n-gram analysis of biological sequences in R

Michał Burdukiewicz

Introduction

Tools of the trade: R

R is robust high-level programming language for data analysis.

- Controls full workflow: from data collection till report generation.
- Interfaces for high performance computing and data storage systems.
- Easy generation of web servers.

biogram package

biogram: the **R** package for the n-gram analysis of biological sequences.

n-grams: composite features describing the sequence.

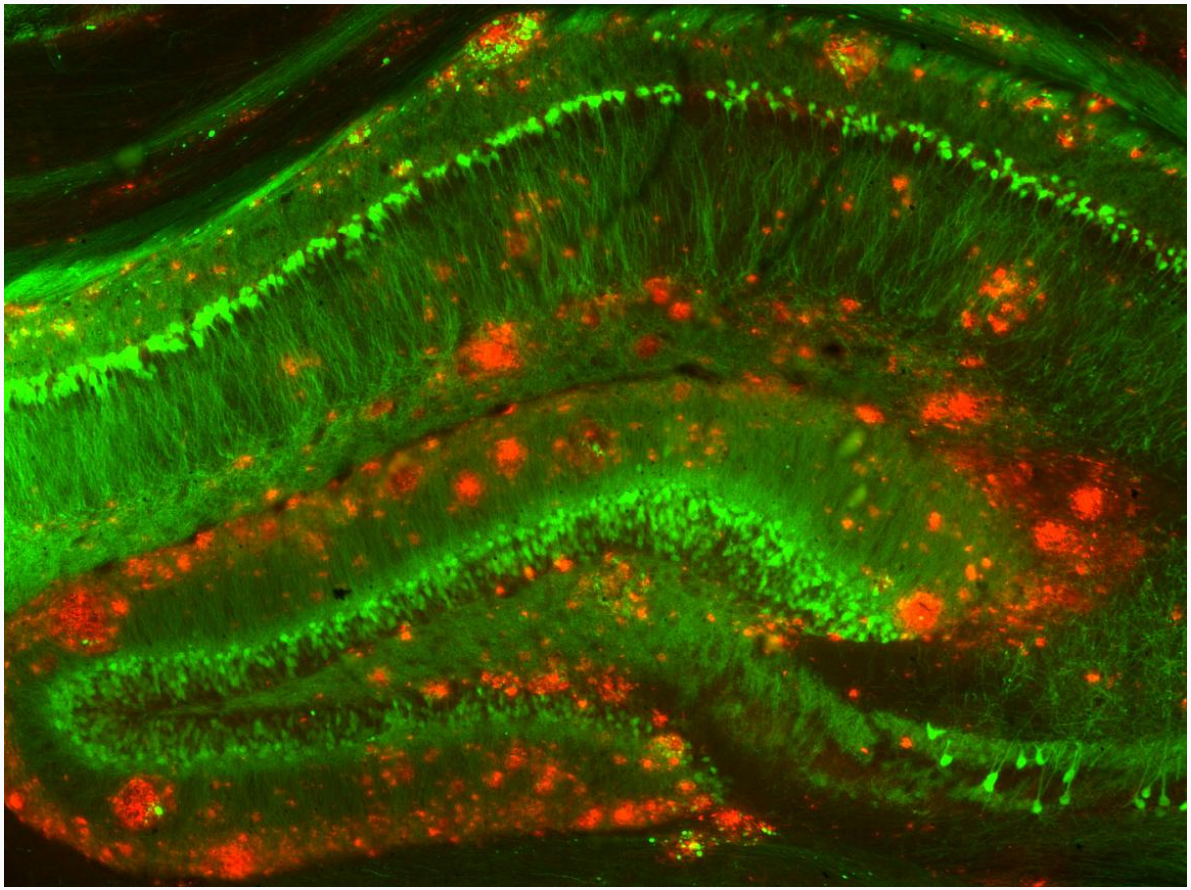
biogram workflow:

1. Extract n-grams.
2. Change composition of n-grams
3. Filter n-grams.

n-grams efficiently and transparently model relationships between protein and its function/structure using only primary sequence.

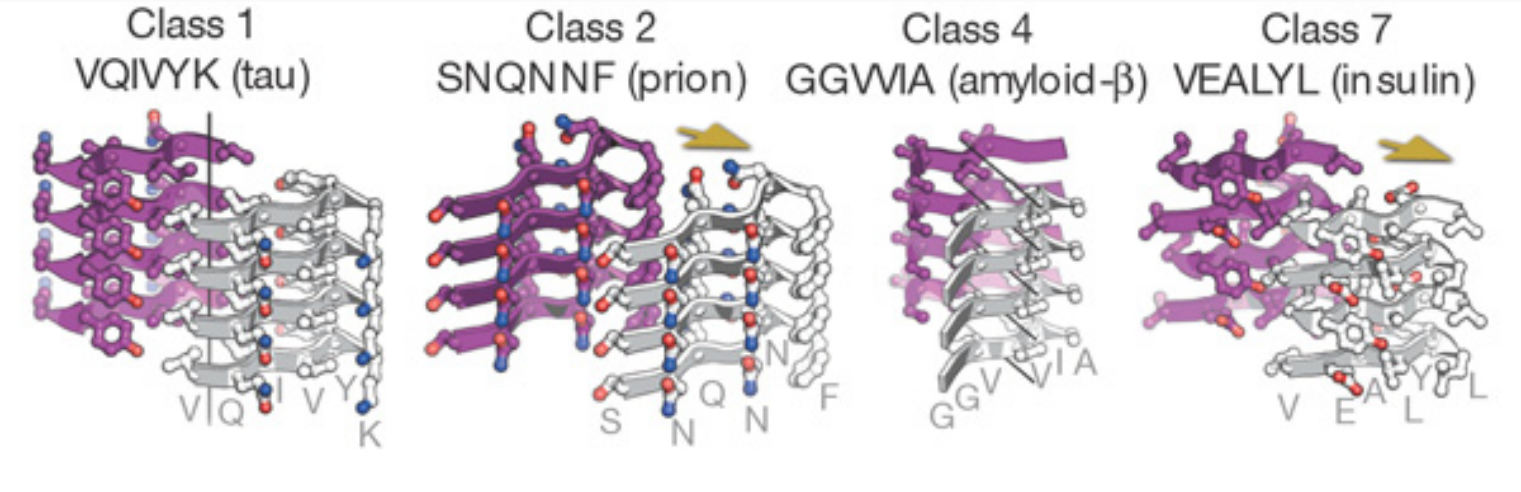
Amyloids

Proteins associated with various neurodegenerative disorders (e.g., Alzheimer's, Parkinson's, Creutzfeldta-Jakob's diseases) creating harmful aggregates.



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

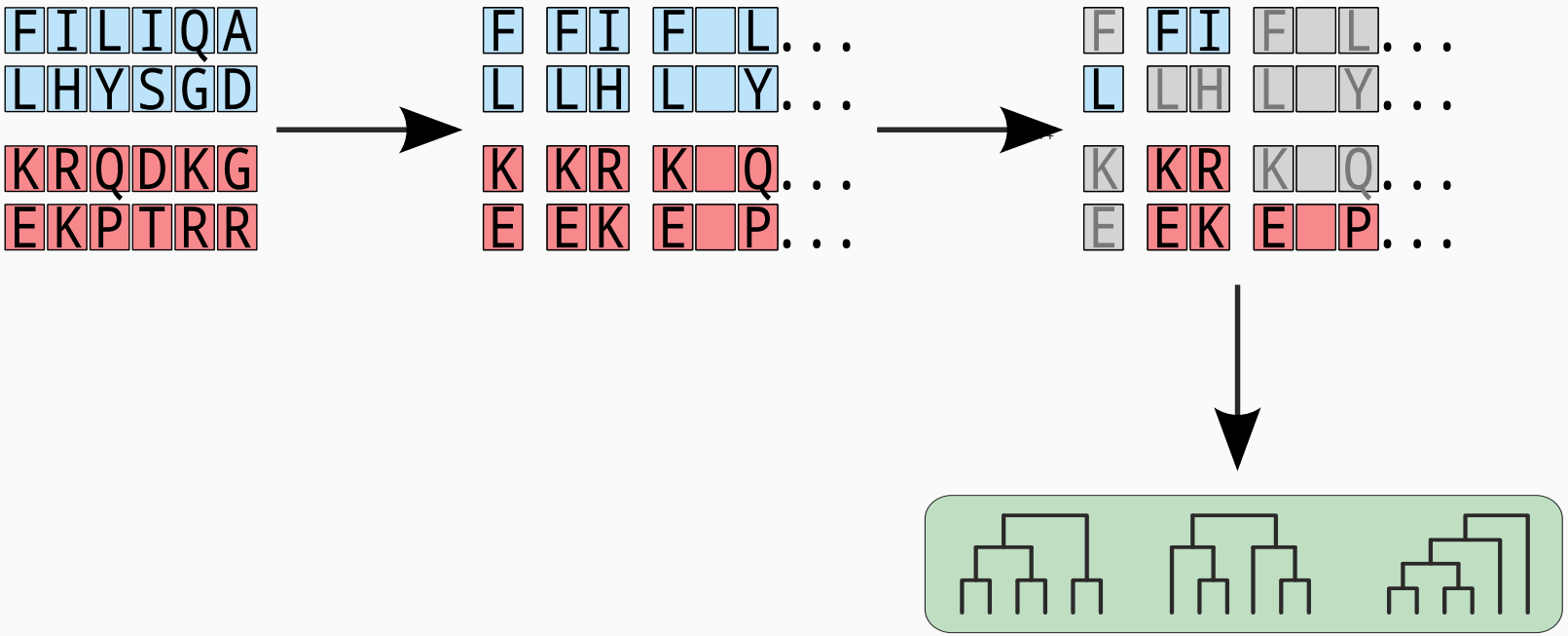
The aggregation of amyloids is initiated by 6- to 15-residue segments called hot spots, diverse subsequences that form unique zipper-like β-structures.



Sawaya et al. (2007)

Amyloidogenic motifs

Which motifs (countinous or gapped subsequences of amino acids) are associated with amyloidogenicity?


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...

n-grams

n-grams (k-tuples) are vectors of n characters derived from input sequence(s).

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

1-grams: M , M , M , M , R , G , A , E

n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

2-grams: MR , MG , MA , ME , RK , GL , AF , ER

n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

3-grams: MRK, MGL, MAF, MER, RKL, GLF, AFG, ERG

n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

2-grams (with a single gap): M-K , M-L , M-F , M-R , R-L , G-F , A-G , E-G

n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
S1	M	R	K	L	Y	C	V	L	L
S2	M	G	L	F	N	I	S	L	L
S3	M	A	F	G	S	L	L	A	F
S4	M	E	R	G	A	G	A	K	L

3-grams (with gaps): M - K - - C , M - L - - I , M - F - - L ,
M - R - - G , R - L - - V , G - F - - S , A - G - - L , E - G - - A

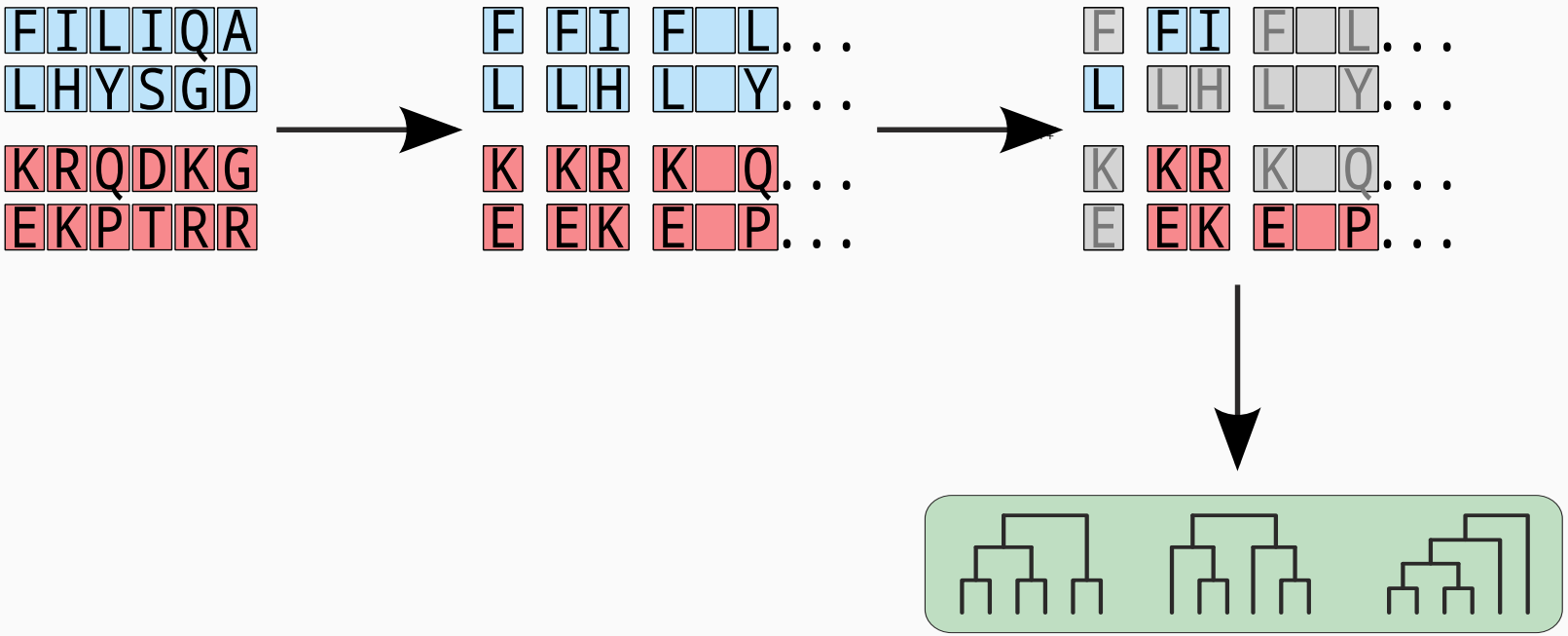
Proposed model

$$\mathrm{y} \sim \text{n-gram}_1 + \text{n-gram}_2 + \ldots + \text{n-gram}_n$$

Problems:

- large number of possible n-grams,
- the majority of n-grams is noninformative.

Filtering n-grams


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...

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 the statistic for permuted data are compared with the value of statistic for original data.

Permutation Test

target	Original data	Permuted data 1	Permuted data 2	...
0	0	1	1	...
0	1	0	1	...
0	1	1	1	...
1	0	0	0	...
1	1	1	0	...
1	0	0	0	...

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

$N_{T_P > T_R}$: 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 computes a probability for a given contigency table providing the exact p-value for the specic value level of the test statistic.

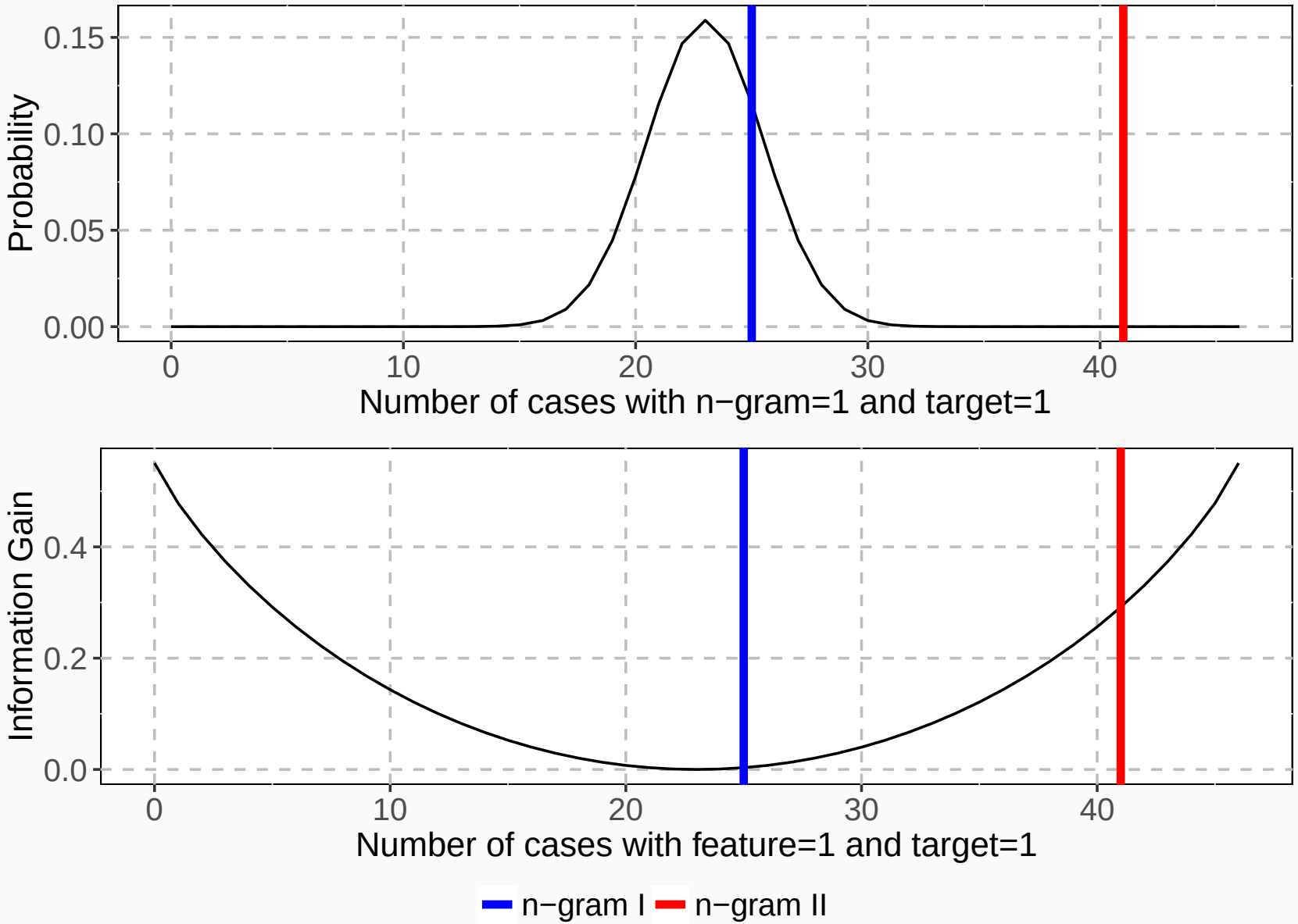
Table 1: IG: 0.0032

Target	n-gram I	Count
0	0	29
1	0	25
0	1	21
1	1	25

Table 2: IG: 0.2917

Target	n-gram II	Count
0	0	45
1	0	9
0	1	5
1	1	41

QuiPT

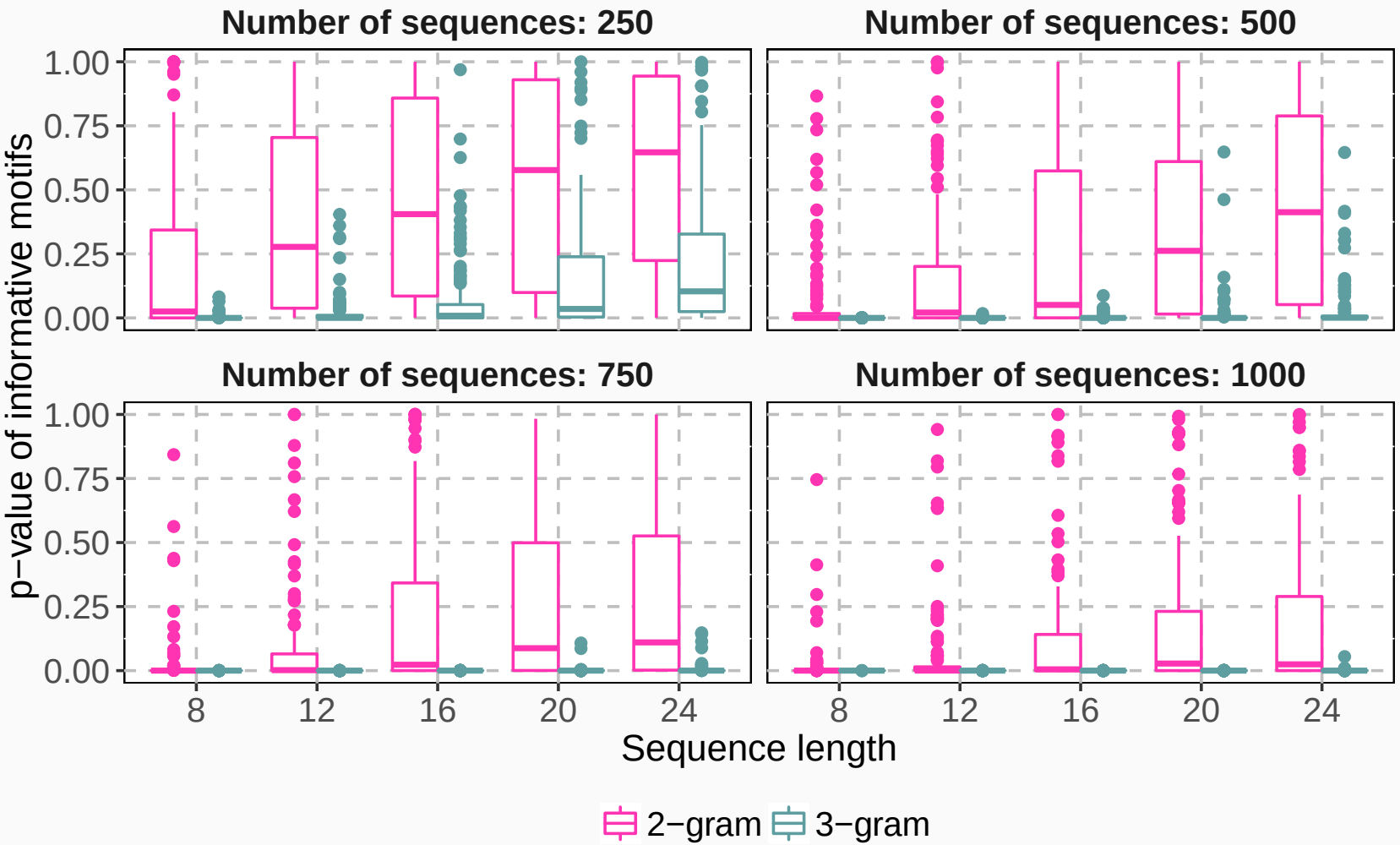

Probability
0.15
0.10
0.05
0.00
0
10
20
30
40
Number of cases with n−gram=1 and target=1
Information Gain
0.4
0.2
0.0
0
10
20
30
40
Number of cases with feature=1 and target=1
n−gram I
n−gram II

Sensitivity of QuiPT

Simulation:

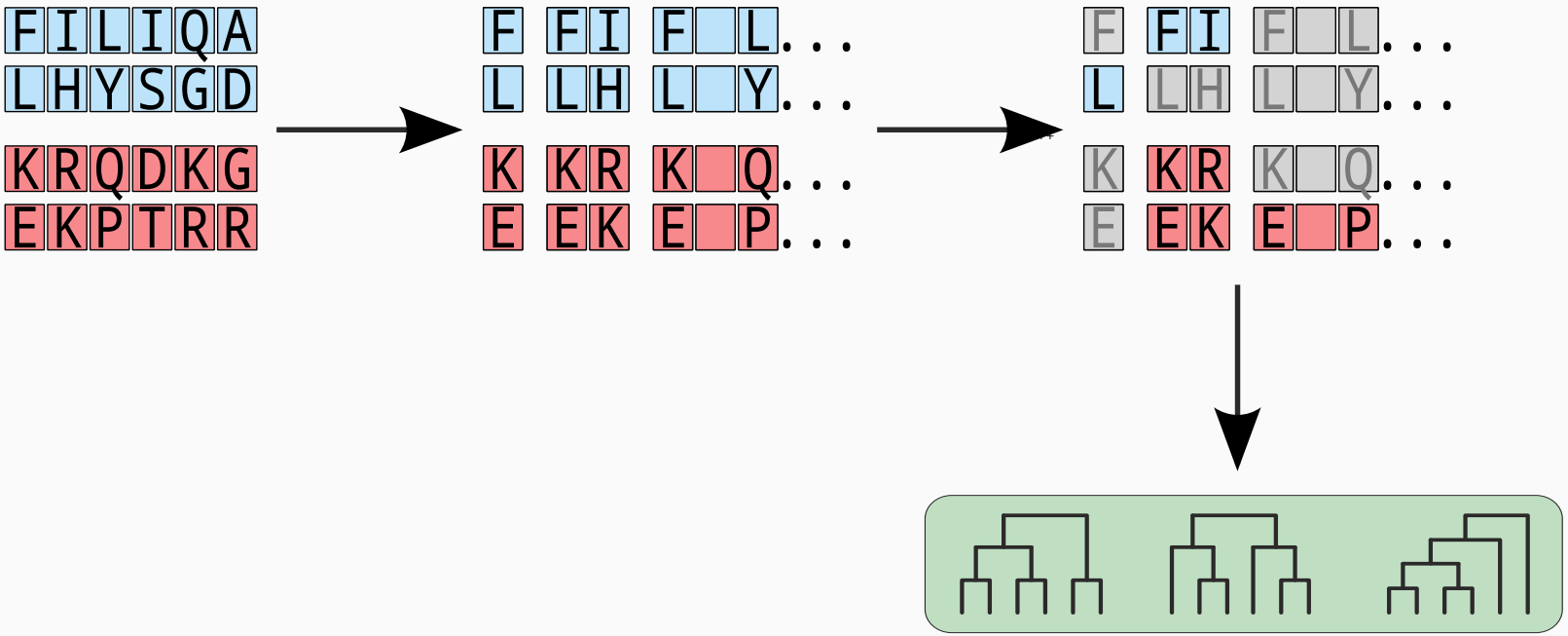
1. Generating random sequences of different lengths.
2. Adding set of informative motifs to the half of sequences.
3. Performing QuiPT.

Sensitivity of QuiPT

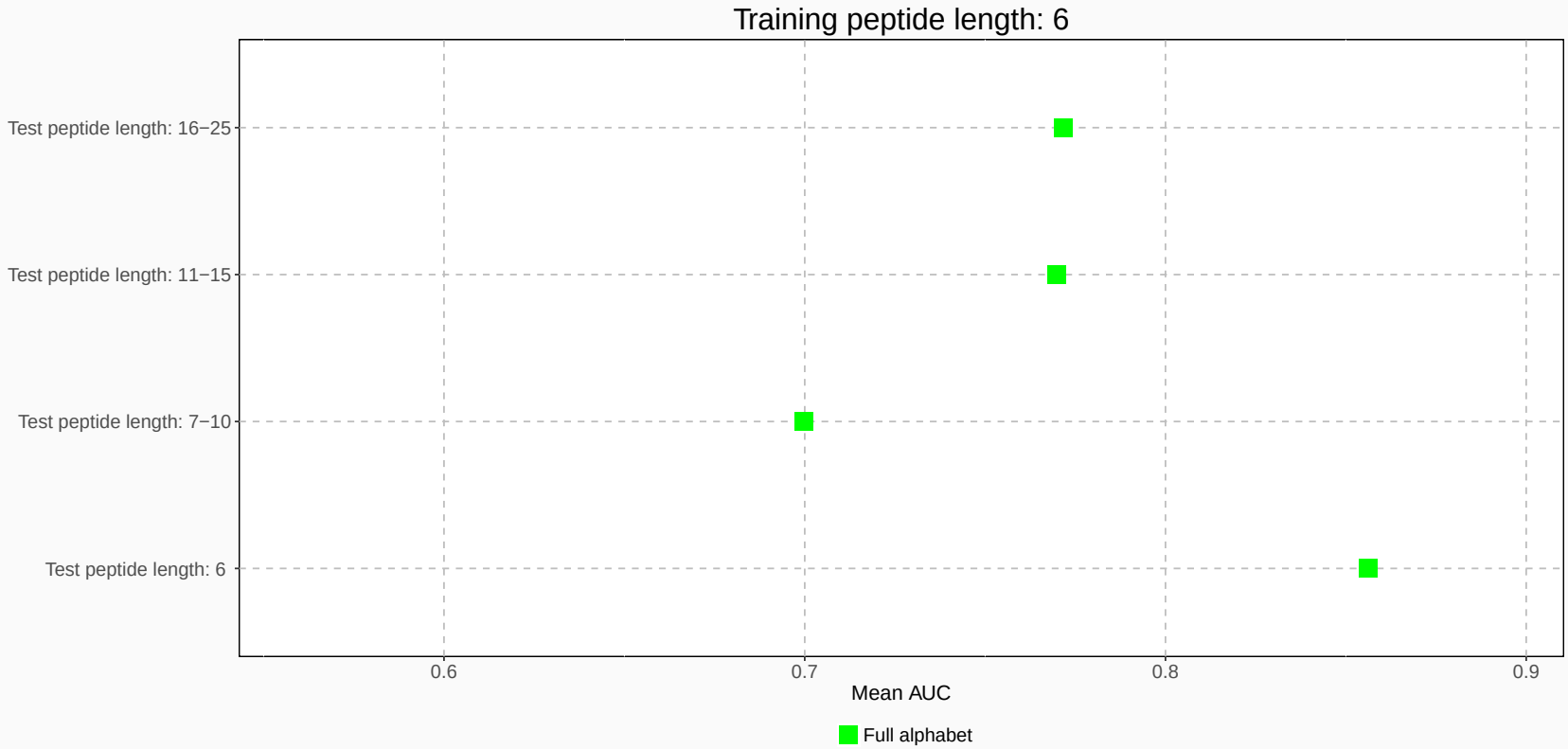


QuiPT works best for large data sets of short sequences.

Validation of models


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



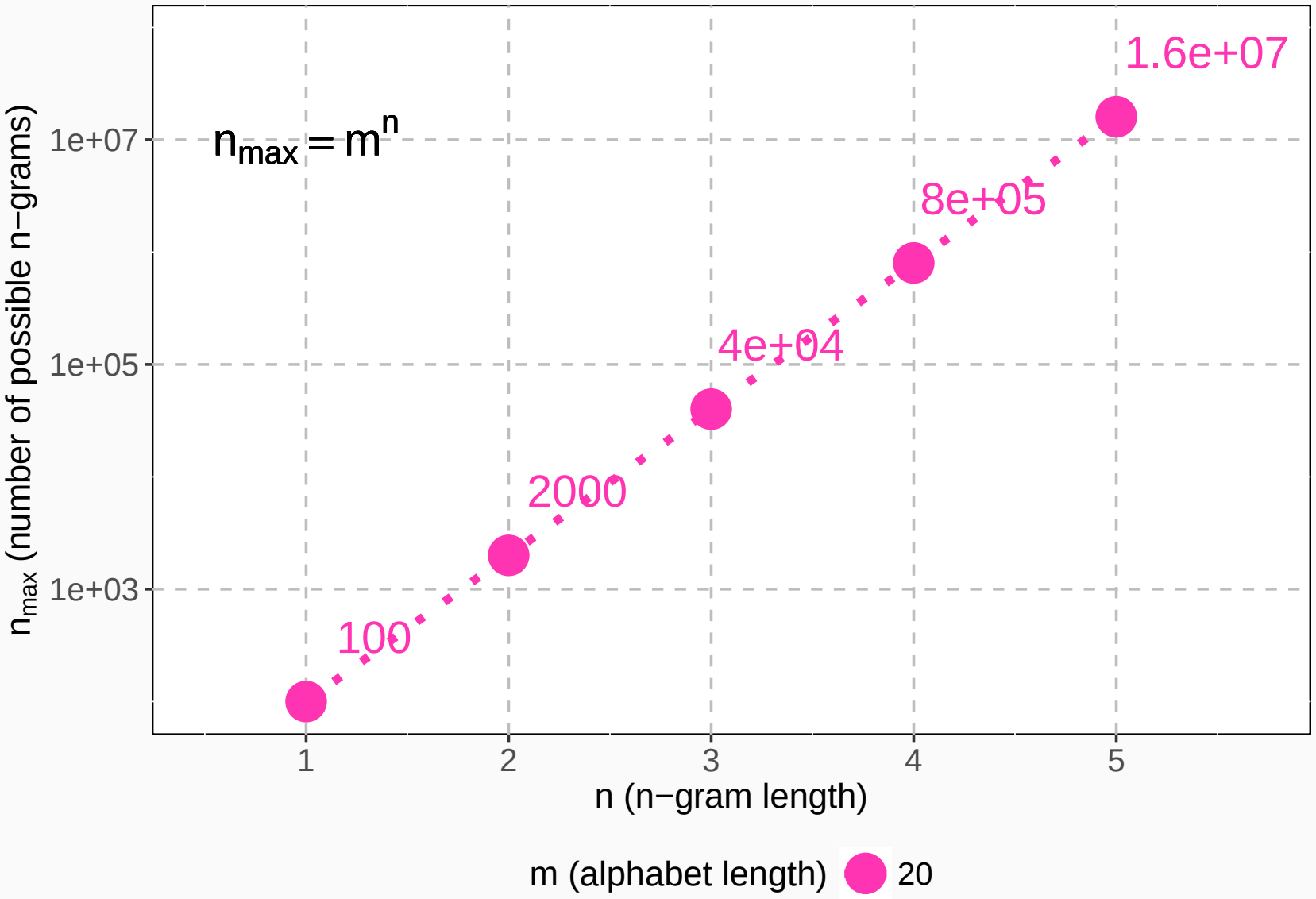
Reduced amino acid alphabets

Does amyloidogenicity depend on the exact sequence of amino acids?

Standard reduced amino acid alphabets

To date, several reduced 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).

n-gram counts using reduced alphabets

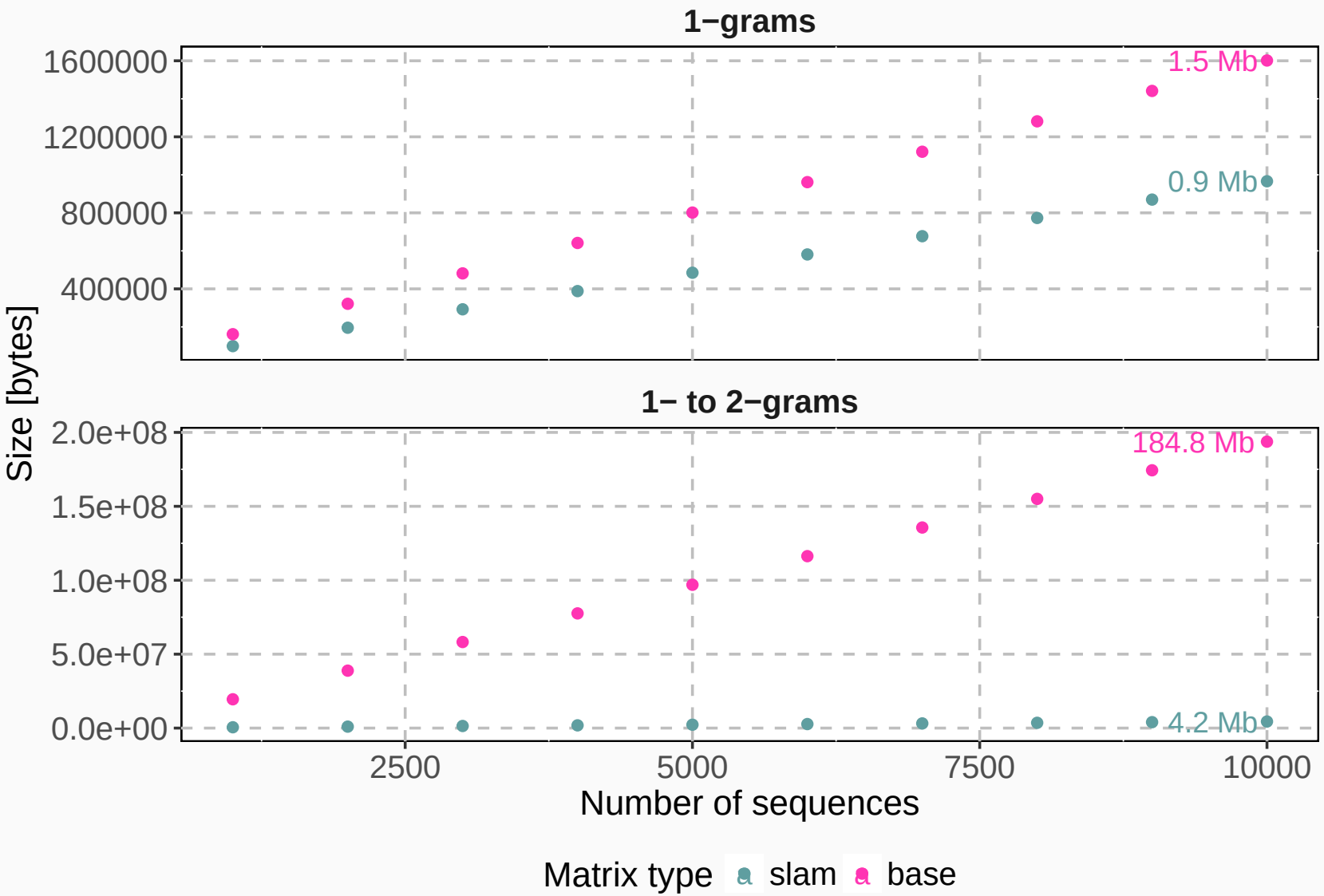


n-gram counts using reduced alphabets

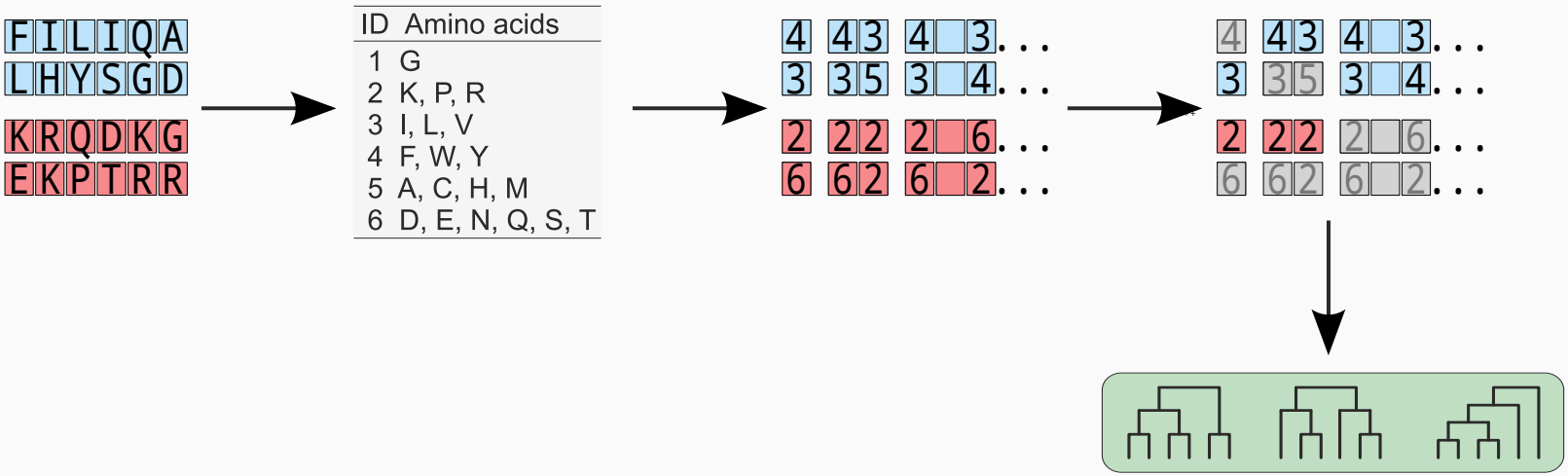
	A_0	C_0	D_0	E_0	F_0	...	L.F_0	M.F_0	N.F_0	...
PF	1	1	0	1	3	...	0	0	1	...
HS	2	2	0	1	2	...	1	0	0	...
SL	9	1	0	1	2	...	0	0	0	...
SP	4	1	1	1	1	...	0	0	0	...

The sparcity of the n-gram count matrix grows with n.

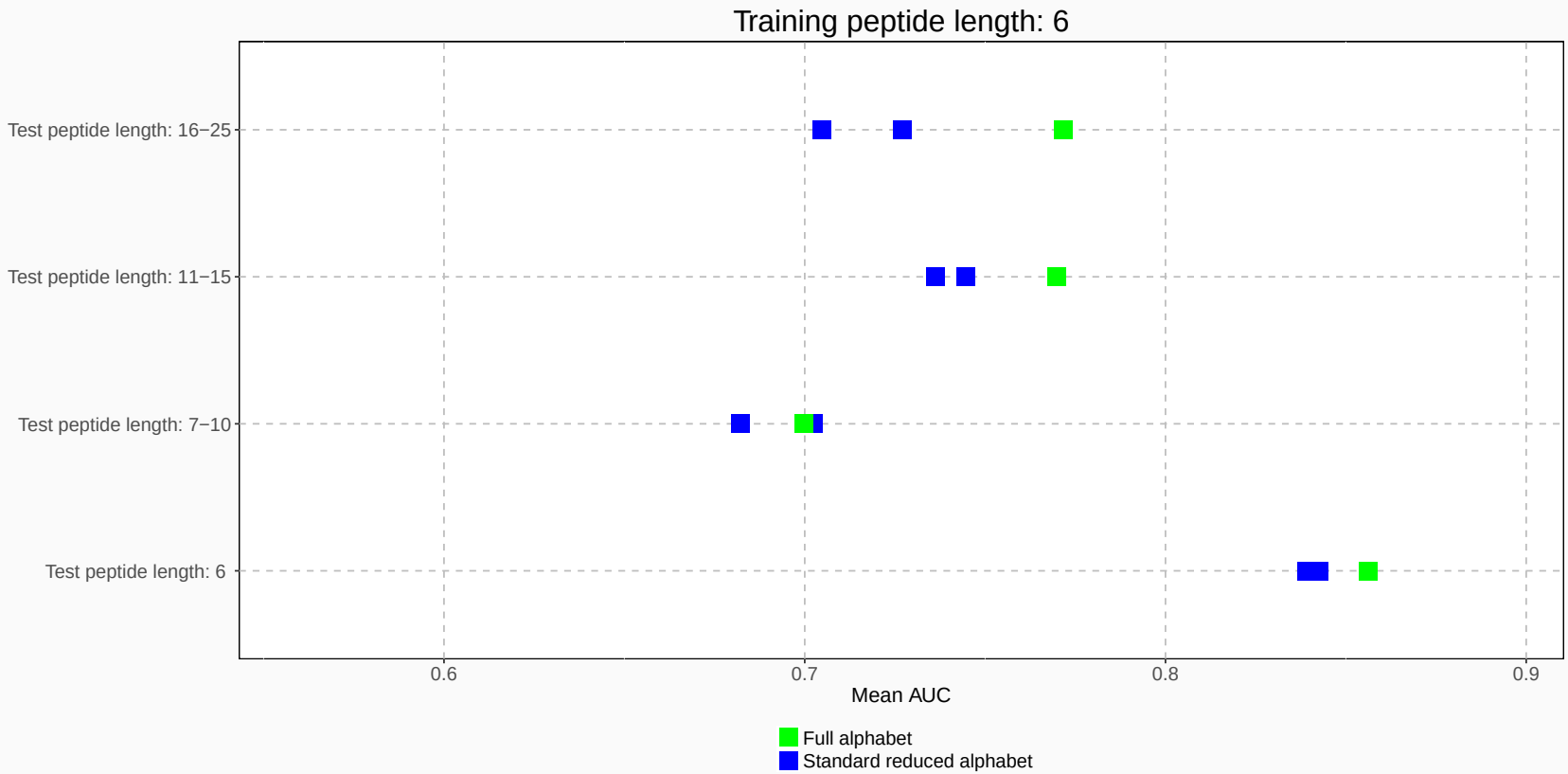
sparse matrix representation



Standard reduced amino acid alphabets



Cross-validation

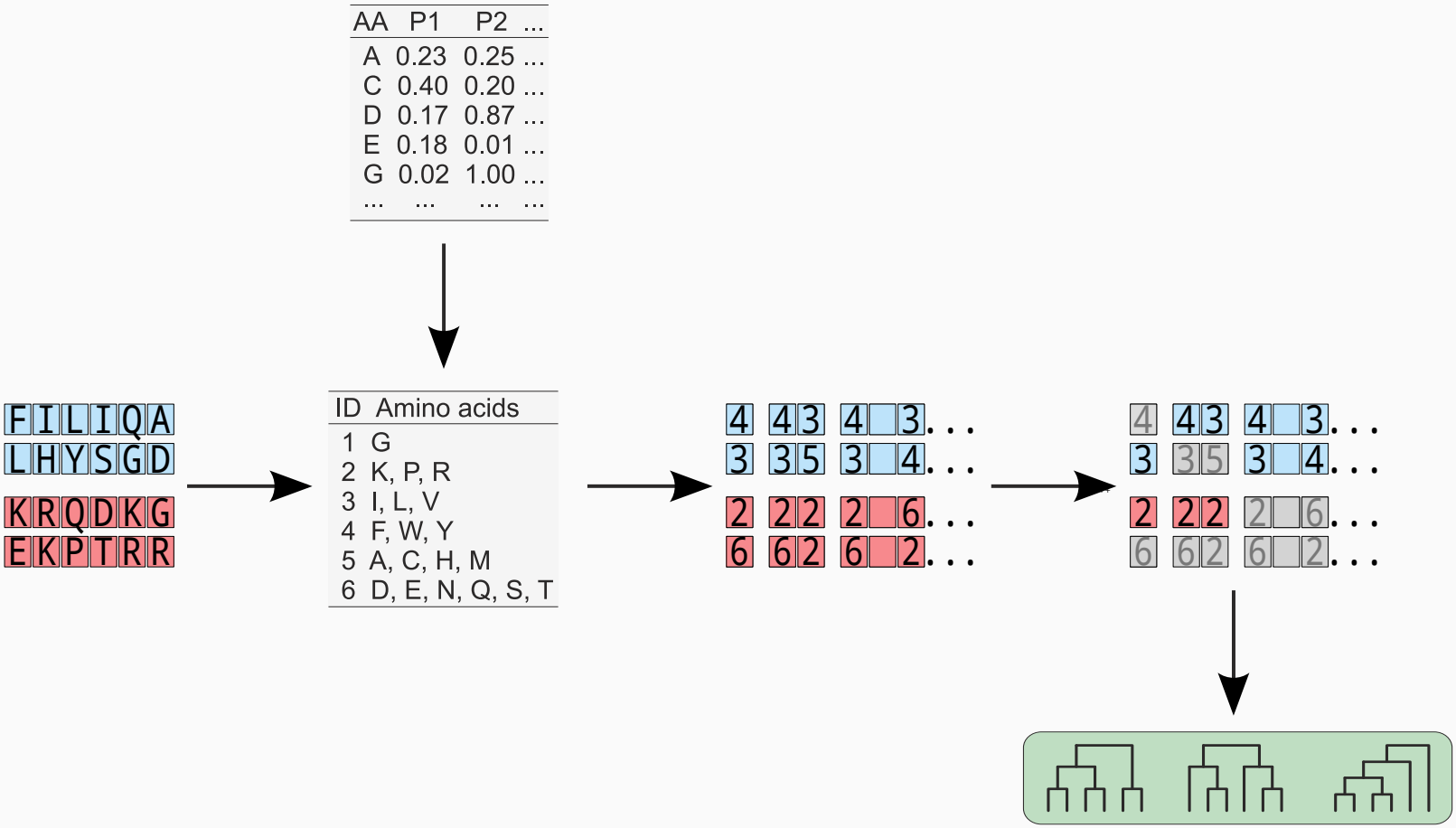


Standard reduced amino acid alphabets do not enhance discrimination between amyloidogenic and non-amyloidogenic proteins.

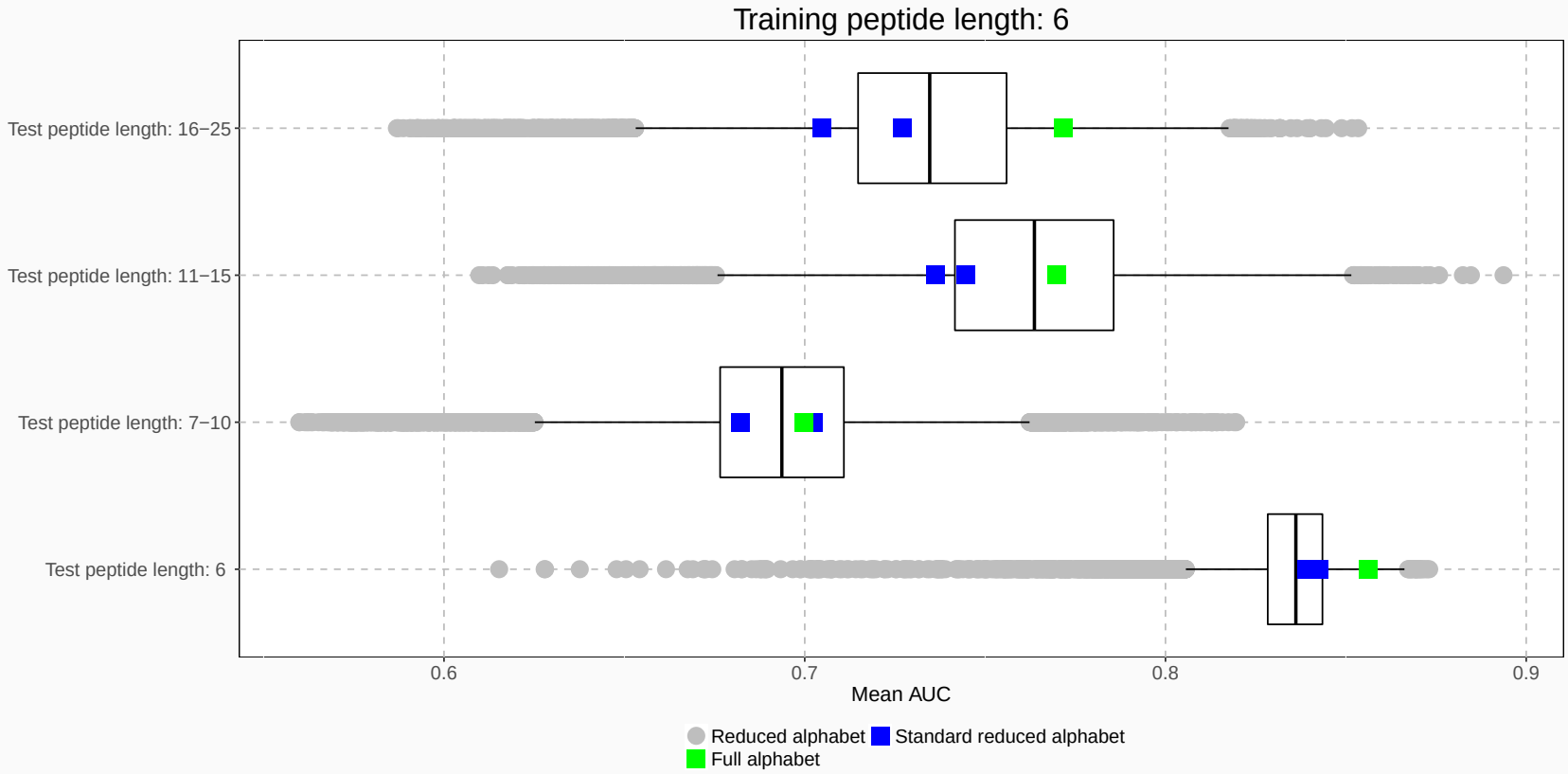
Novel reduced amino acid alphabets

- 17 measures handpicked from AAIndex database:
  - size of residues,
  - hydrophobicity,
  - solvent surface area,
  - frequency in β-sheets,
  - contactivity.
- 524 284 amino acid reduced alphabets with different level of amino acid alphabet reduction (three to six amino acid groups).

Novel reduced amino acid alphabets

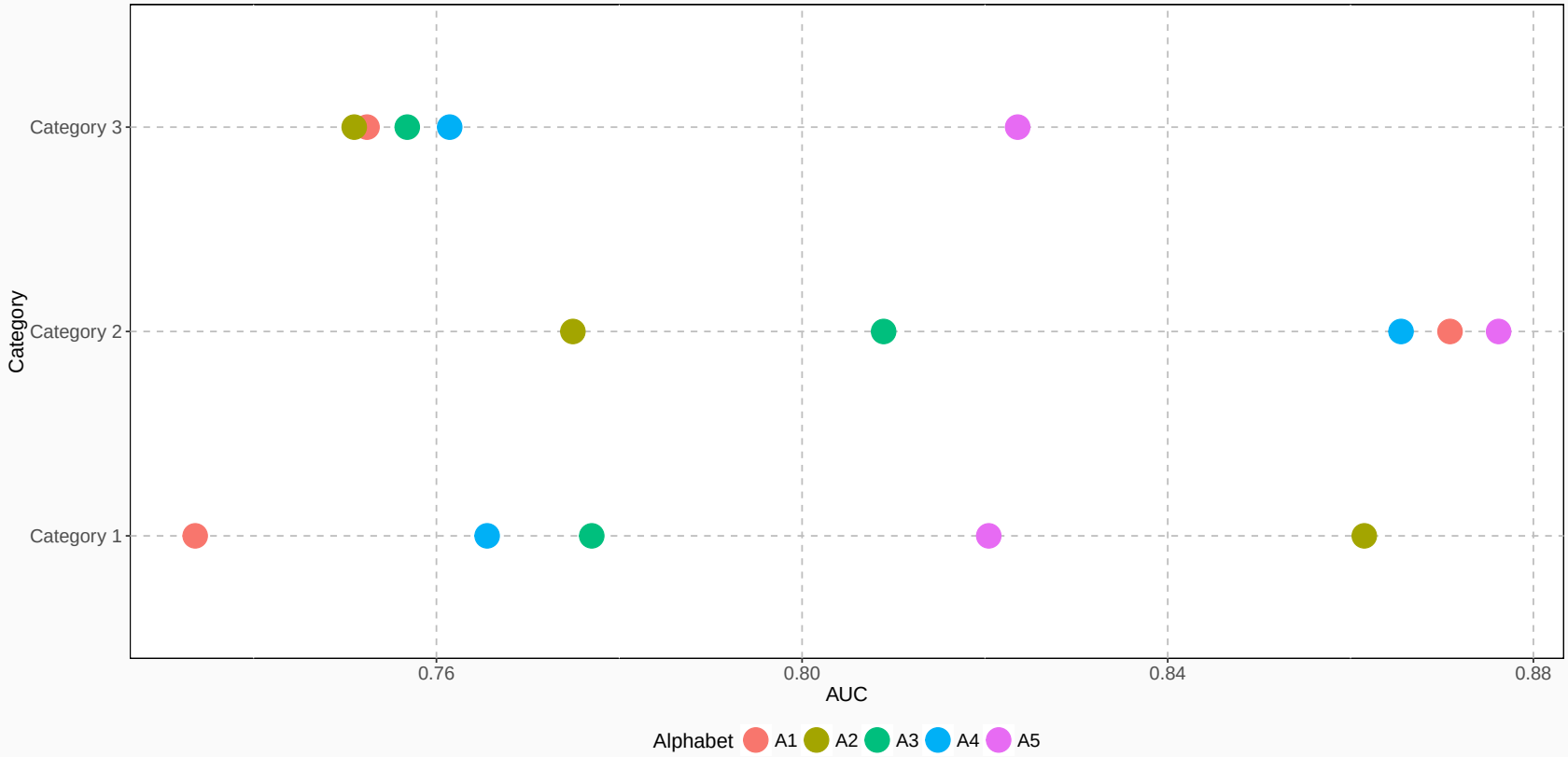


Cross-validation

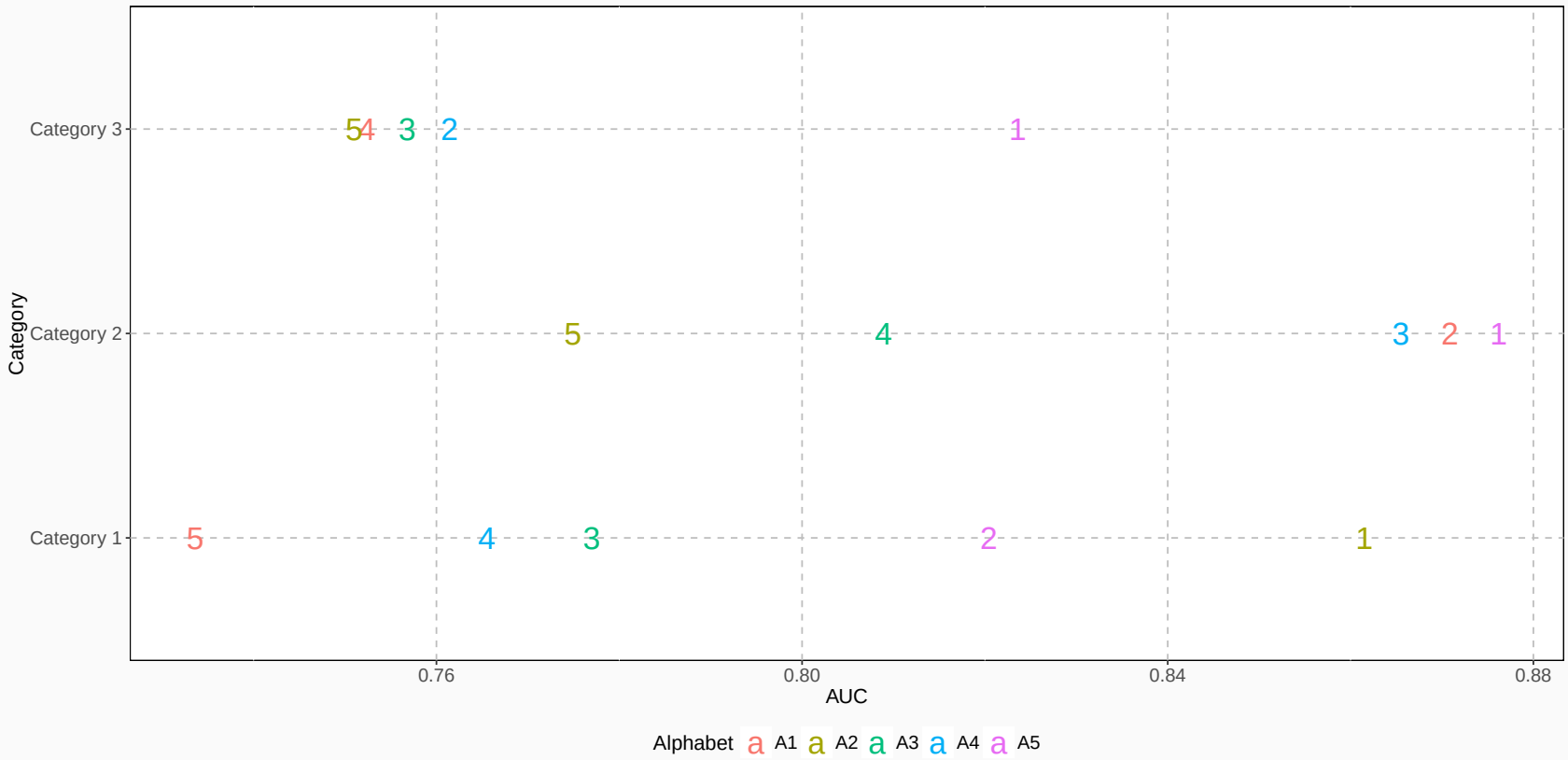


Hinges of boxes correspond to the 0.25 and 0.75 quartiles. The bar inside the box represents the median. The gray circles correspond to the reduced alphabets with the AUC outside the 0.95 confidence interval.

Ranking alphabets

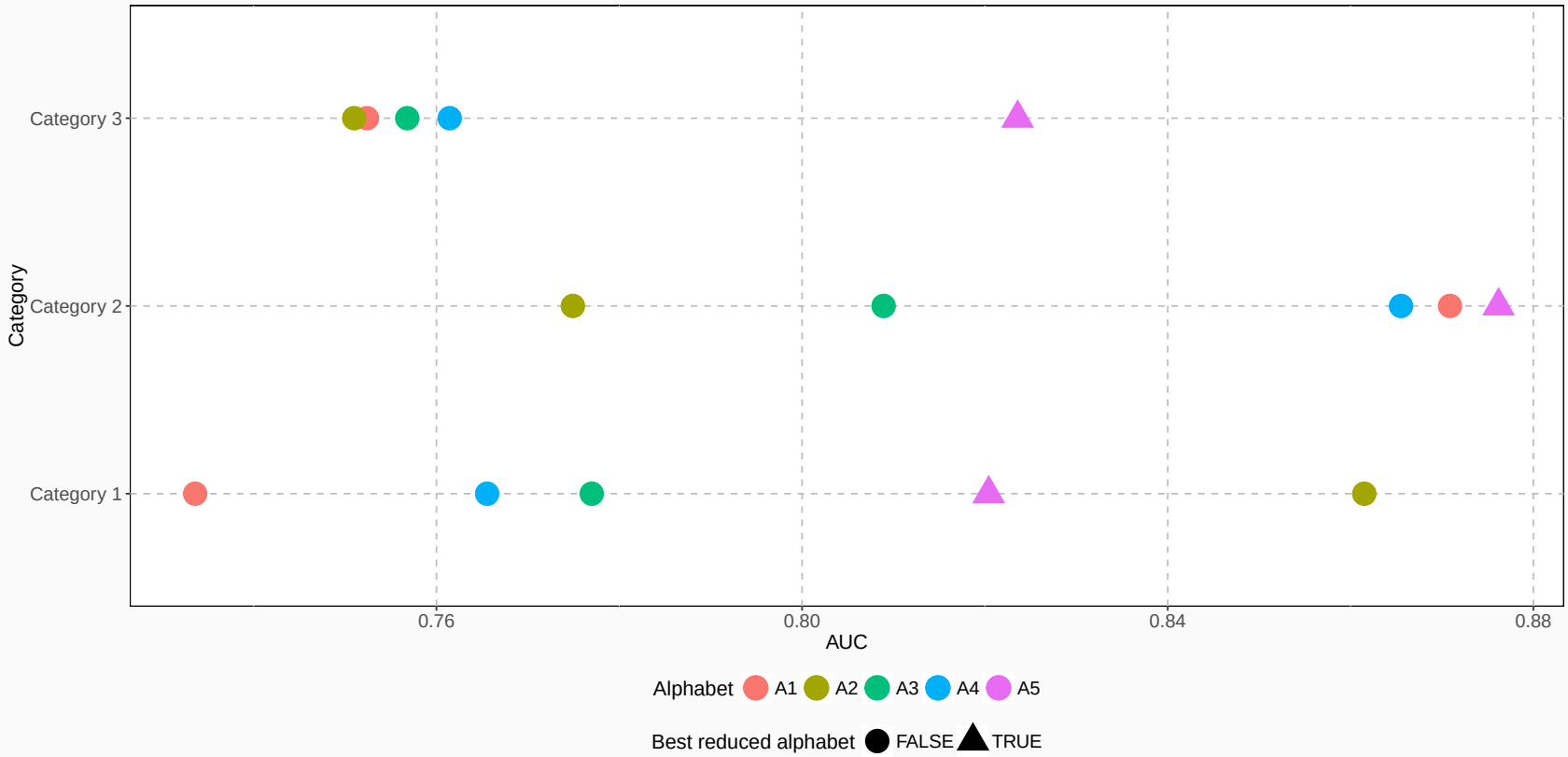


Ranking alphabets



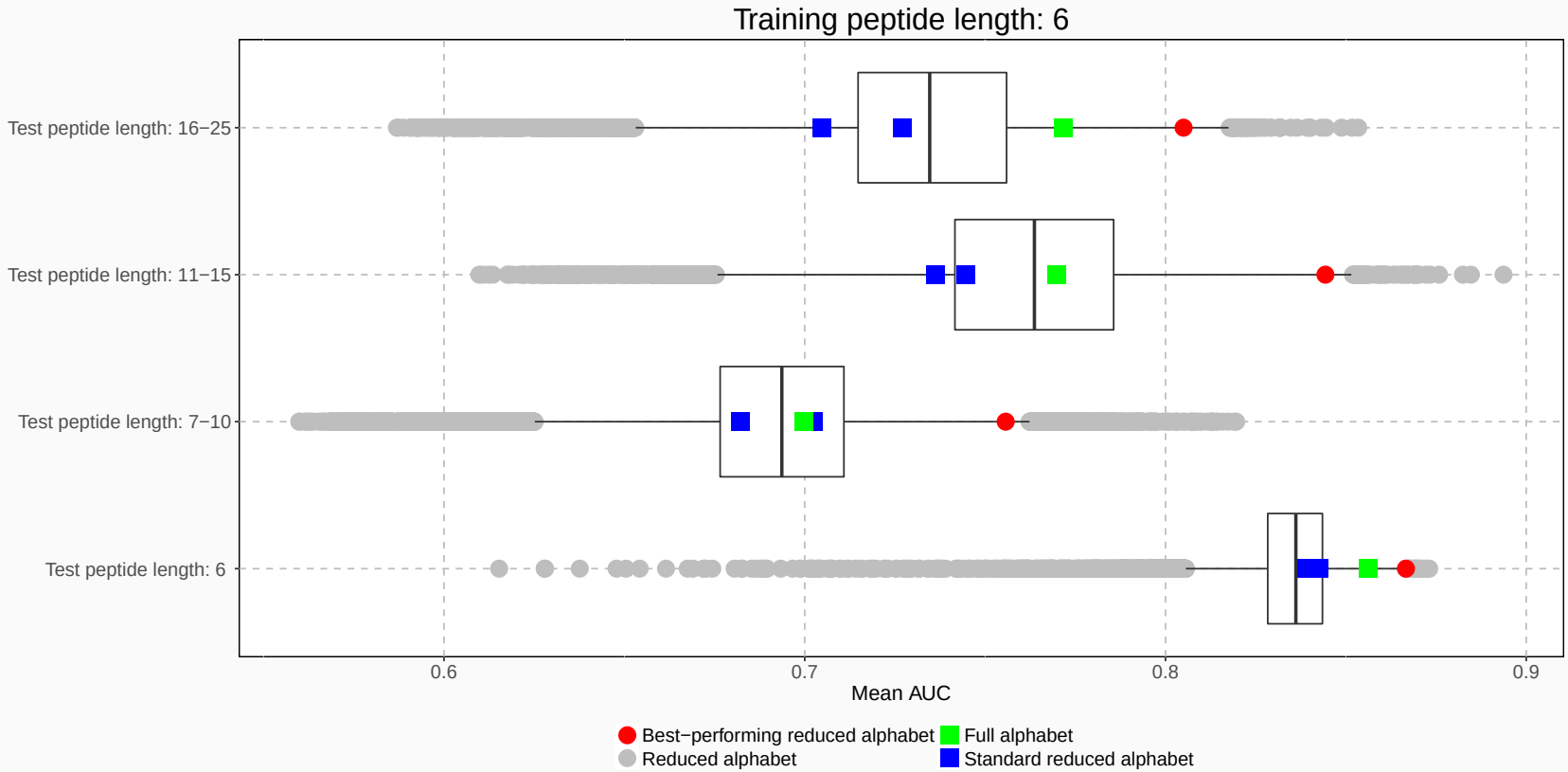
We rank alphabets separately in all length categories assuming the rank 1 for the best AUC, rank 2 for the second best AUC and so on.

Ranking alphabets



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

The best-perfoming reduced alphabet



The best-perfoming reduced 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 reduced 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 reduced alphabet

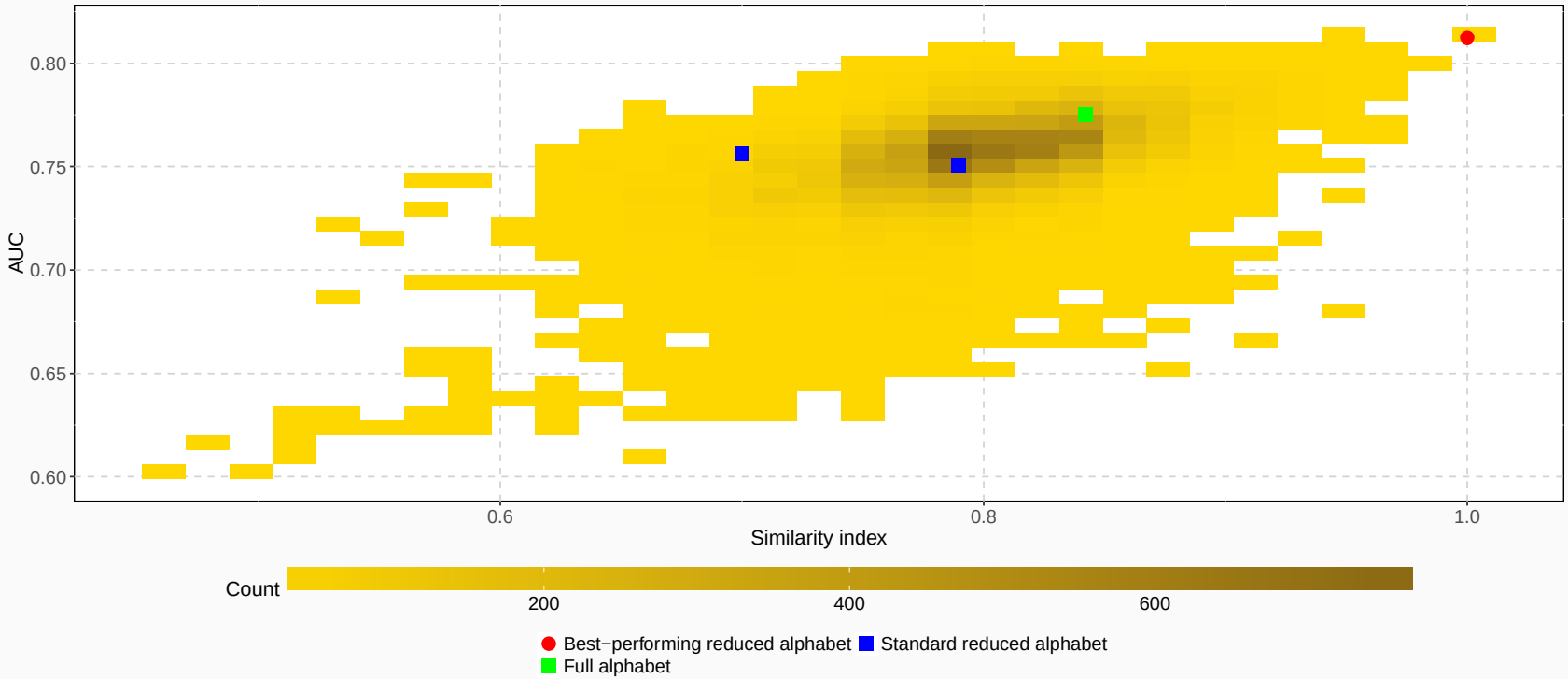
Subgroup ID	Amino acids
1	G
2	K, P, R
3	I, L, V
4	F, W, Y
5	A, C, H, M
6	D, E, N, Q, S, T

Group 2 - charged breakers of β-structures.

Alphabet similarity and performance

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

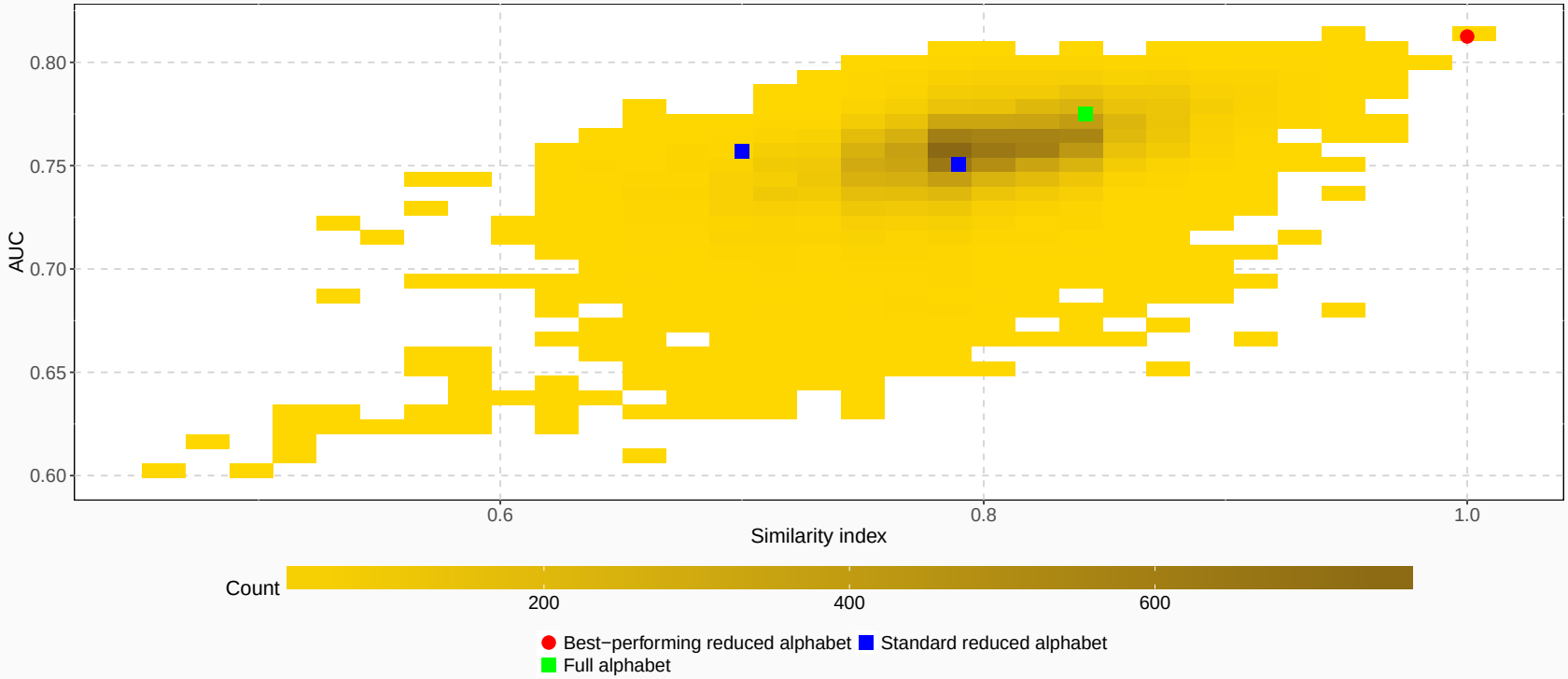
Similarity index



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

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

Similarity index

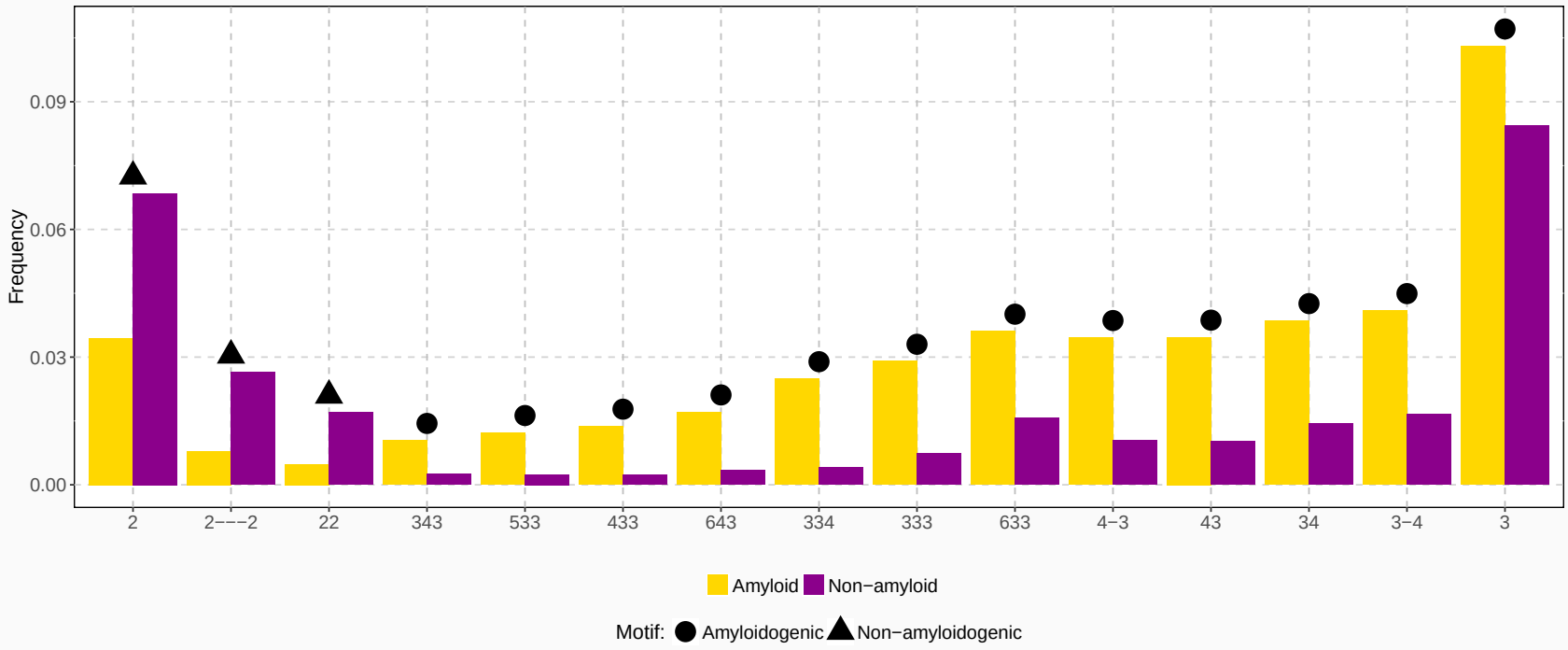


The correlation between mean AUC an similarity index is significant ($\text{p-value} \leq 2.2^{-16}$; $\rho = 0.51$).

Knowledge-discovery

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 and summary

Is performance of the AmyloGram, the classifier based on the best-performing reduced amino acid alphabet, also adequate on the independent dataset?

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*.

Summary

We identified a group of reduced amino acid alphabets which capture properties of amyloids.

Our algorithm was also capable of extracting n-gram associated with amyloidogenicity, partially confirming experimental results.

Our software is available as a web-server: `http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/`.

Summary

biogram: the **R** package for the n-gram analysis of biological sequences.

`https://CRAN.R-project.org/package=biogram` (1.3)
`http://github.com/michbur/biogram` (1.4)

Acknowledgements and funding

This research was partially funded by the KNOW Consortium and National Science Center (2015/17/N/NZ2/01845).

- Małgorzata Kotulska.
- Paweł Mackiewicz,
- Stefan Rödiger,
- **biogram** package (`https://cran.r-project.org/package=biogram`):
  - Piotr Sobczyk,
  - Chris Lauber,
- **AmyLoad** database (`comprec-lin.iiar.pwr.edu.pl/amyload`):
  - Paweł Woźniak,

References

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.

References II

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.

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 III

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*, page gku399.