams



n-grams create very large datasets, which are hard to process and analyze.

Quick Permutation Test

Informative n-grams are usually selected using permutation tests.

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).

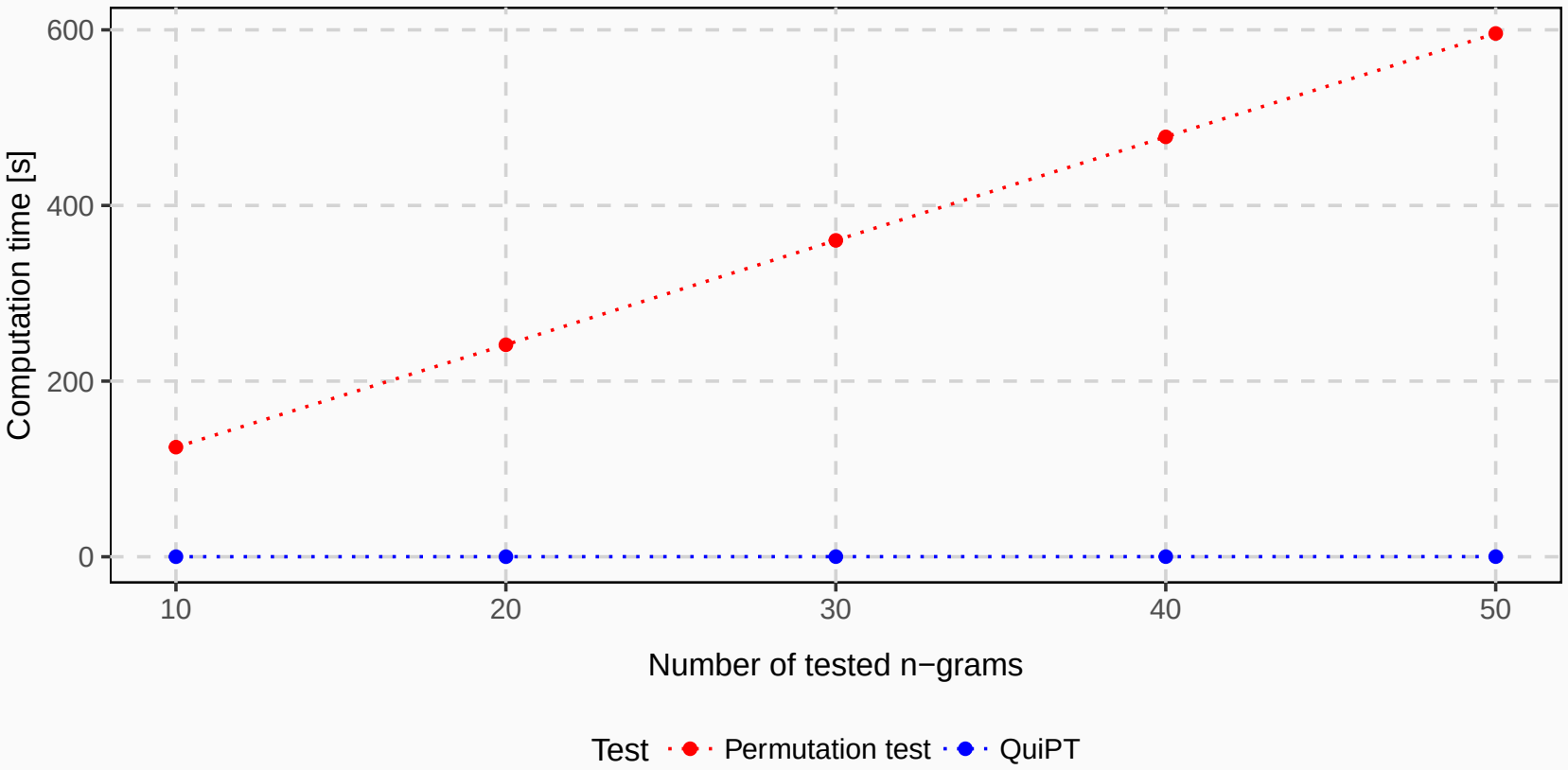
N: number of permutations.

QuiPT

Quick **P**ermutation **T**est is a fast alternative to permutation tests for n-gram data. It also allows precise estimation of p-value.

QuiPT is avaible as part of the **biogram** R package.

QuiPT



QuiPT is faster than classical permutation tests and returns exact p-values.

Simplified alphabets

Simplified alphabets:

- are based on grouping amino acids with similar physicochemical properties,
- ease computational analysis of a sequence (Murphy et al., 2000),
- create more interpretable models.

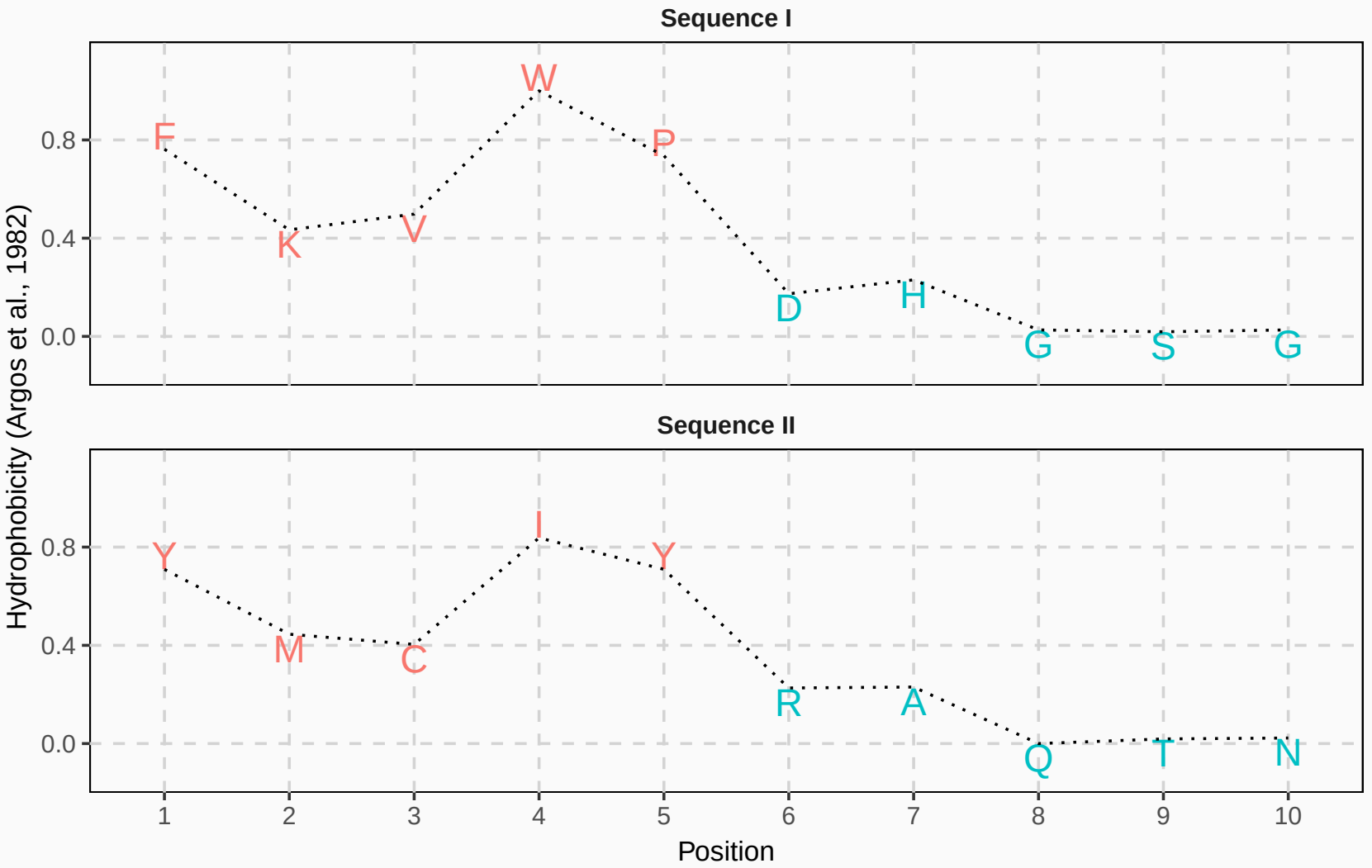
Two sequences that are drastically different considering their amino acids composition can have the same physicochemical properties.

Sequence I:

`FKVWPDHGSG`

Sequence II:

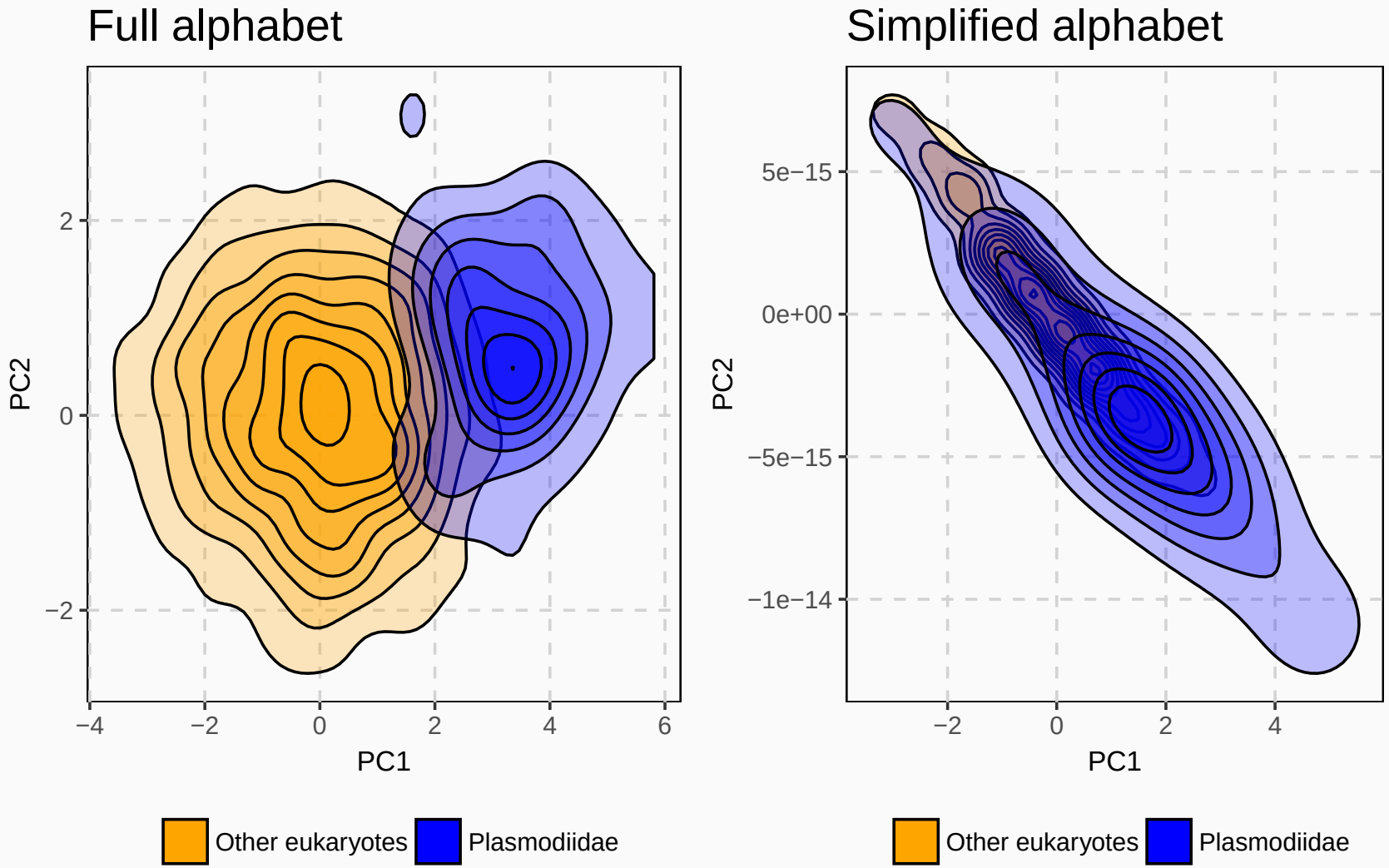
`YMCIYRAQTN`


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

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

Sequence I: `FKVWPDHGSG` $\rightarrow$ `1111122222`

Sequence II: `YMCIYRAQTN` $\rightarrow$ `1111122222`



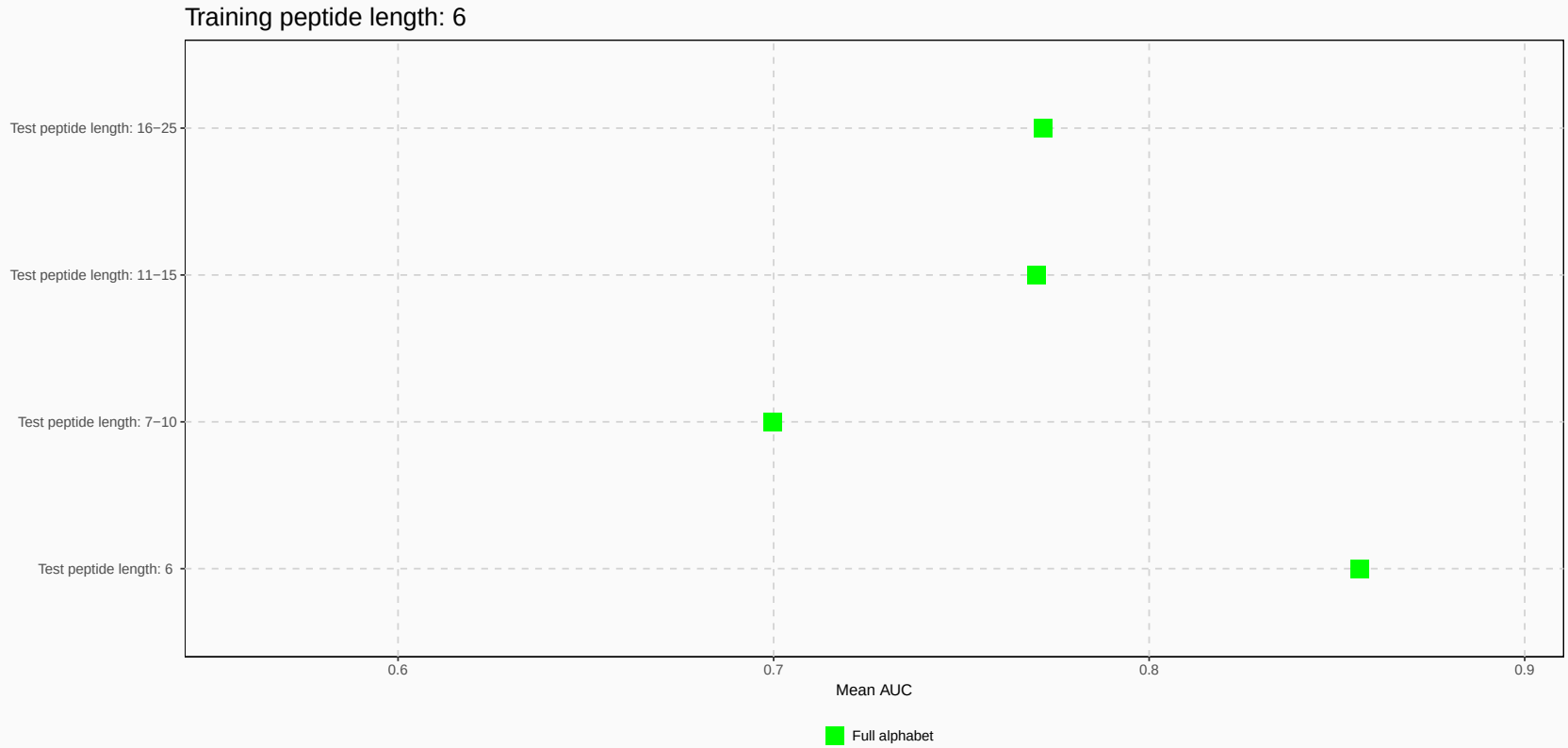
PCA of amino acid frequency in signal peptides.

Prediction of amyloidogenicity

AmyloGram

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

Cross-validation

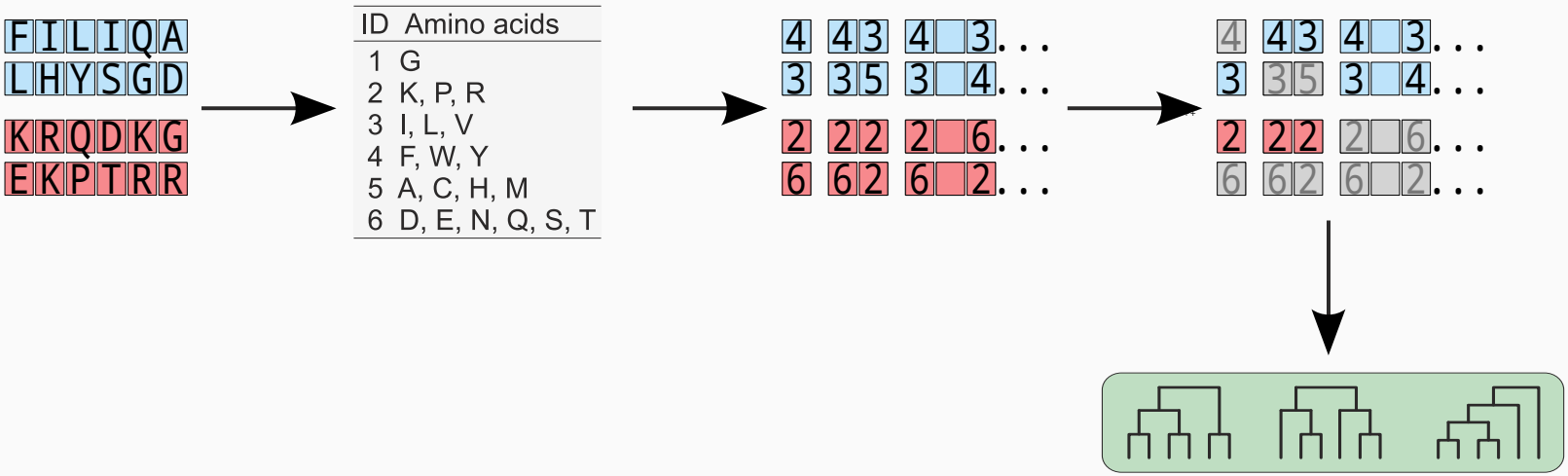


Does amyloidogenicity depend on the exact sequence of amino acids?

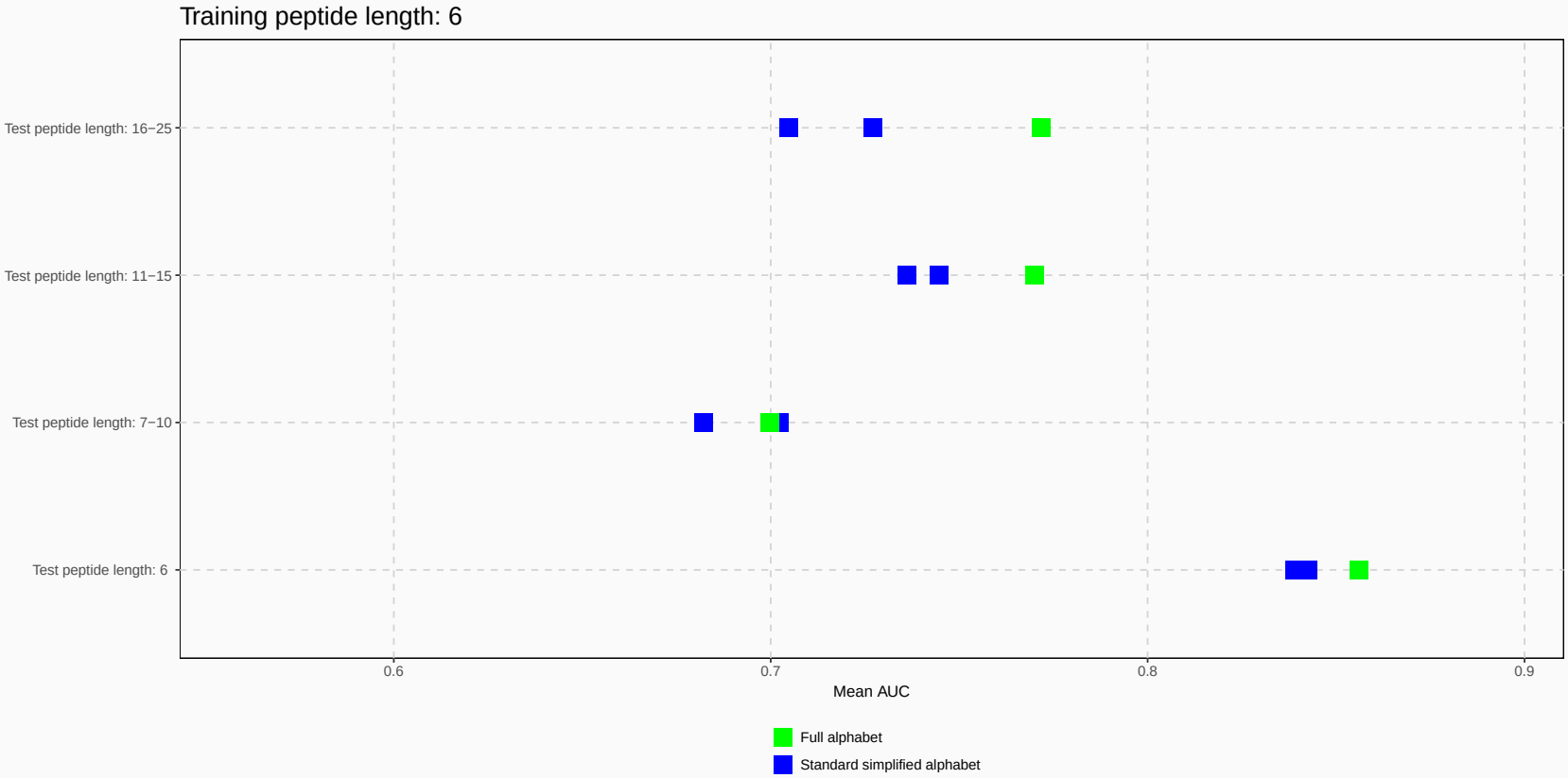
Standard simplified amino acid alphabets

To date, several simplified amino acid alphabets have been proposed, which have been applied to (among others) protein folding and protein structure prediction (Kosiol et al., 2004; Melo and Marti-Renom, 2006).

Standard simplified amino acid alphabets



Cross-validation

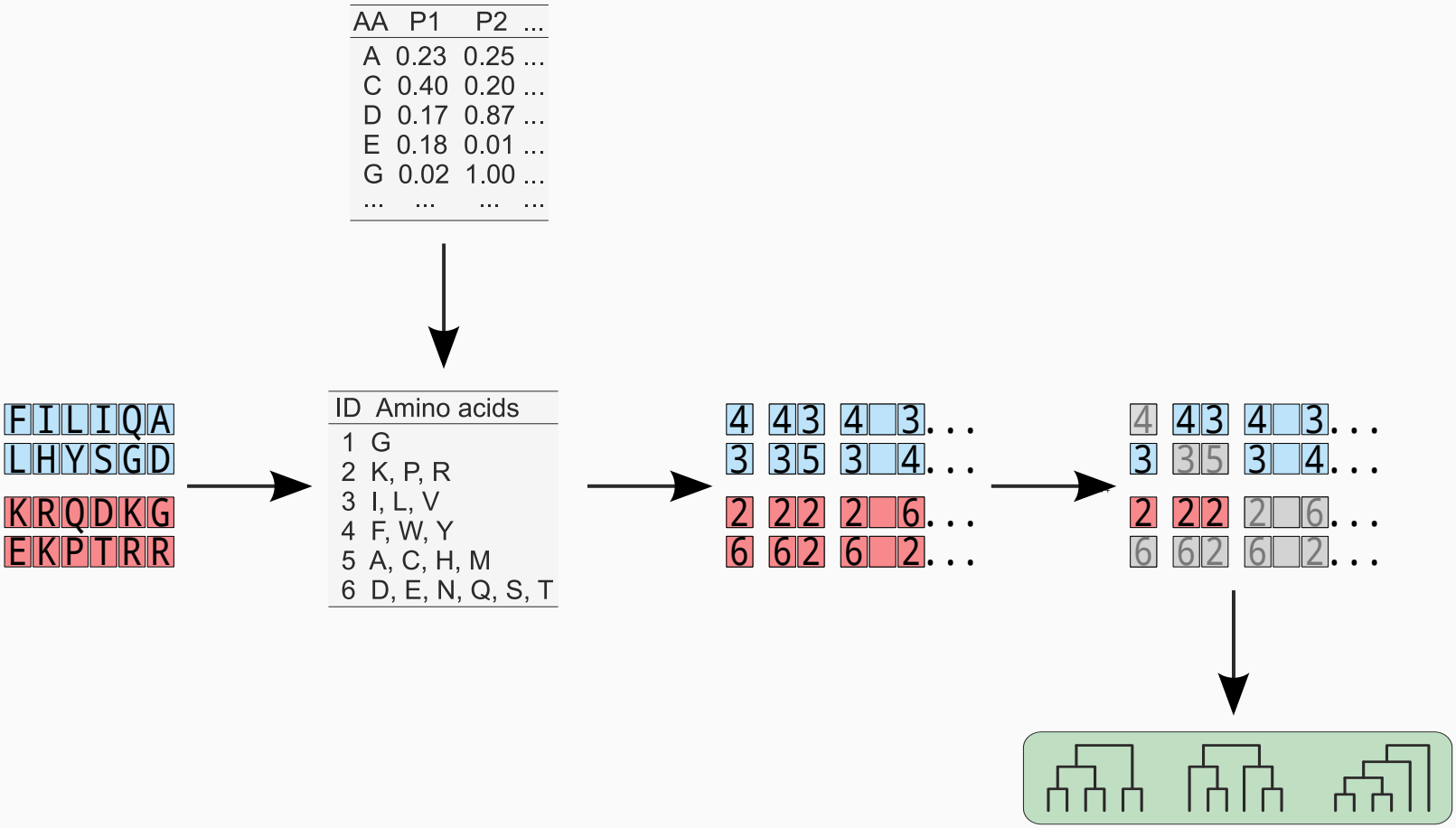


Standard simplified amino acid alphabets do not enhance discrimination between amyloidogenic and non-amyloidogenic proteins.

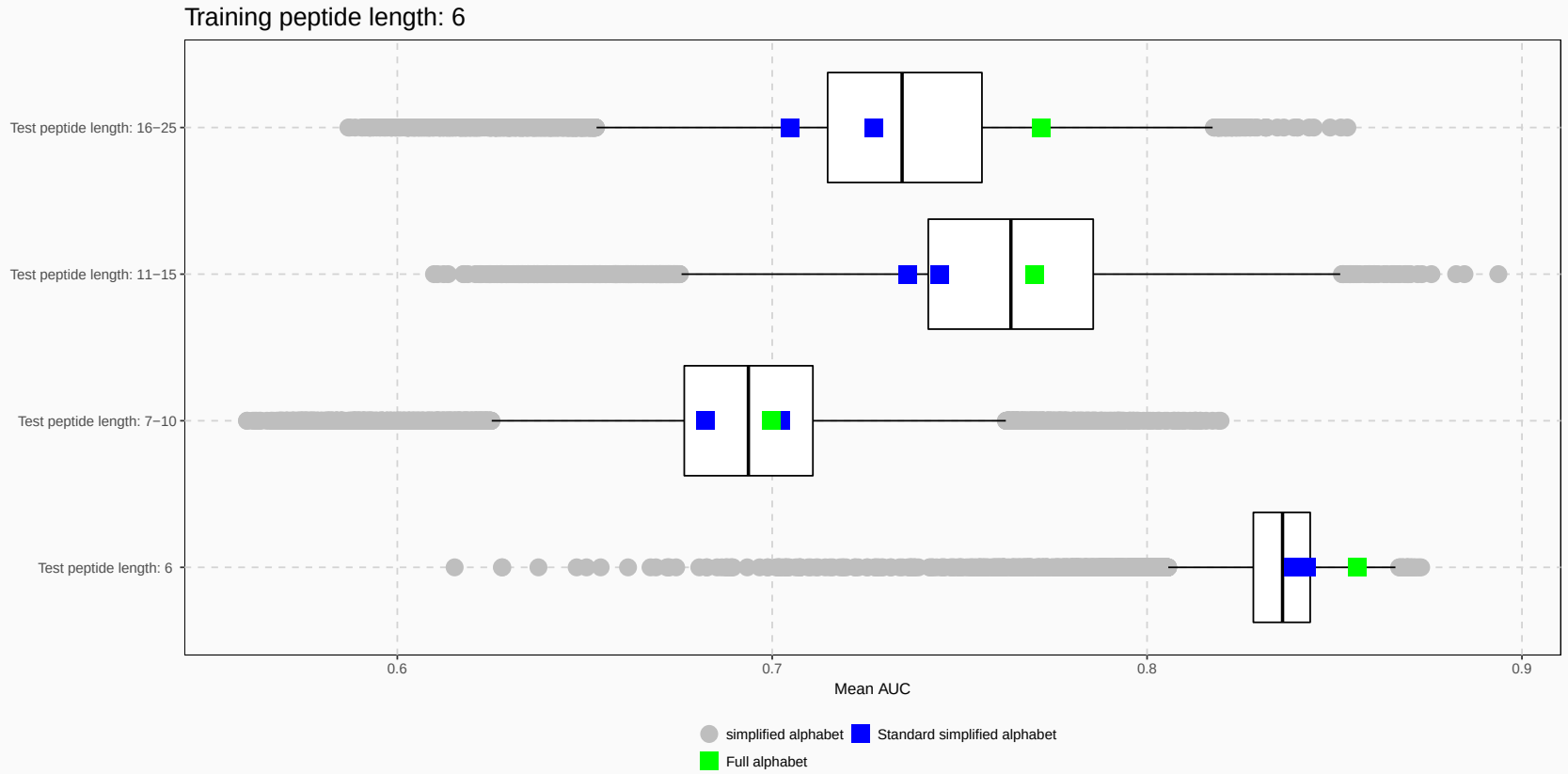
Novel simplified amino acid alphabets

- 17 measures handpicked from AAIndex database:
  - 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

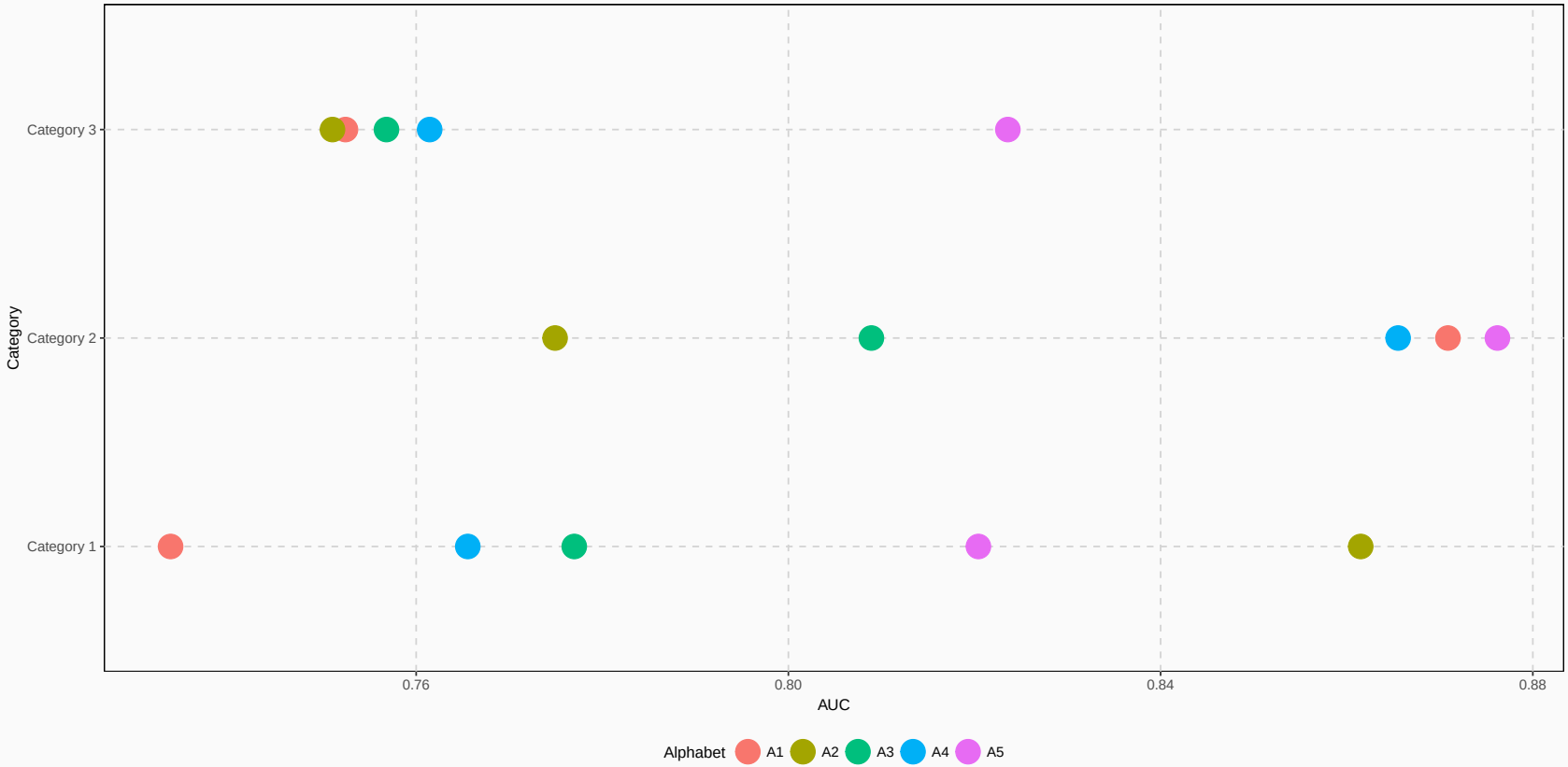


Cross-validation

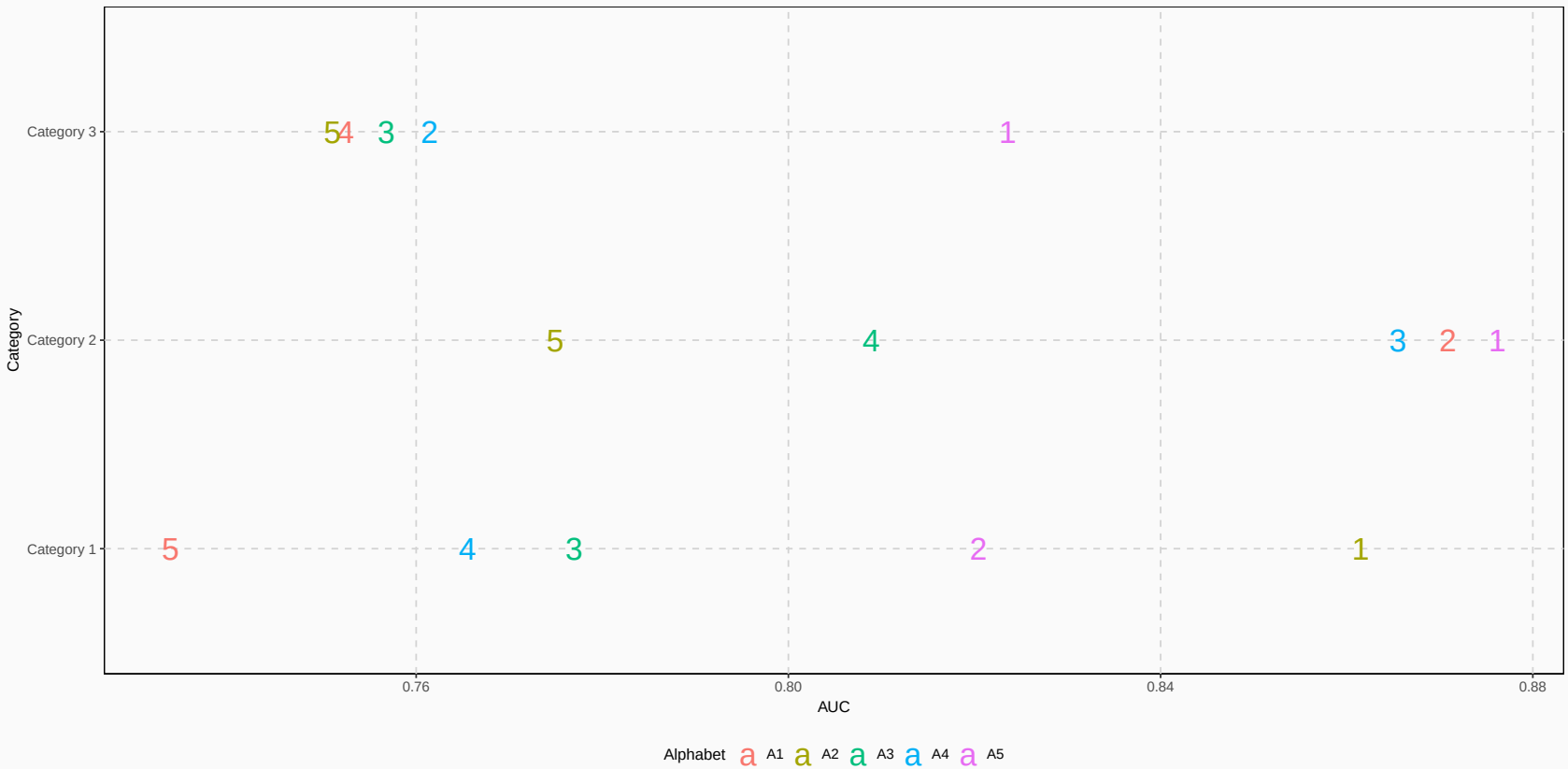


Hinges of boxes correspond to the 0.25 and 0.75 quartiles. The bar inside the box represents the median. The gray circles correspond to the simplified alphabets with the AUC outside the 0.95 confidence interval.

Ranking alphabets

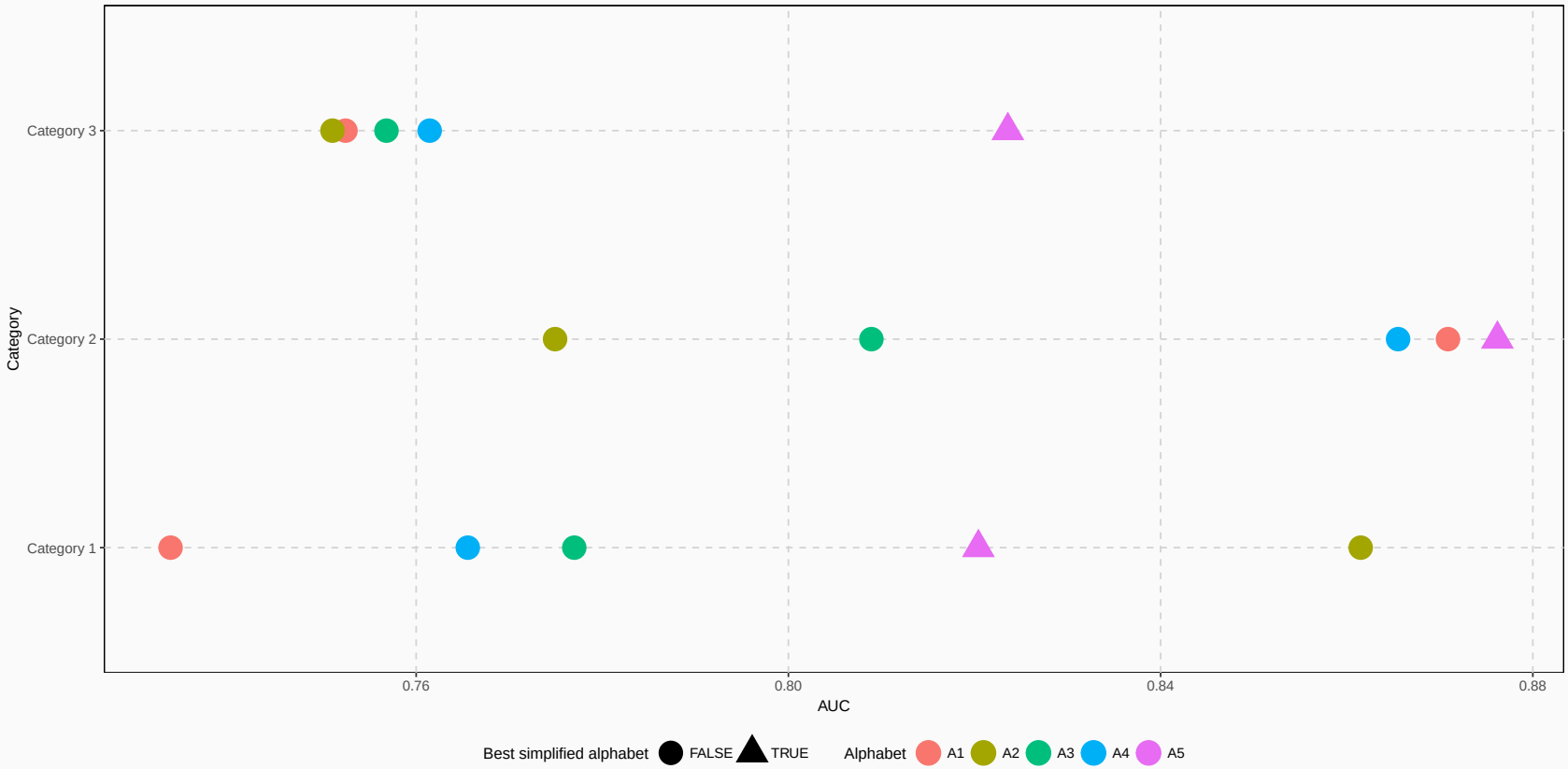


Ranking alphabets



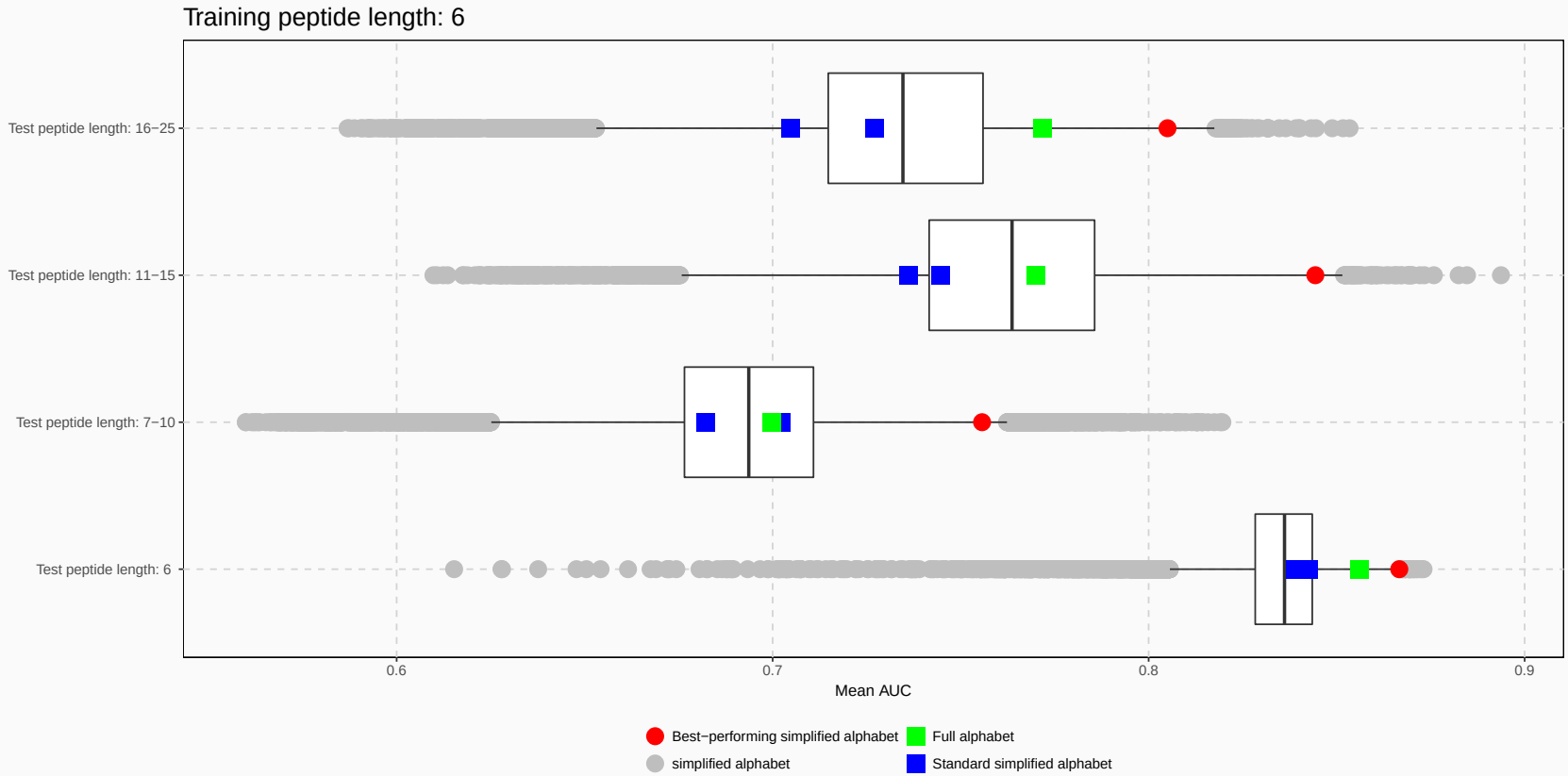
We rank alphabets separately in all length categories assuming the rank 1 for the best AUC, rank 2 for the second best AUC and so on.

Ranking alphabets



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

The best-perfoming simplified alphabet



The best-perfoming 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-perfoming 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-perfoming 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.

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).

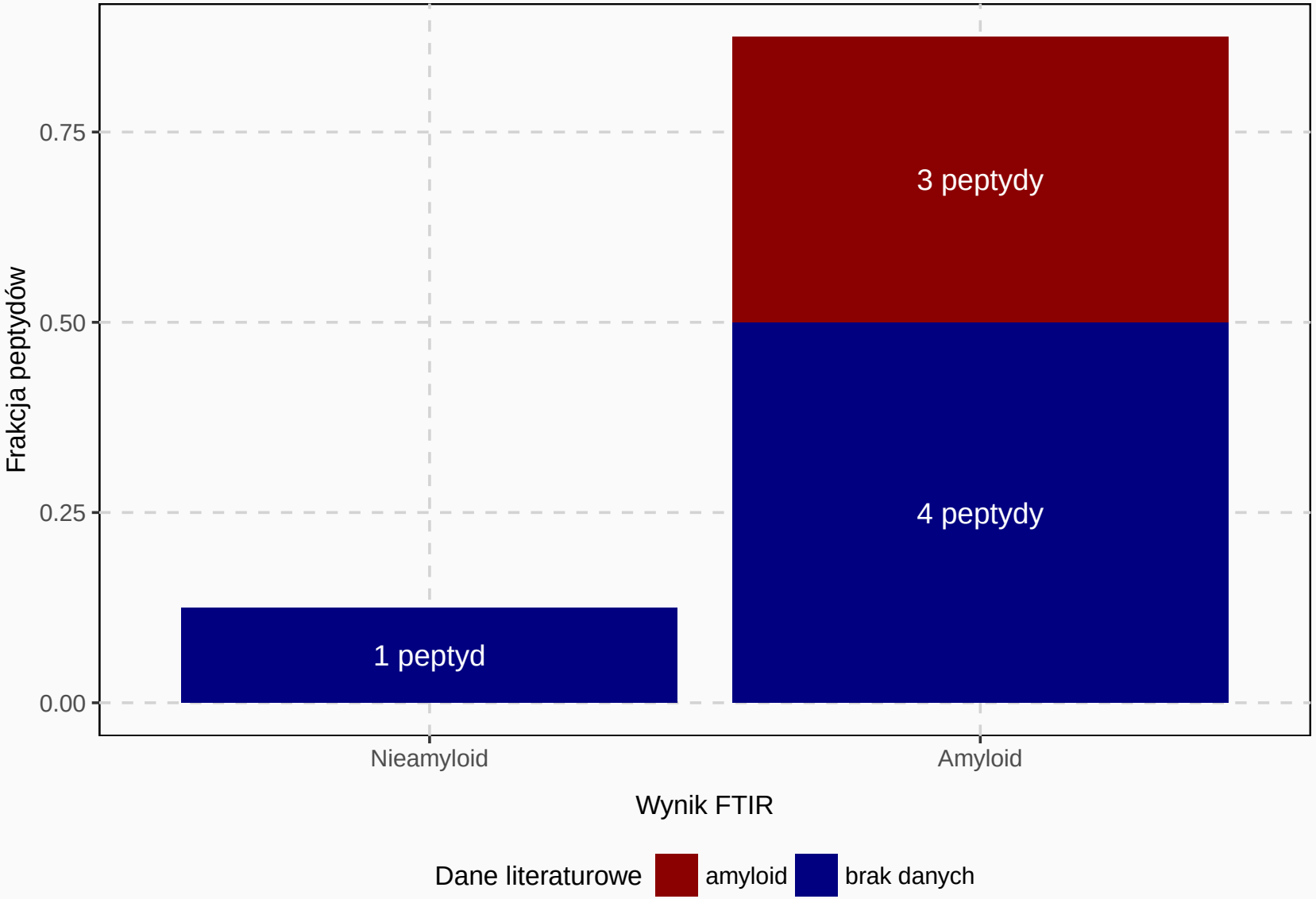
MCC (Matthew's Correlation Coefficient): miara jakości predykcji (1: idealny dobry klasyfikator, -1: idealnie zły klasyfikator).

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

Walidacja eksperymentalna

1. Wszystkie nieamyloidowe peptydy z bazy AmyLoad zanalizowano AmyloGramem.
2. Wybrano 8 peptydów z najwyższym prawdopodobieństwem amyloidogenności.
3. Peptydy zbadano przy pomocy spektroskopii fourierowskiej (FTIR).
4. Wyniki potwierdzono esejami z czerwienią Kongo i tioflawiną.

Walidacja eksperymentalna



Podsumowanie

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

Perspektywy

1. Zastosowanie opracowanej metodologii do przewidywania lokalizacji subkomórkowej białek.
2. Stworzenie oprogramowania wspierającego interpretowanie wyników analizy n-gramów i uproszczonych alfabetów.
3. 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.

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.

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

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.