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 and conduct experimental studies.

Examples:

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

Machine learning models can help in the understanding of biological phenomenons provided that they are not black boxes.

Aim

Create efficient methods for analysis of amyloids that have human-readable decision rules.

Outline

n-grams

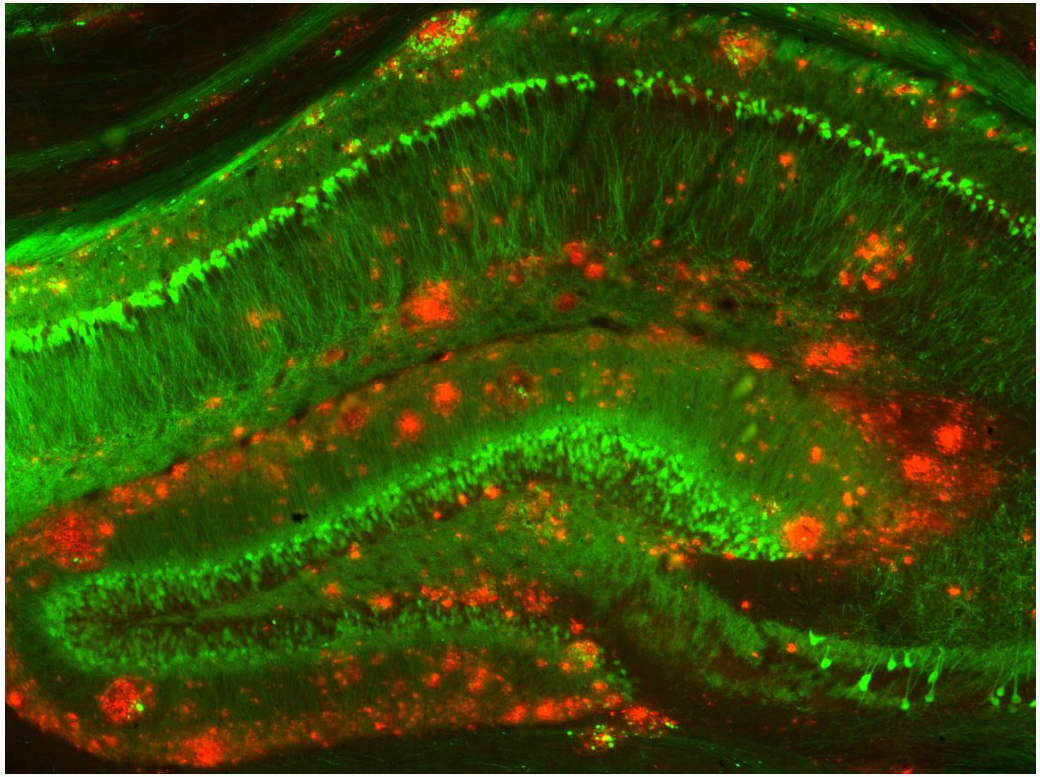
Simplified alphabets

Prediction of amyloidogenicity

Perspectives and summary

Amyloid proteins

Amyloid are aggregate-forming proteins associated with various diseases (e.g., Alzheimer's, Creutzfeldt-Jakob's and Huntington's diseases).



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

Amyloid proteins

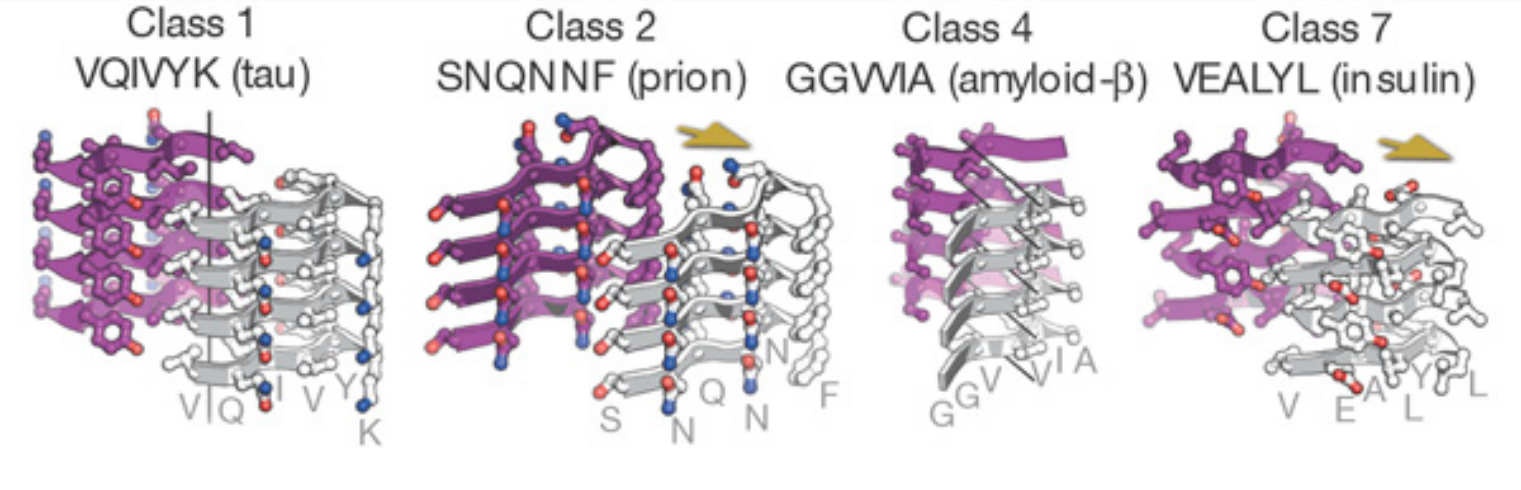
Functional amyloids:

- Pmel17,
- RIP1 and RIP3,
- acrosomal matrix proteins,
- HET-s,
- proteinaceous scaffolds of biofilms.

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)

n-grams

Computational analysis of biological sequences requires converting them to features understandable by machines.

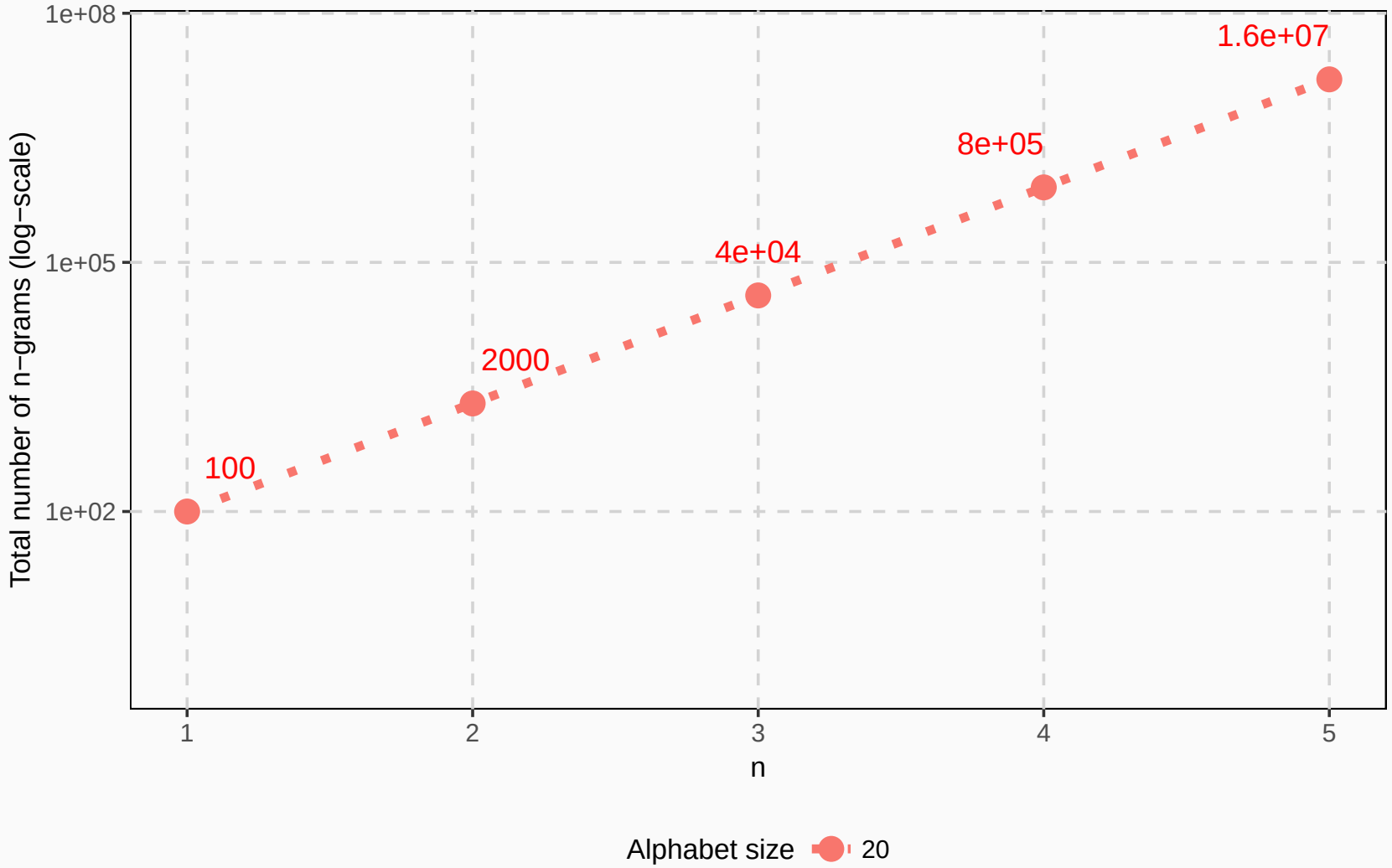
The optimal conversion of information:

- loss-less,
- concise.

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

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

```
## Error in file(con, "r"):  cannot open the
connection

## Error in lapply(seqs, function(i) i[1L:27]):
object 'seqs' not found

## Error in colnames(seq_df)[-1] <- paste0("P",
1L:27):  object 'seq_df' not found

## Error in melt(seq_df, id.vars = "name", value.name
= "aa", variable.name = "pos"):  object 'seq_df' not
found

## Error in as.matrix(seq_df[, 2L:10]):  object
'seq_df' not found

## Error in ncol(sample.seq): object 'sample.seq' not
```



Longer n-grams are more informative, but create larger feature spaces, which are hard to process and analyze.

Permutation Tests

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

```
## Error in readChar(con, 5L, useBytes = TRUE):
cannot open the connection

## Error in lapply(times, function(i) t(sapply(i,
function(j) {:  object 'times' not found

## Error in ggplot(times_dat, aes(x = size, y =
value, color = variable)):  object 'times_dat' not
found
```

QuiPT (available as a part of the **biogram** R package) 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 explicite models.

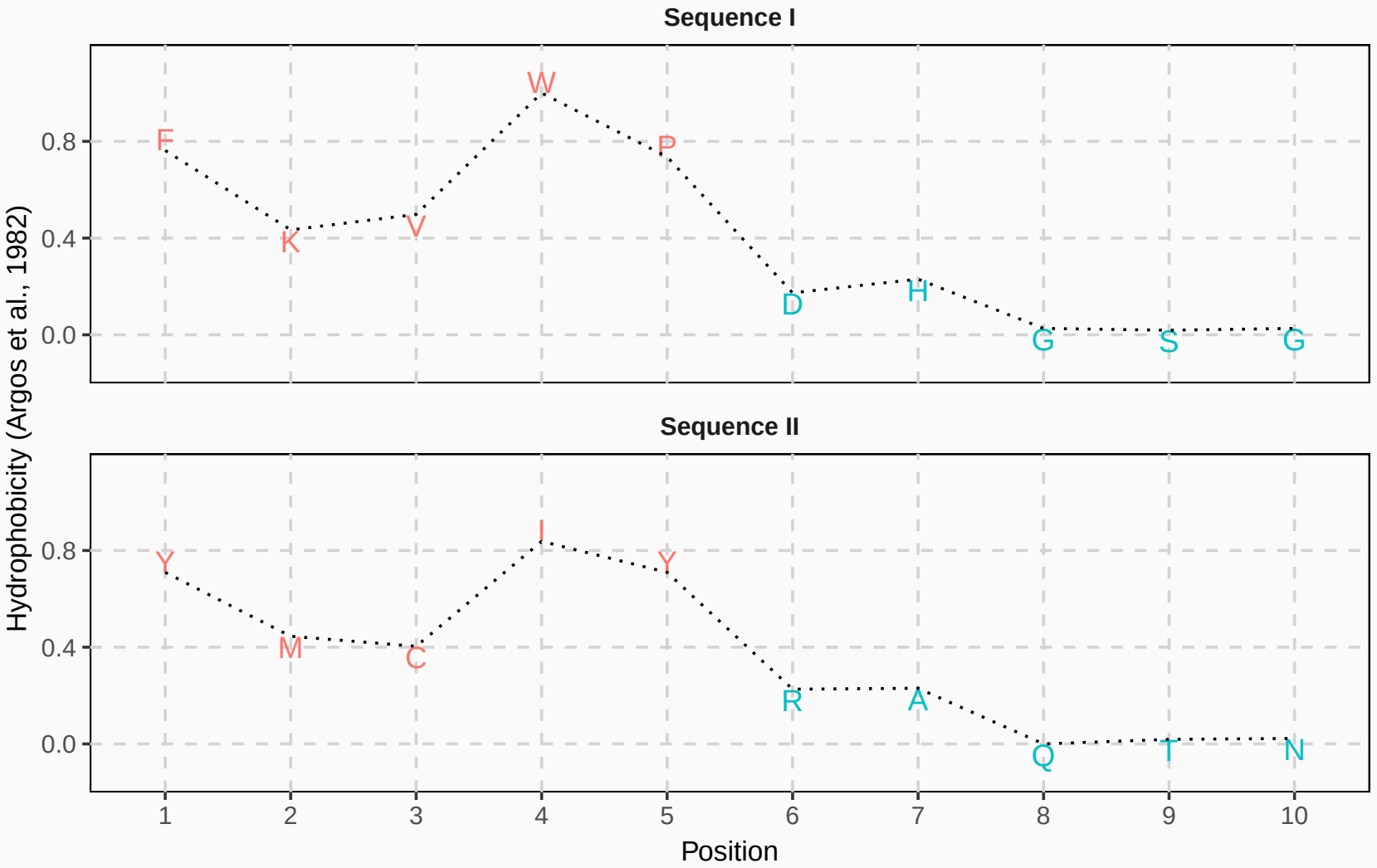
Two sequences that have drastically different amino acids composition may have very similar physicochemical properties.

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` `1111122222`

Sequence II: `YMCIYRAQTN` `1111122222`

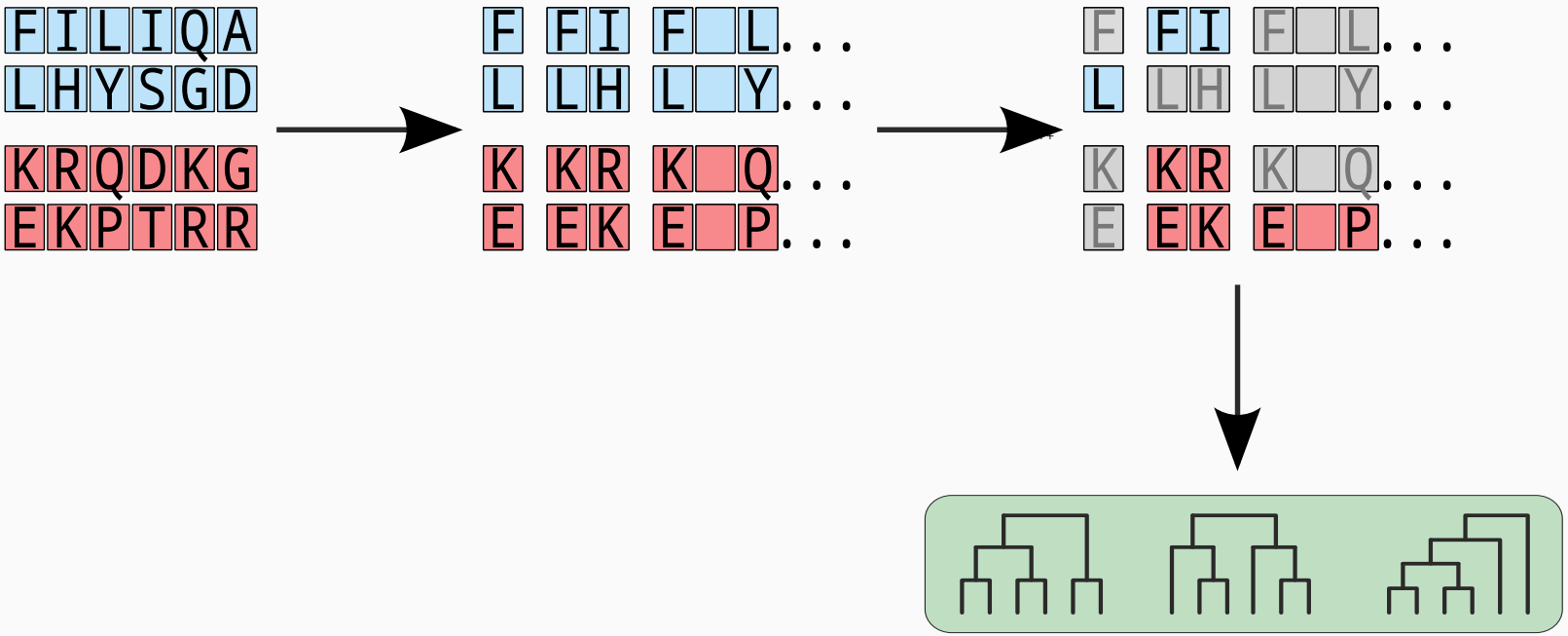
Prediction of amyloidogenicity

Data

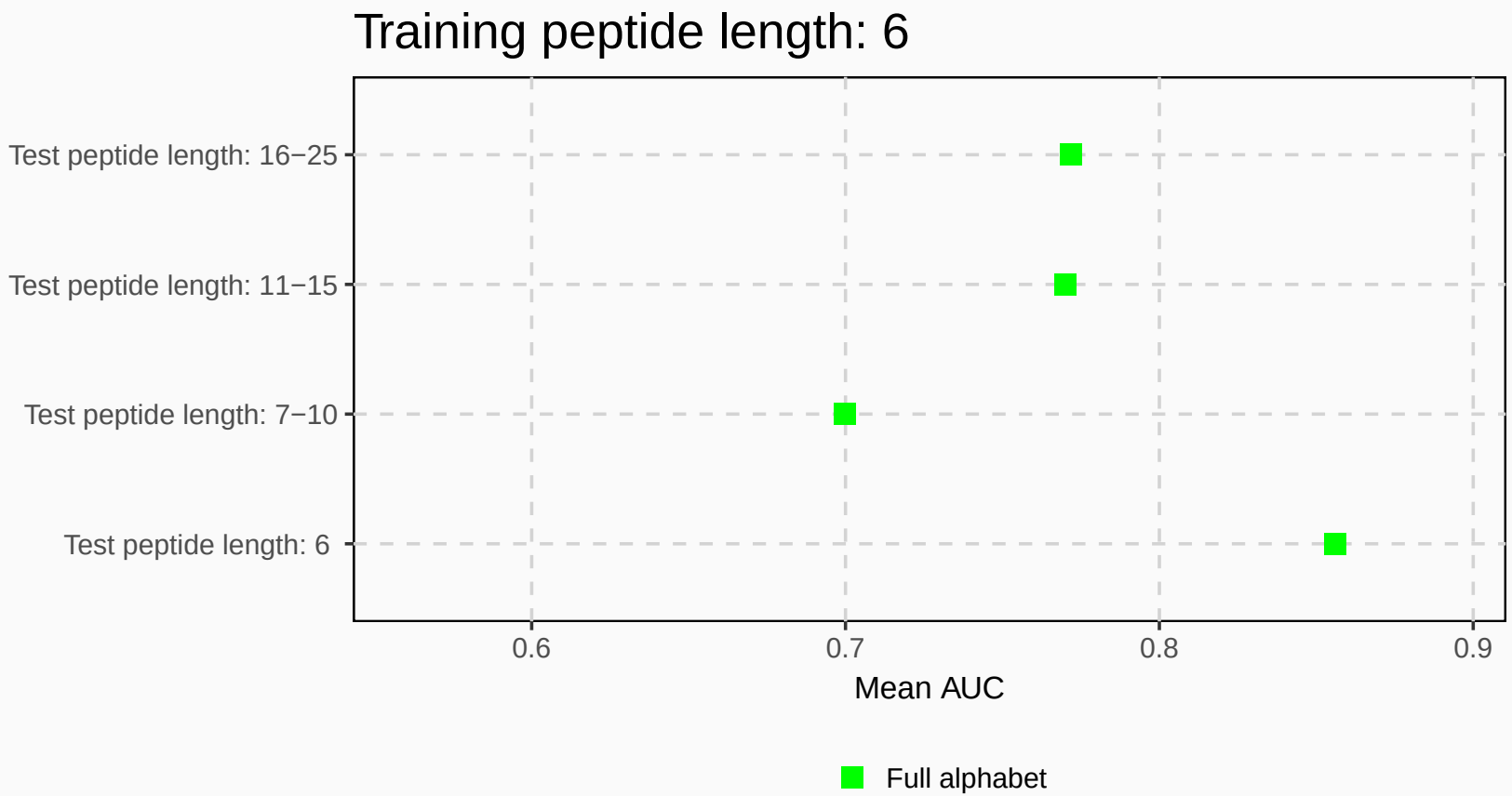
AmyLoad: a database of amyloid fragments (Wozniak and Kotulska, 2015).

- 1465 fragments,
- 11915 residues,
- 421 aggregation-prone fragments,
- 4312 residues (36.19%) in aggregation-prone fragments.

Can we predict amyloid fragments using n-gram data?


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...

Cross-validation



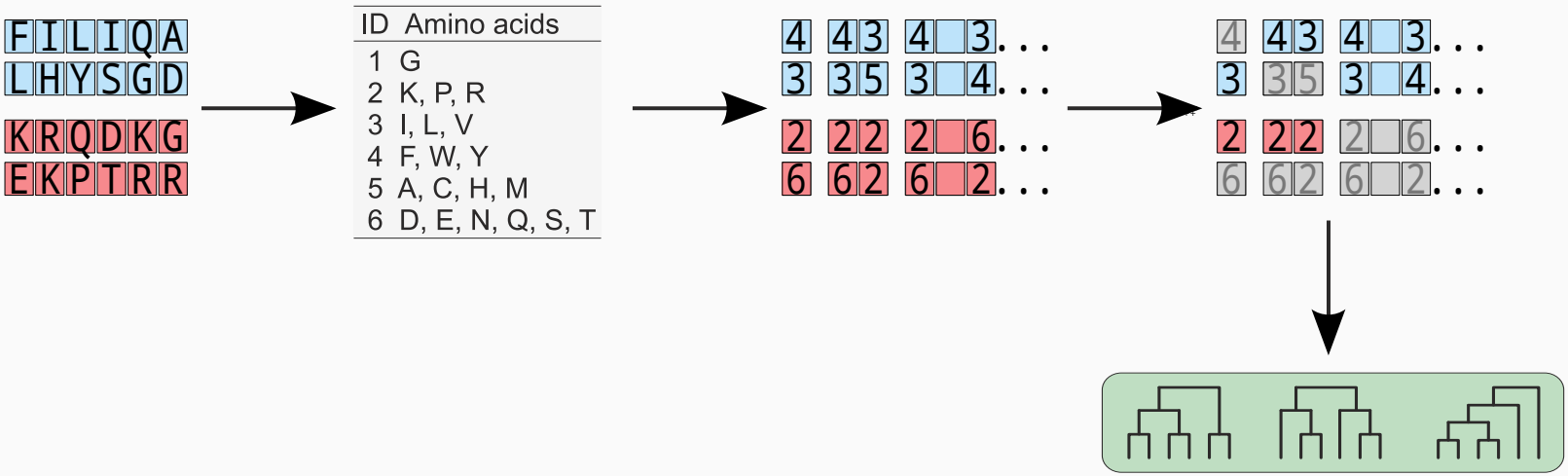
AUC (Area Under the Curve) measures the performance of a classifier (1 - classifier always properly recognizes amyloid proteins, 0 - classifier never properly recognizes amyloid proteins).

Does amyloidogenicity depend on the exact sequence of amino acids?

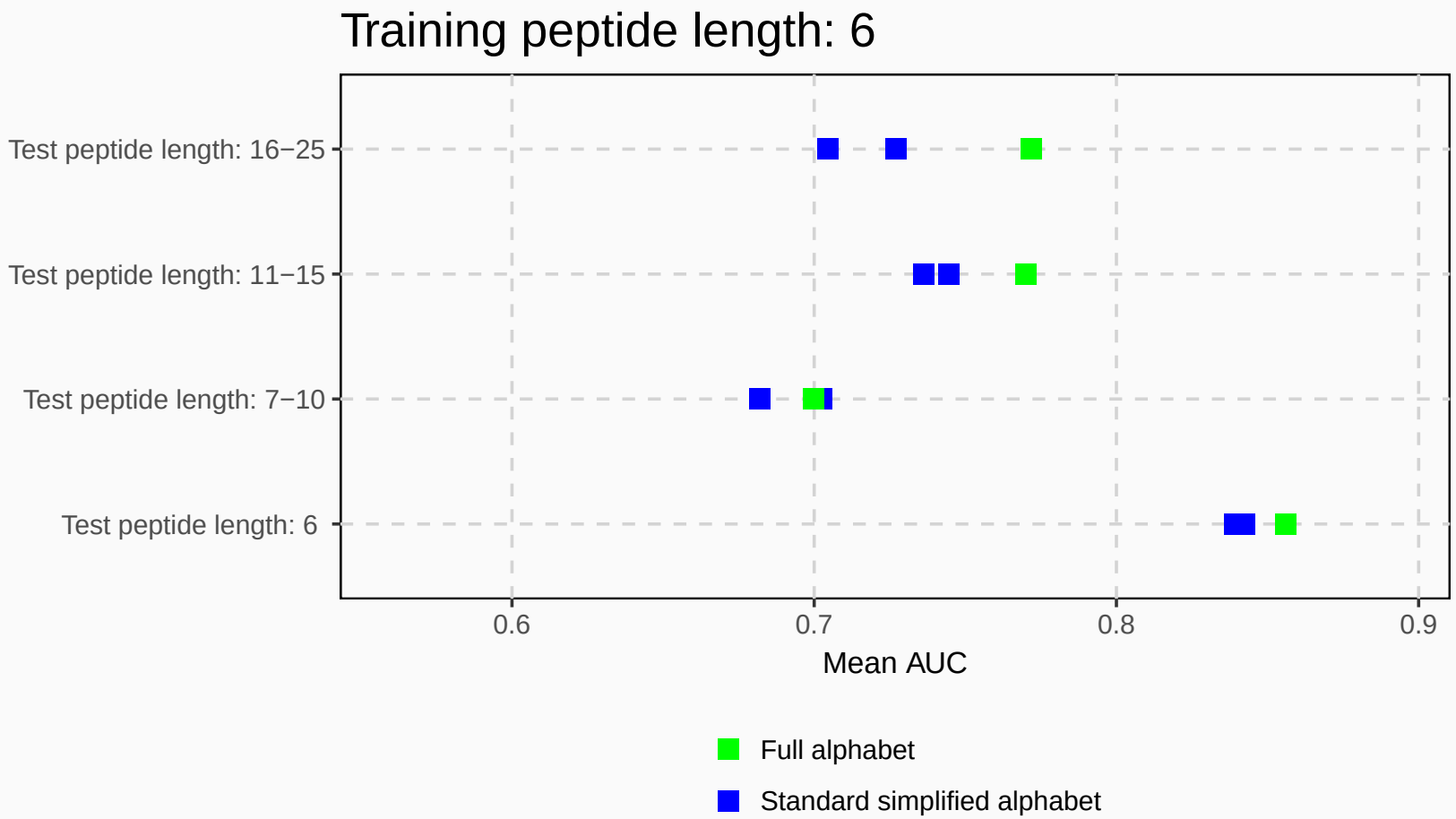
Standard simplified amino acid alphabets

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

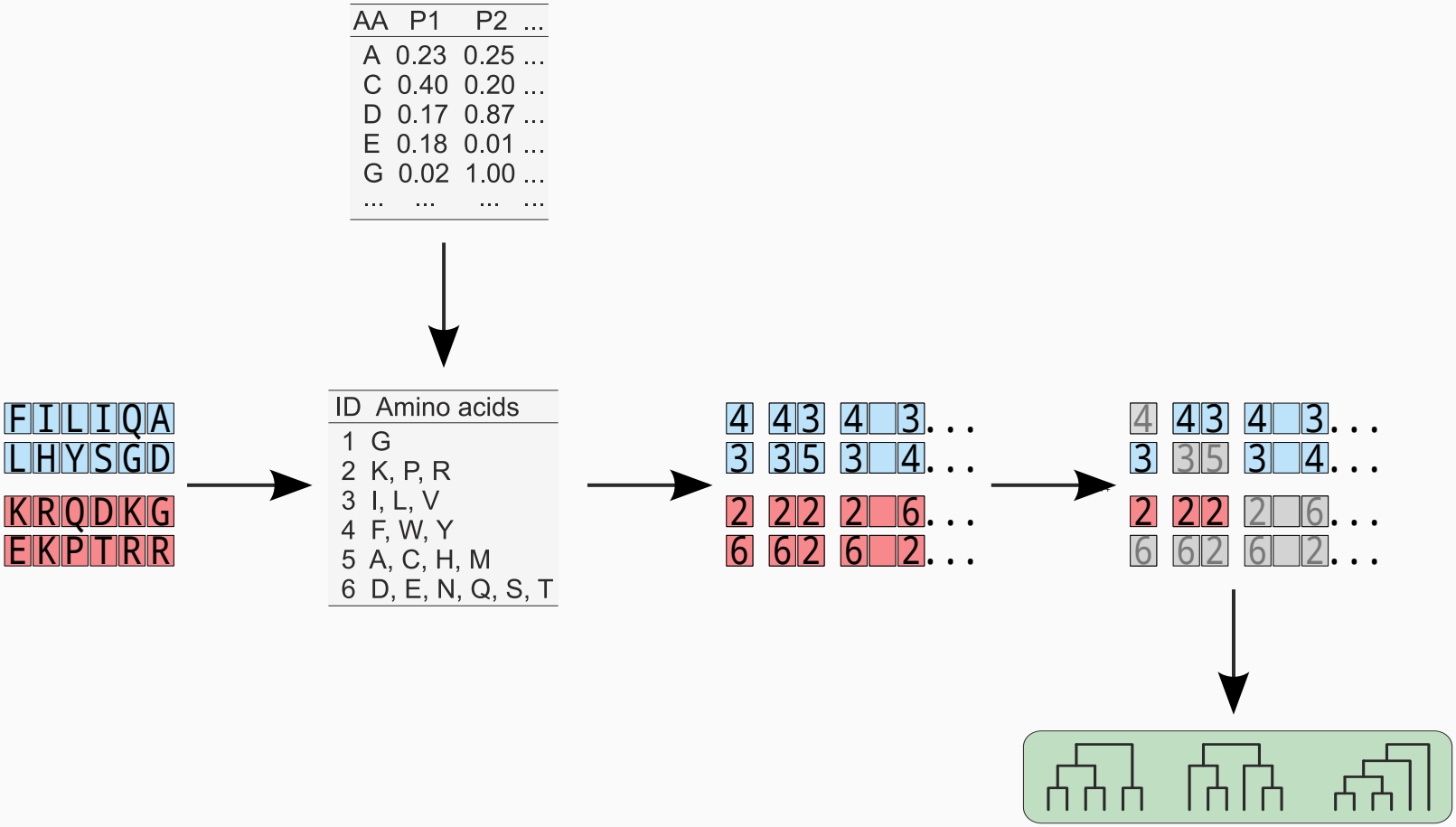


Standard simplified amino acid alphabets do not enhance discrimination between amyloidogenic and non-amyloidogenic proteins.

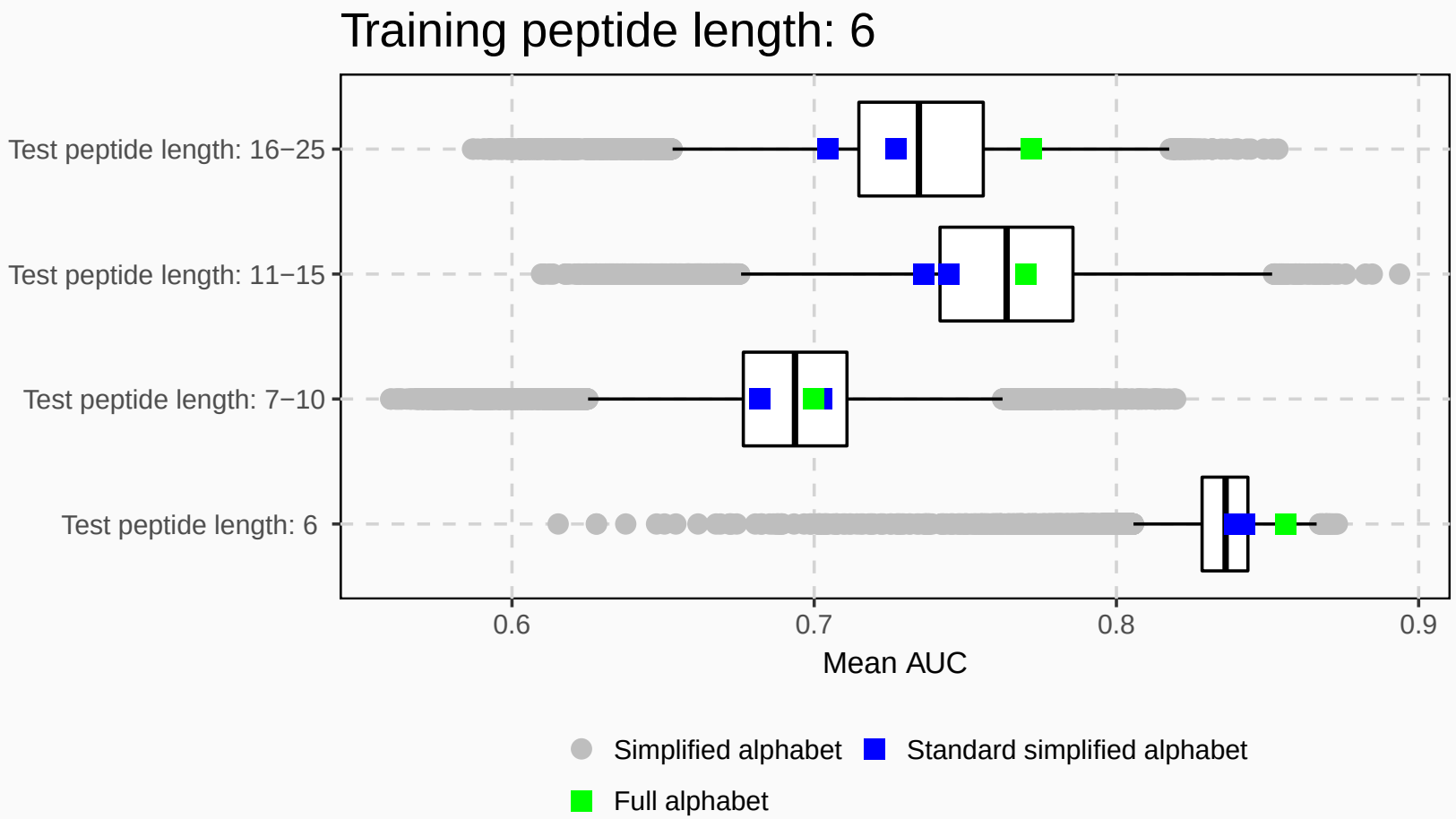
Novel simplified amino acid alphabets

- 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

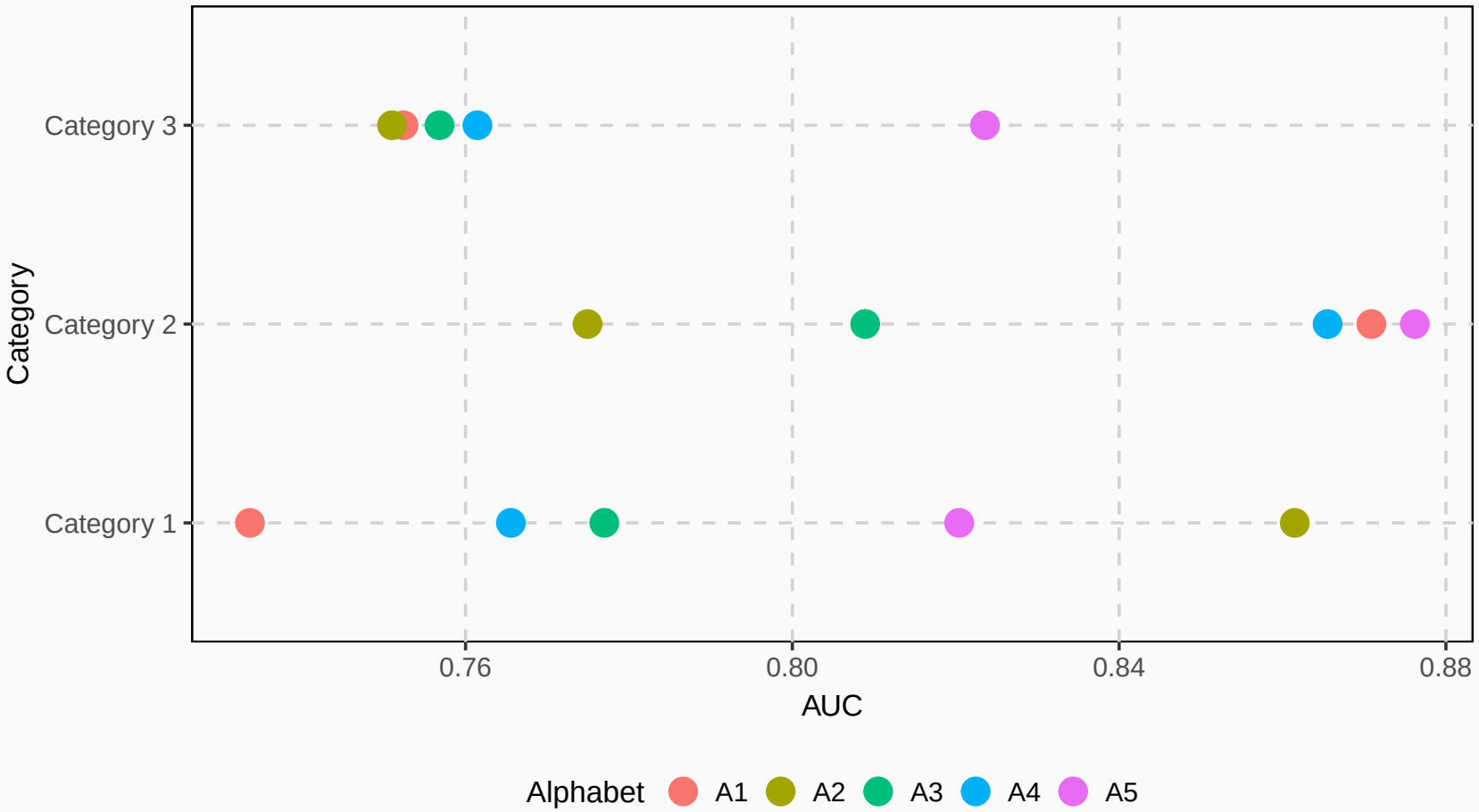


Cross-validation

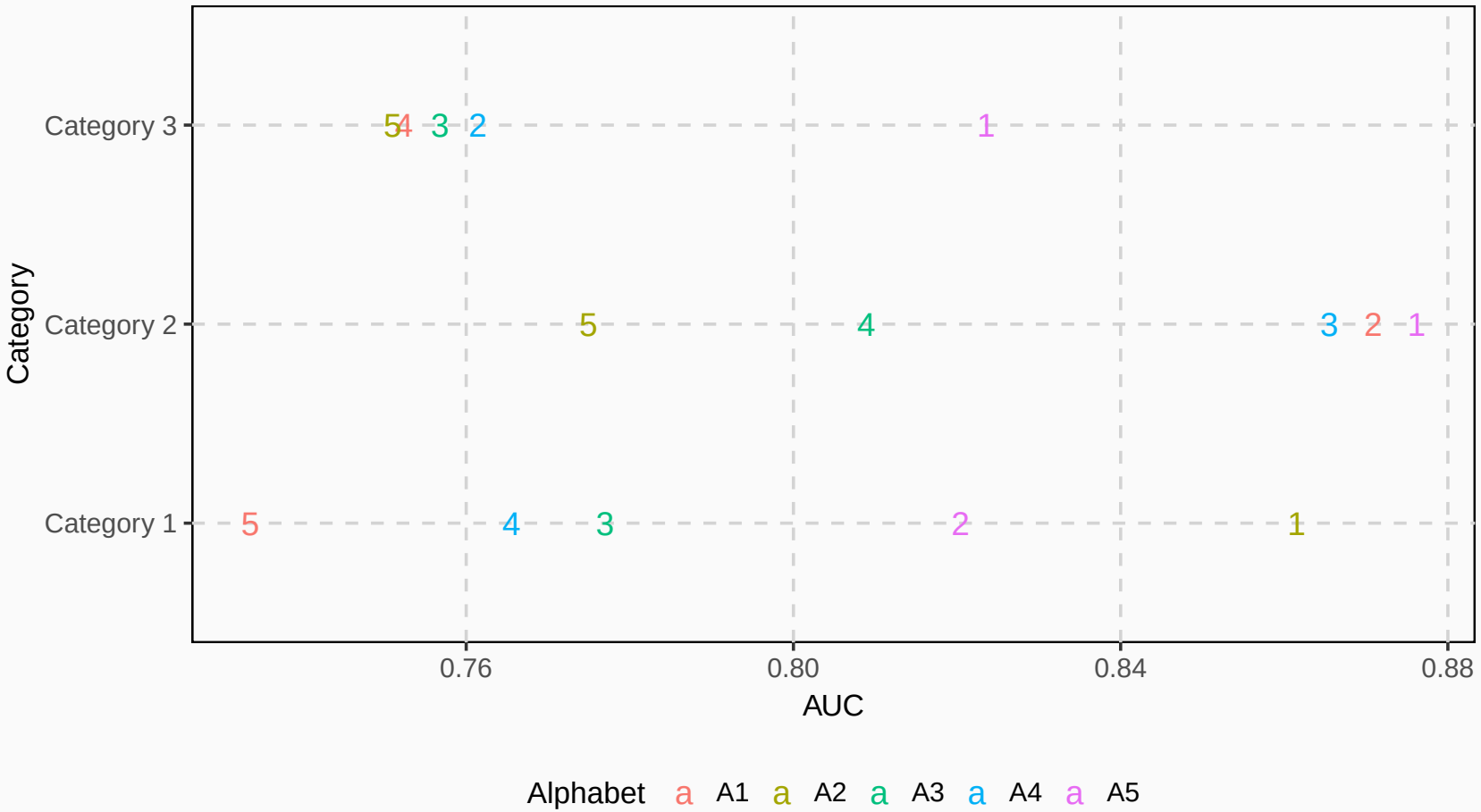


Hinges of boxes correspond to the 0.25 and 0.75 quartiles. The bar inside the box represents the median. The gray circles correspond to the simplified alphabets with the AUC outside the 0.95 confidence interval.

Ranking alphabets

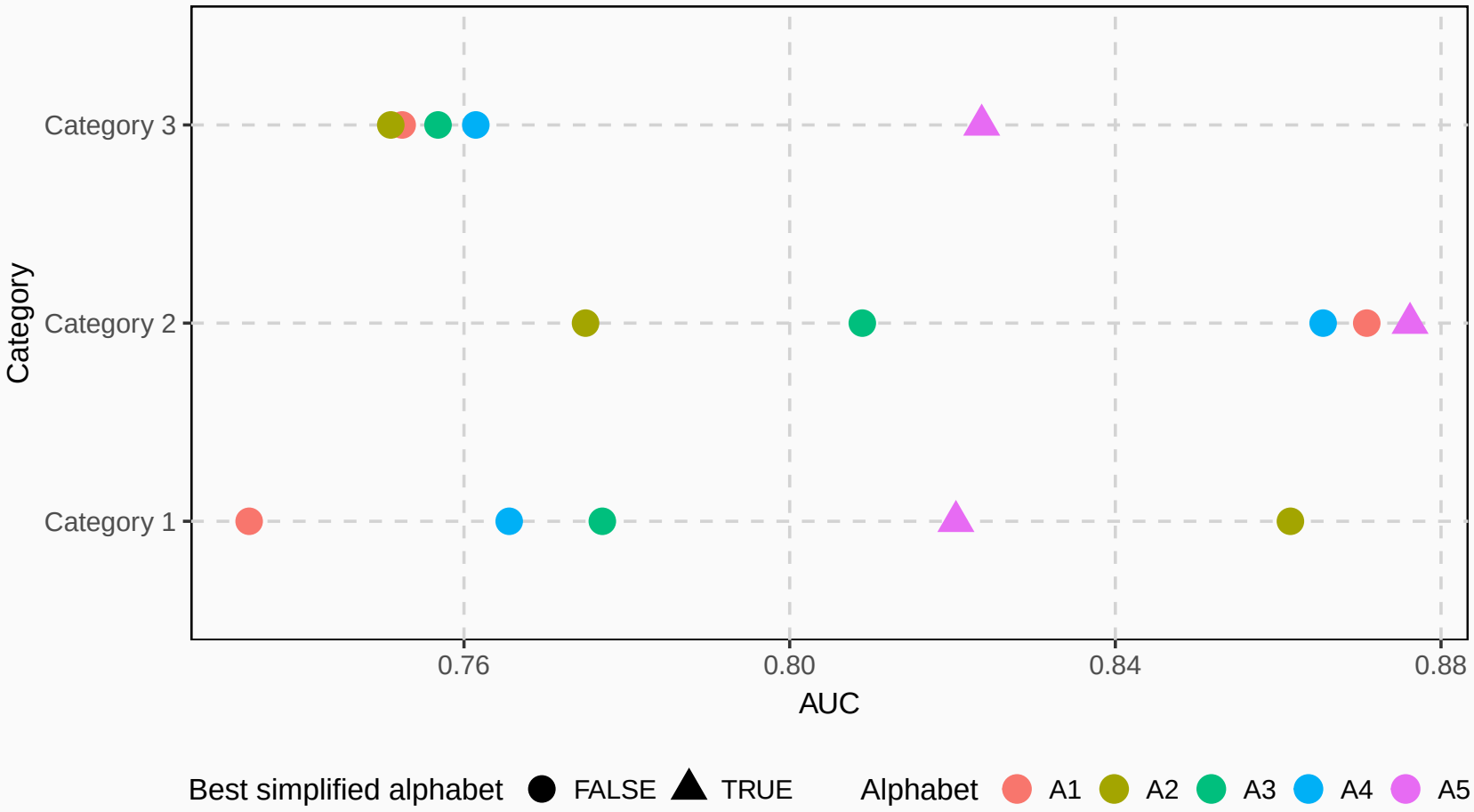


Ranking alphabets



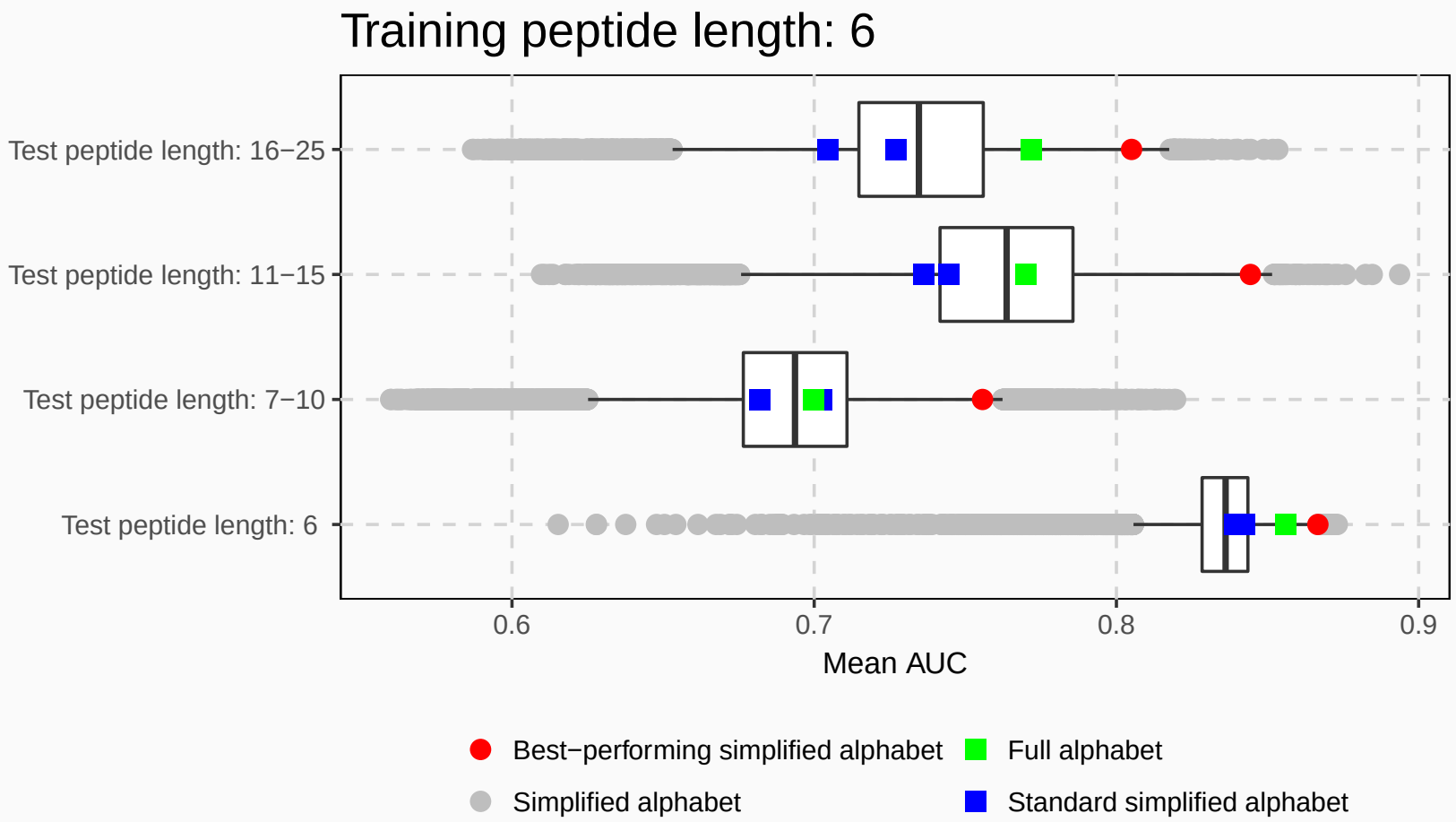
We rank alphabets separately in all length categories assuming the rank 1 for the best AUC, rank 2 for the second best AUC and so on.

Ranking alphabets



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

The best-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

```
## Error in file(file, "rt"):  cannot open the
connection

## Error in levels(si_dat[["et"]]) <-
c("Best-performing simplified alphabet", :  object
'si_dat' not found

## Error in ggplot(si_dat, aes(x = si, y = AUC_mean)):
object 'si_dat' not found

## Error in print(simil_plot):  object 'simil_plot'
not found
```

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

Similarity index

```
## Error in print(simil_plot): object 'simil_plot'
not found
```

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

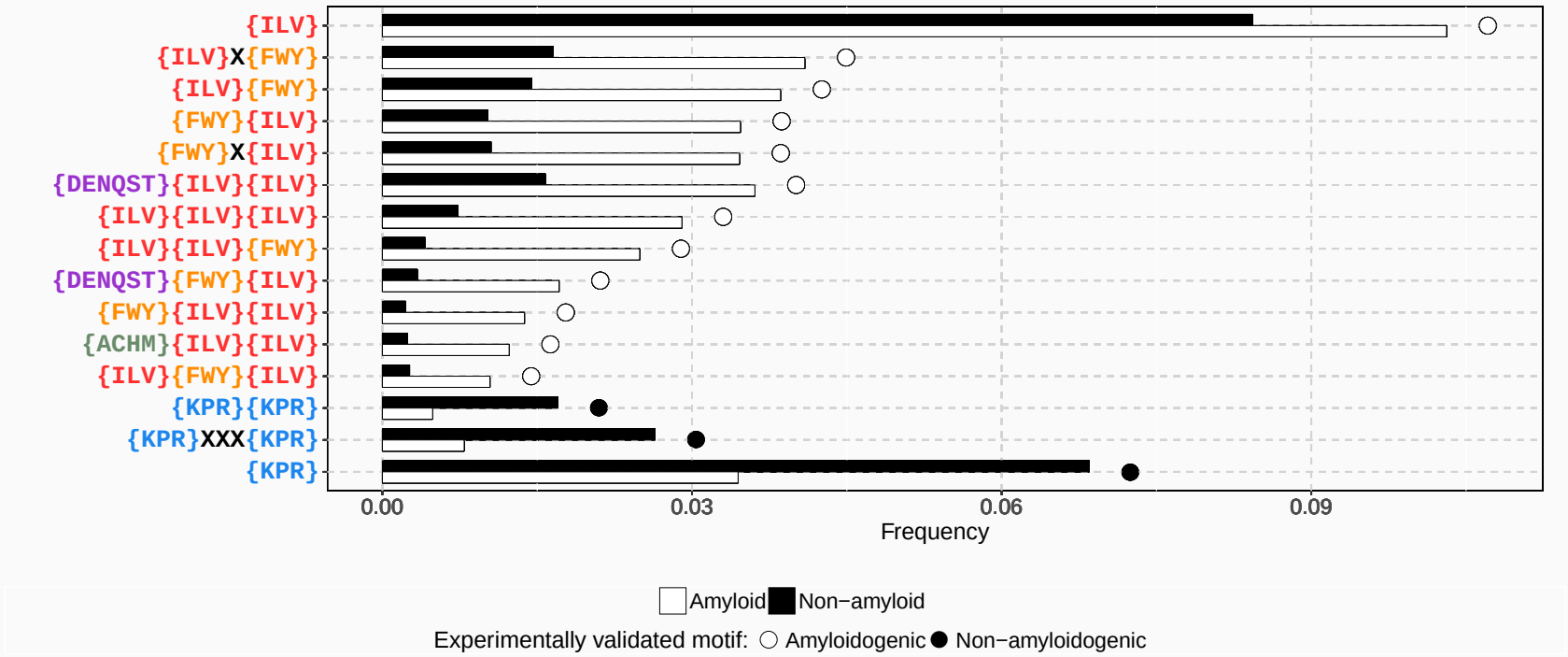
Similarity index

```
## Error in print(simil_plot): object 'simil_plot'
not found
```

The correlation between mean AUC and 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

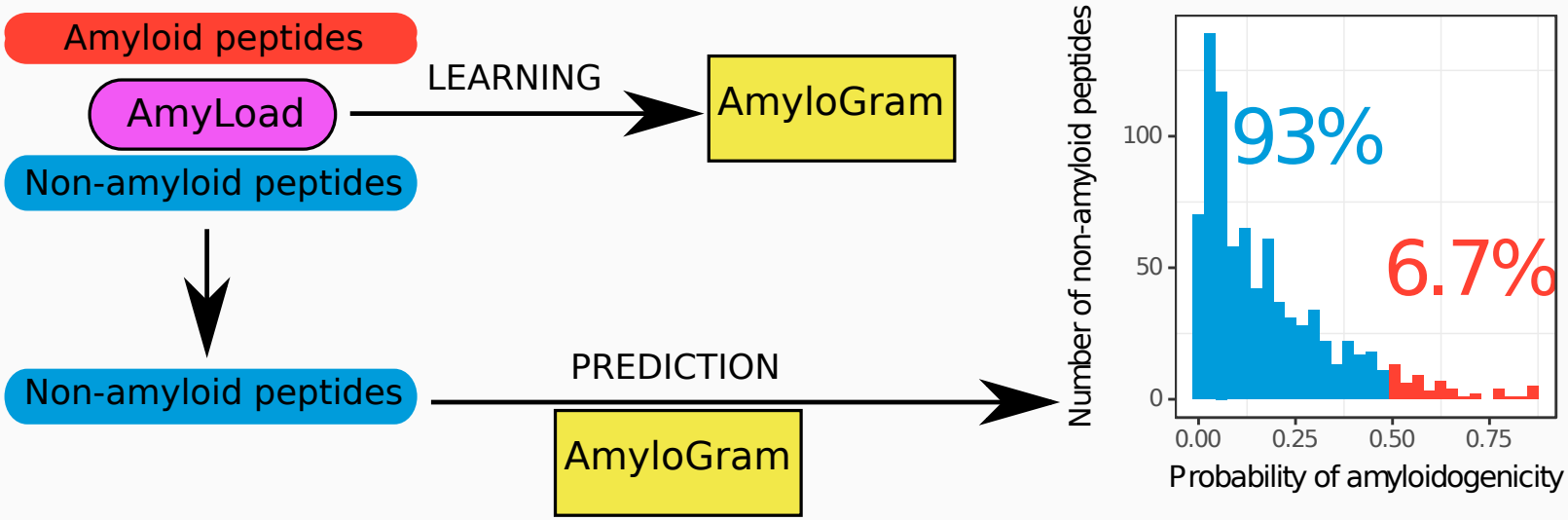
The predictor based on the best-performing alphabet, called AmyloGram, was benchmarked against the most popular tools for the detection of amyloid peptides using an external data set *pep424*.

Benchmark results

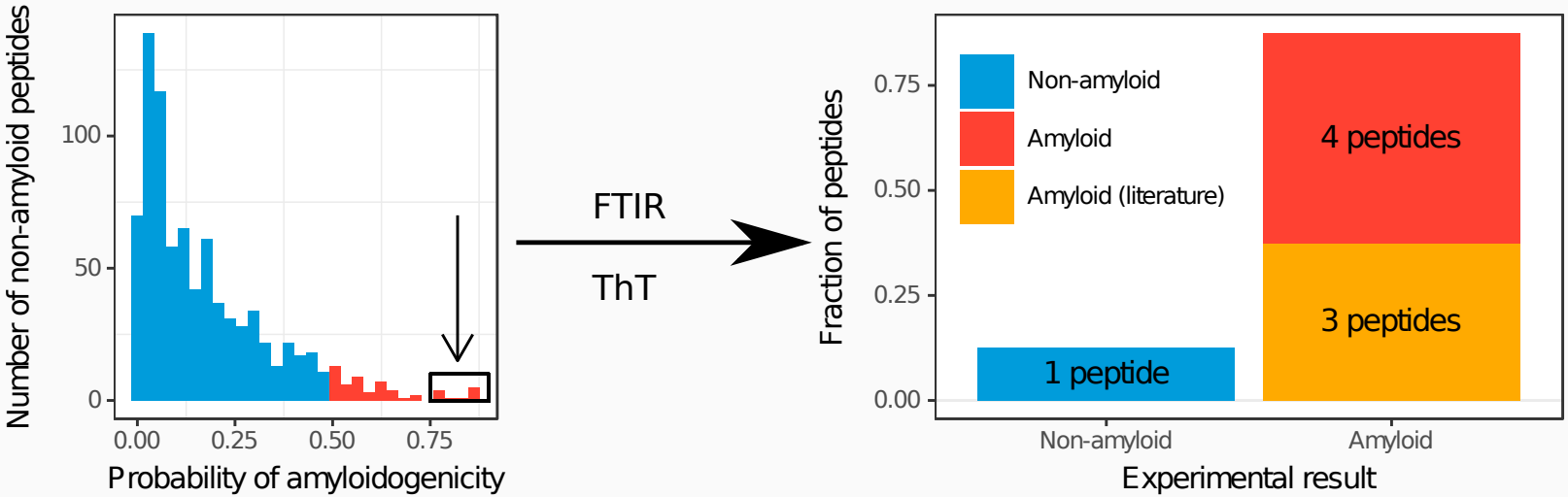
Classifier	AUC	MCC
AmyloGram	**0.8972**	**0.6307**
PASTA 2.0 (Walsh et al., 2014)	0.8550	0.4291
FoldAmyloid (Garbuzynskiy et al., 2010)	0.7351	0.4526
APPNN (Família et al., 2015)	0.8343	0.5823

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



Perspectives and summary

Perspectives

Improved prediction of amyloid proteins:

- hot-spots in the context of the whole protein,
- association of amino acid motifs and amyloidogenicity.

Goal: proteome-wide *in silico* detection of amyloid proteins.

Limitations: very few proteins with known aggregation-prone regions.

AmyPro: a database of amyloid proteins.

- 143 proteins,
- 40719 residues,
- 174 aggregation-prone regions,
- 5645 residues (13.86%) in aggregation-prone regions.

Perspectives

Seeding and cross-seeding: families of hot spots and relaxed seeding specificity.

- CsgA, CsgB, α-synuclein,
- FapC and FapB.

Goal: identification of potential cross-seeding proteins in human microbiome.

Perspectives

Hot-spot specific inhibitors of amyloidogenicity (CsgC, TTR).

Goal: co-evolution of amyloid and its inhibitor.

Availability

Software packages:

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

Web servers:

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

Summary

1. Created a new, accurate predictor of amyloids (Burdukiewicz et al., 2017).
2. Found a set of amino acid motifs associated with the amyloidogenic properties or the lack of them.
3. Found a simplified alphabet suitable for prediction of amyloid proteins.

Acknowledgments

Mentors:

- **Paweł Mackiewicz (University of Wrocław)**.
- Lars Kaderali (University of Greifswald).
- Małgorzata Kotulska (Wrocław University of Science and Technology).
- Henrik Nielsen (Technical University of Denmark).
- Stefan Rödiger (Brandenburg University of Technology Cottbus-Senftenberg).
- Vytauas Smirnovas (University of Vilnus).

Acknowledgments

Peers:

- Agata Błaszczyńska (Wrocław University of Science and Technology).
- Anna Duda-Madej (Wrocław Medical University).
- Marlena Gasior-Głogowska (Wrocław University of Science and Technology).
- 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 (grants 2015/17/N/NZ2/01845 and 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.

Summary

1. Created AmyloGram, a new, accurate predictor of amyloids (Burdukiewicz et al., 2017).
2. Found a set of amino acid motifs associated with the amyloidogenic properties or the lack of them.
3. Found a simplified alphabet suitable for prediction of amyloid proteins.

AmyloGram:

`http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/`.

References

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

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.

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.

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.

References III

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.

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

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.

Wozniak, P. P. and Kotulska, M. (2015). AmyLoad: Website dedicated to amyloidogenic protein fragments. *Bioinformatics*, 31(20):3395–3397.