AmyloGram: Analysis of proteins in R

Jarek Chilimoniuk

Department of Bioinformatics and Genomics, University of Wrocław

Presentation plan

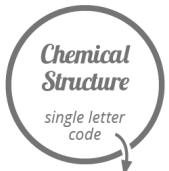
1. Aminoacids and proteins
2. n-grams and simplified alphabets
3. Amyloid prediction
4. Shiny aplication

Aminoacids and proteins

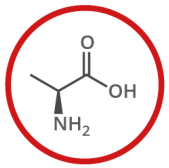
A GUIDE TO THE TWENTY COMMON AMINO ACIDS

AMINO ACIDS ARE THE BUILDING BLOCKS OF PROTEINS IN LIVING ORGANISMS. THERE ARE OVER 500 AMINO ACIDS FOUND IN NATURE - HOWEVER, THE HUMAN GENETIC CODE ONLY DIRECTLY ENCODES 20. 'ESSENTIAL' AMINO ACIDS MUST BE OBTAINED FROM THE DIET, WHILST NON-ESSENTIAL AMINO ACIDS CAN BE SYNTHESISED IN THE BODY.

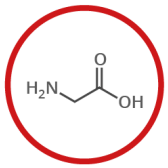
Chart Key: ALIPHATIC · AROMATIC · ACIDIC · BASIC · HYDROXYLIC · SULFUR-CONTAINING · AMIDIC · NON-ESSENTIAL · ESSENTIAL



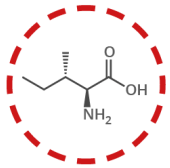
NAME (A)
three letter code
DNA codons



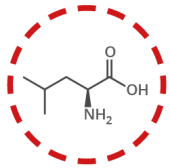
ALANINE (A)
Ala
GCT, GCC, GCA, GCG



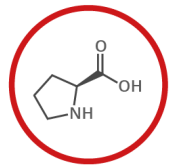
GLYCINE (G)
Gly
GGT, GGC, GGA, GGG



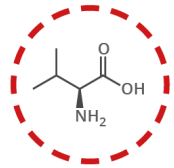
ISOLEUCINE (I)
Ile
ATT, ATC, ATA



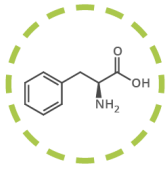
LEUCINE (L)
Leu
CTT, CTC, CTA, CTG, TTA, TTG



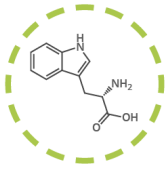
PROLINE (P)
Pro
CCT, CCC, CCA, CCG



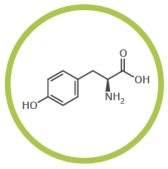
VALINE (V)
Val
GTT, GTC, GTA, GTG



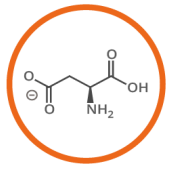
PHENYLALANINE (F)
Phe
TTT, TTC



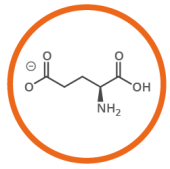
TRYPTOPHAN (W)
Trp
TGG



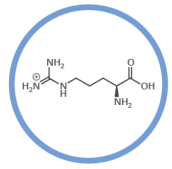
TYROSINE (Y)
Tyr
TAT, TAC



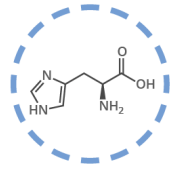
ASPARTIC ACID (D)
Asp
GAT, GAC



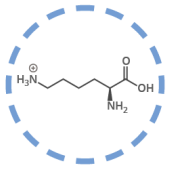
GLUTAMIC ACID (E)
Glu
GAA, GAG



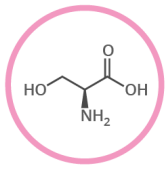
ARGININE (R)
Arg
CGT, CGC, CGA, CGG, AGA, AGG



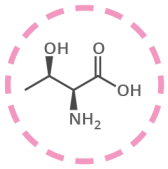
HISTIDINE (H)
His
CAT, CAC



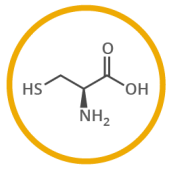
LYSINE (K)
Lys
AAA, AAG



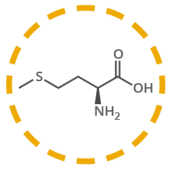
SERINE (S)
Ser
TCT, TCC, TCA, TCG, AGT, AGC



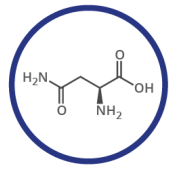
THREONINE (T)
Thr
ACT, ACC, ACA, ACG



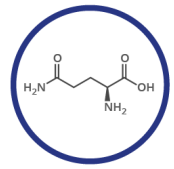
CYSTEINE (C)
Cys
TGT, TGC



METHIONINE (M)
Met
ATG



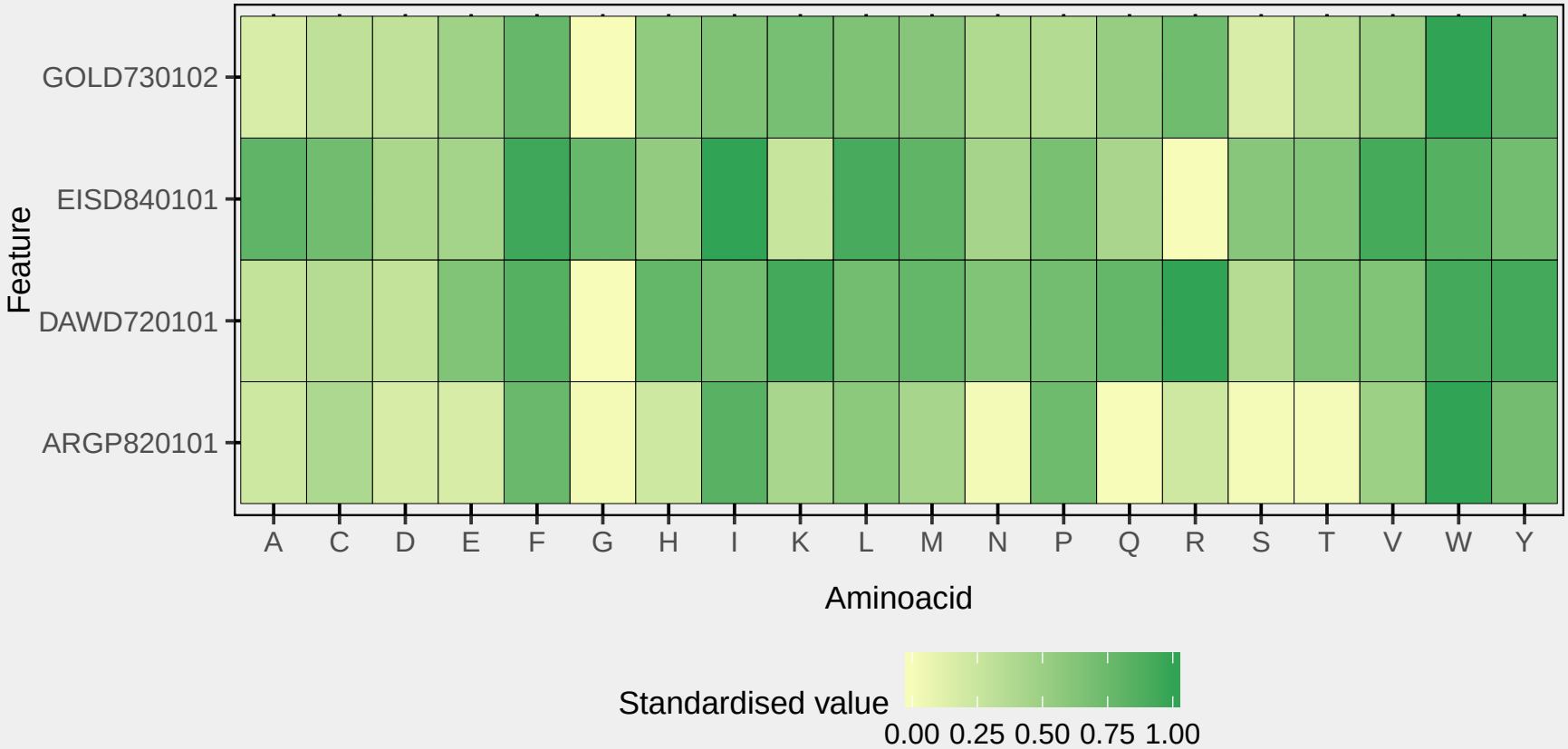
ASPARAGINE (N)
Asn
AAT, AAC



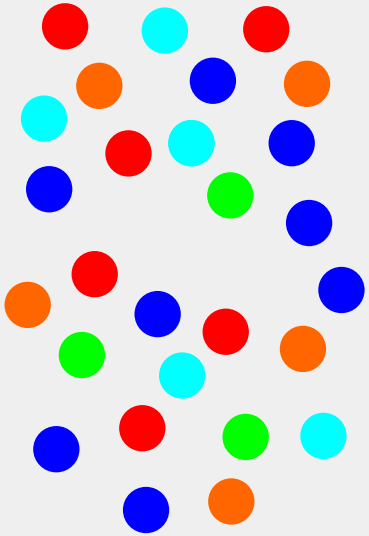
GLUTAMINE (Q)
Gln
CAA, CAG

Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (Z) are respectively used.

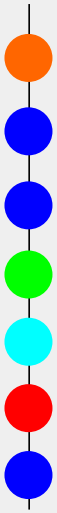
Aminoacids


GOLD730102
EISD840101
DAWD720101
ARGP820101
Feature
A C D E F G H I K L M N P Q R S T V W Y
Aminoacid
Standardised value
0.00 0.25 0.50 0.75 1.00

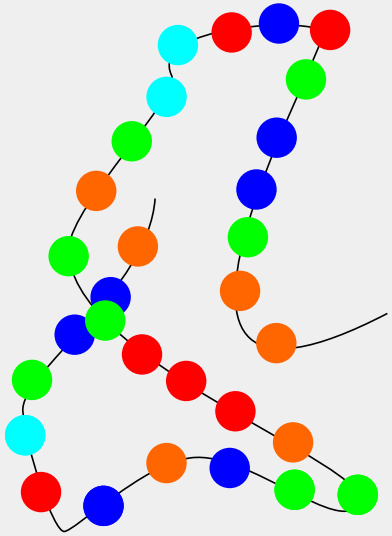
Proteins



Aminoacids

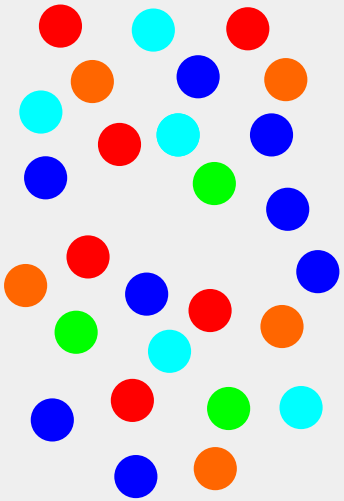


Protein (primary structure)

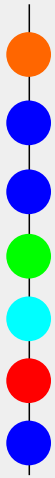


Protein (higher order structures)

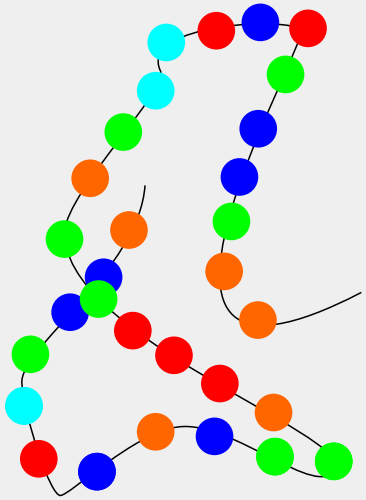
Proteins



Aminoacids



Protein (primary structure)

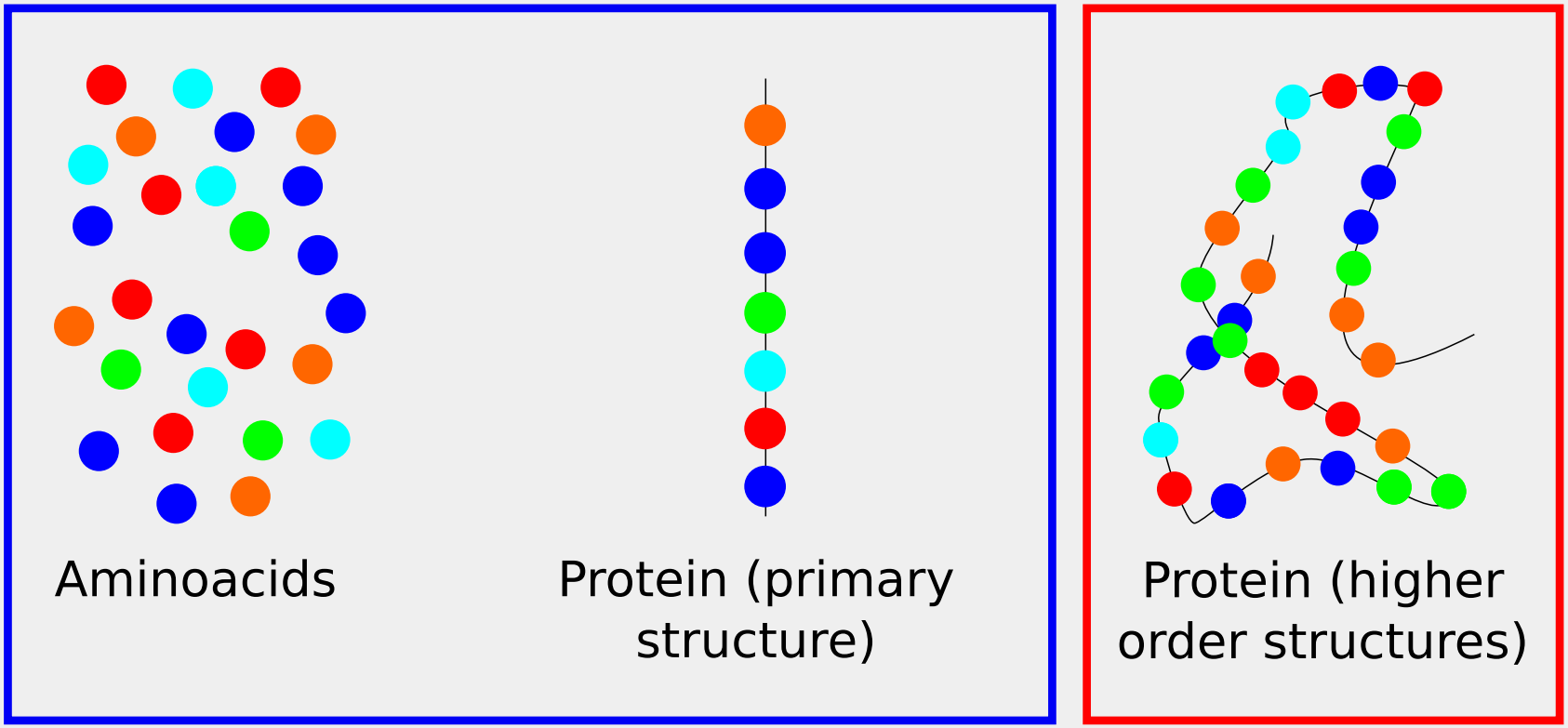


Protein (higher order structures)

KNOWN

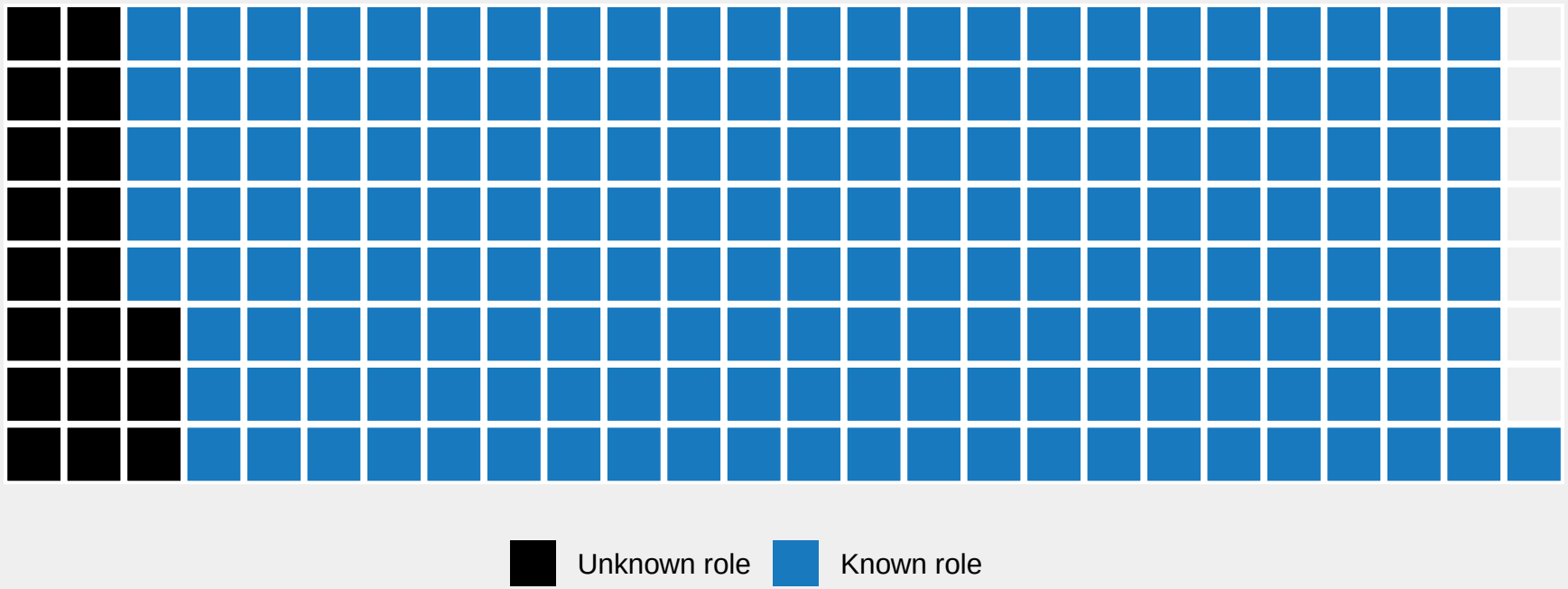
UNKNOWN

Proteins



Protein higher order structures determines its function.

Human proteom



1937 human proteins have unknown role (dark proteome) (Young-Ki Paik et al., 2018).

Goal

Development of methods for predicting protein properties on the basis of their primary structure in a way that is understandable for biologists and experimentally validated.

n-grams and simplified alphabets

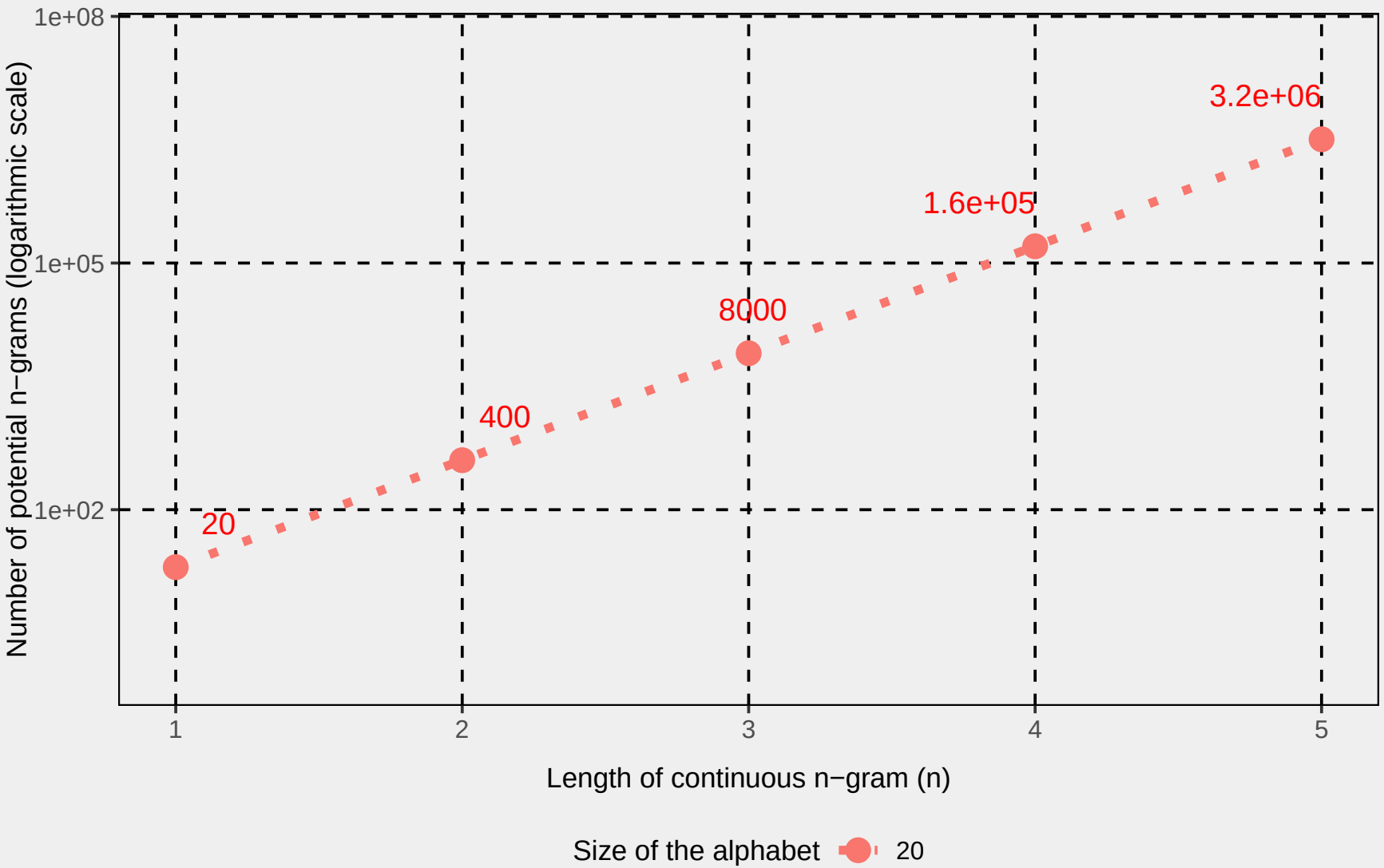
n-grams (k-tuple, k-mers):

- subsequences (continuous or discontinuous) *n* aminoacid or nucleotide residues,
- more informative than the individual residues.

Peptide I: FKVWPDHGSG
Peptide II: YMCIYRAQTN

n-gram examples from peptide I and II:

1. 1-gram: F, Y, K, M,
2. 2-gram: FK, YM, KV, MC,
3. 2-gram (discontinuous): F-V, Y-C, K-W, M-I,
4. 3-gram (discontinuous): F–WP, Y–IY, K–PD, M–YR.



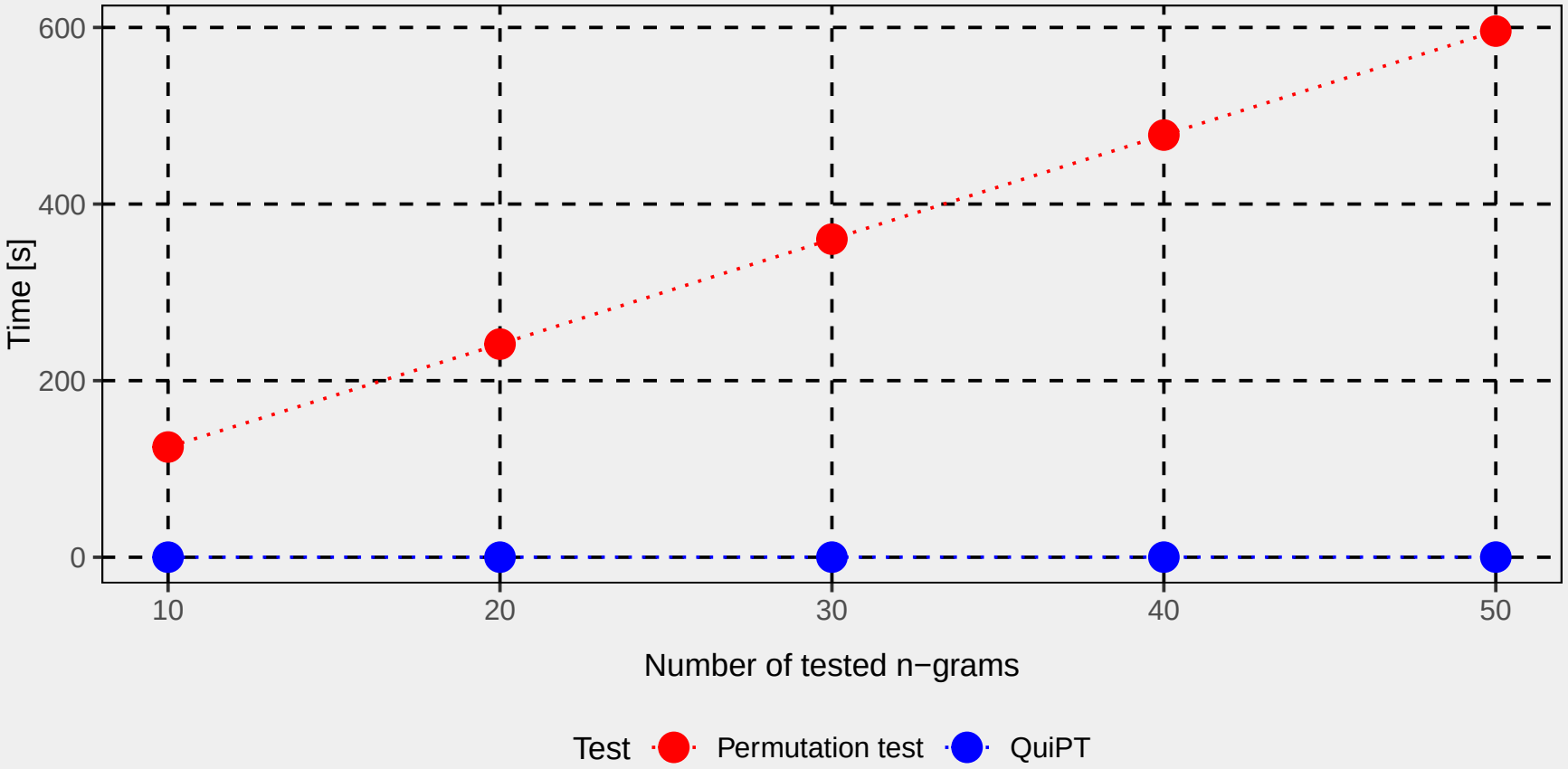
Longer n-games are more informative, but create larger attribute spaces that are more difficult to analyze.

slam: Sparse Lightweight Arrays and Matrices

Counting n-grams creates sparse matrices, that are causing dimensional problems.

	m	storage	value
1	1.00	base	0.000214 Mb
2	1.00	slam	0.001122 Mb
3	10.00	base	0.000969 Mb
4	10.00	slam	0.001312 Mb
5	100.00	base	0.0765 Mb
6	100.00	slam	0.002625 Mb
7	1000.00	base	7.629601 Mb
8	1000.00	slam	0.016357 Mb
9	10000.00	base	762.939659 Mb
10	10000.00	slam	0.153687 Mb

QuiPT



QuiPT (avaible as function in biogram package) is faster than classic permutation tests.

Simplified alphabets

Simplified alphabets:

- amino acids are grouped into larger yields on the basis of specific criteria,
- easier anticipation of structures (Murphy et al., 2000),
- creation of more generalised models.

Simplified alphabets

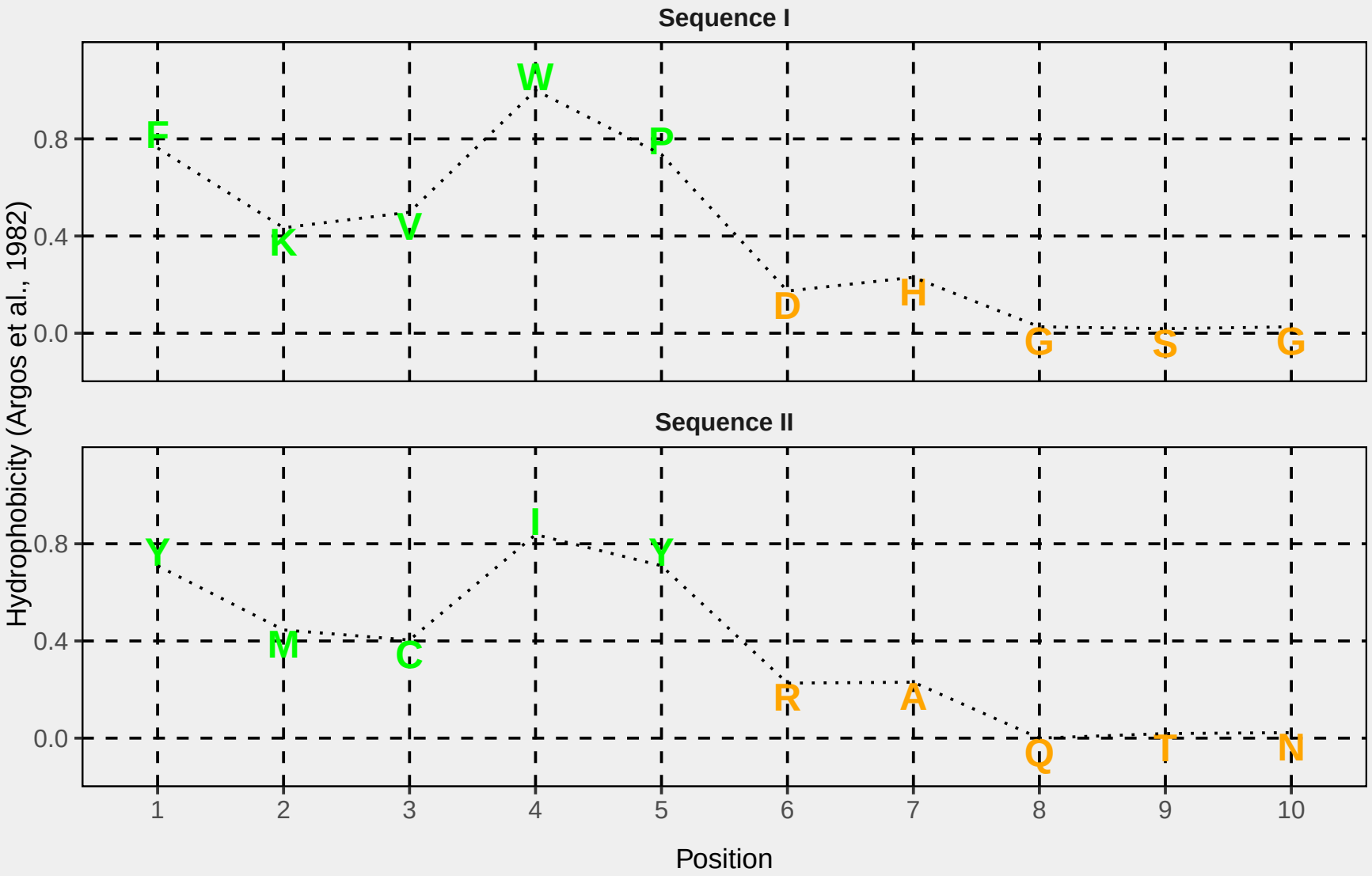
Following peptides appear to be completely different in terms of amino acid composition.

Peptide I:

FKVWPDHGSG

Peptide 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

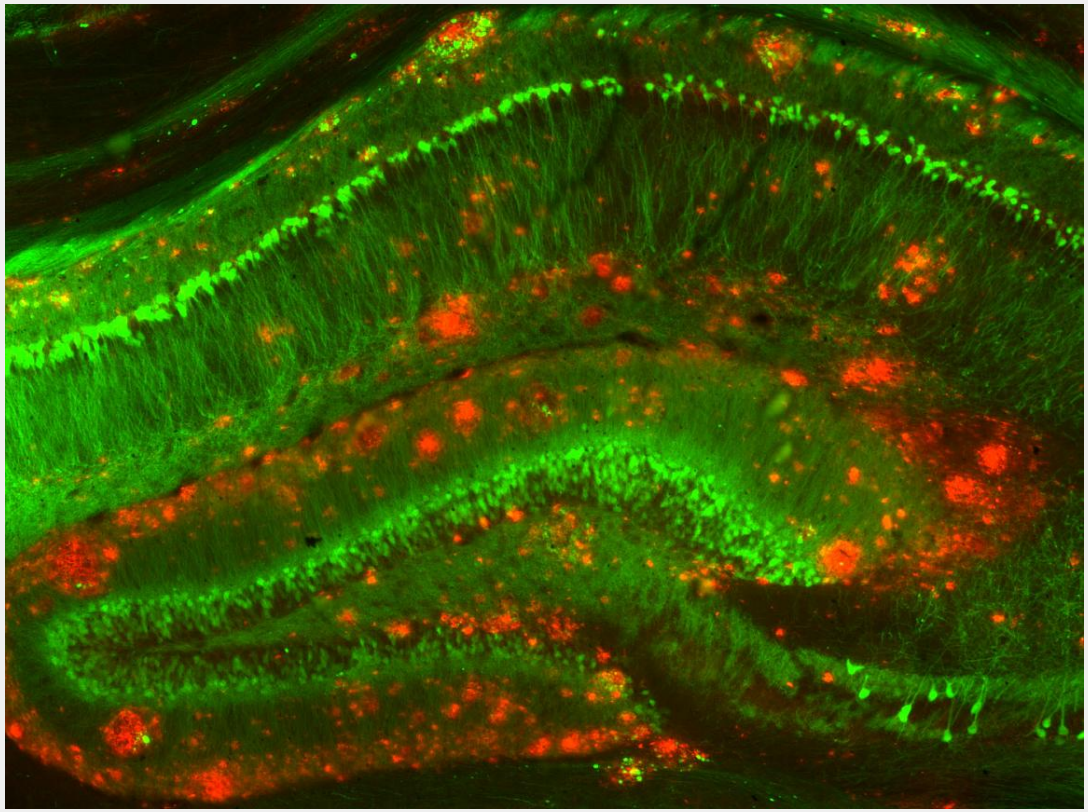
Group	Aminoacids
1	C, I, L, K, M, F, P, W, Y, V
2	A, D, E, G, H, N, Q, R, S, T

Peptide I: FKVWPDHGSG → 1111122222
Peptide II: YMCIYRAQTN → 1111122222

Amyloid prediction

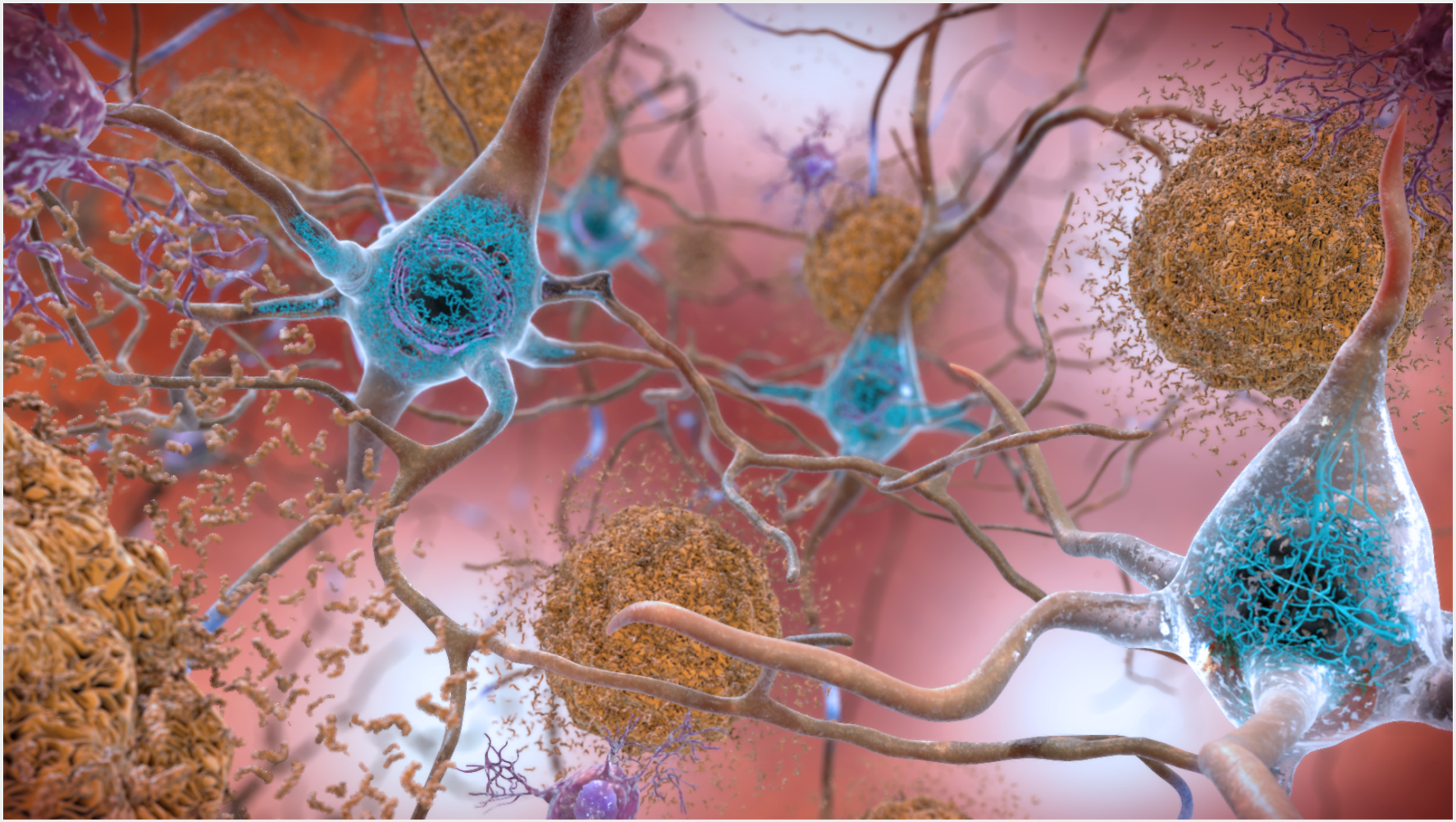
Amyloids

Amyloid aggregates are found in tissues of people suffering from neurodegenerative disorders such as Alzheimer's disease, Parkinson's disease and many other diseases.



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

Amyloids

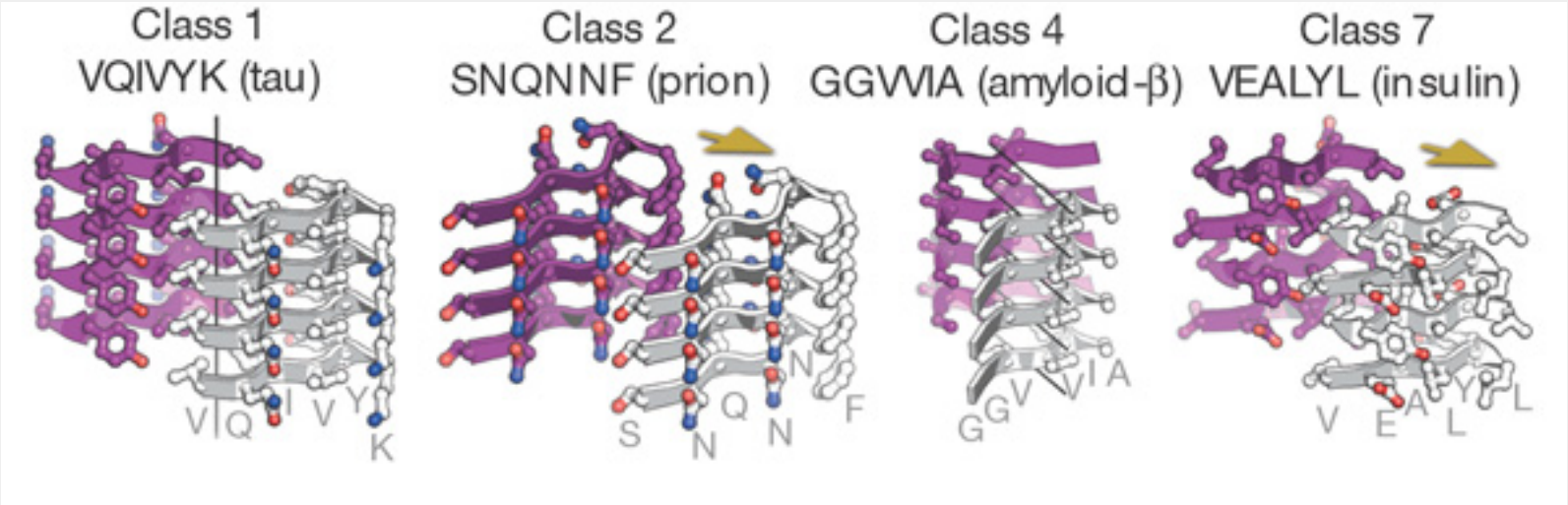


Source: National Institute on Aging (NIA) | National Institutes of Health (NIH)

Amyloid proteins

Peptide sequences with amyloidogenic properties are responsible for the aggregation of amyloidogenic proteins (hot spots):

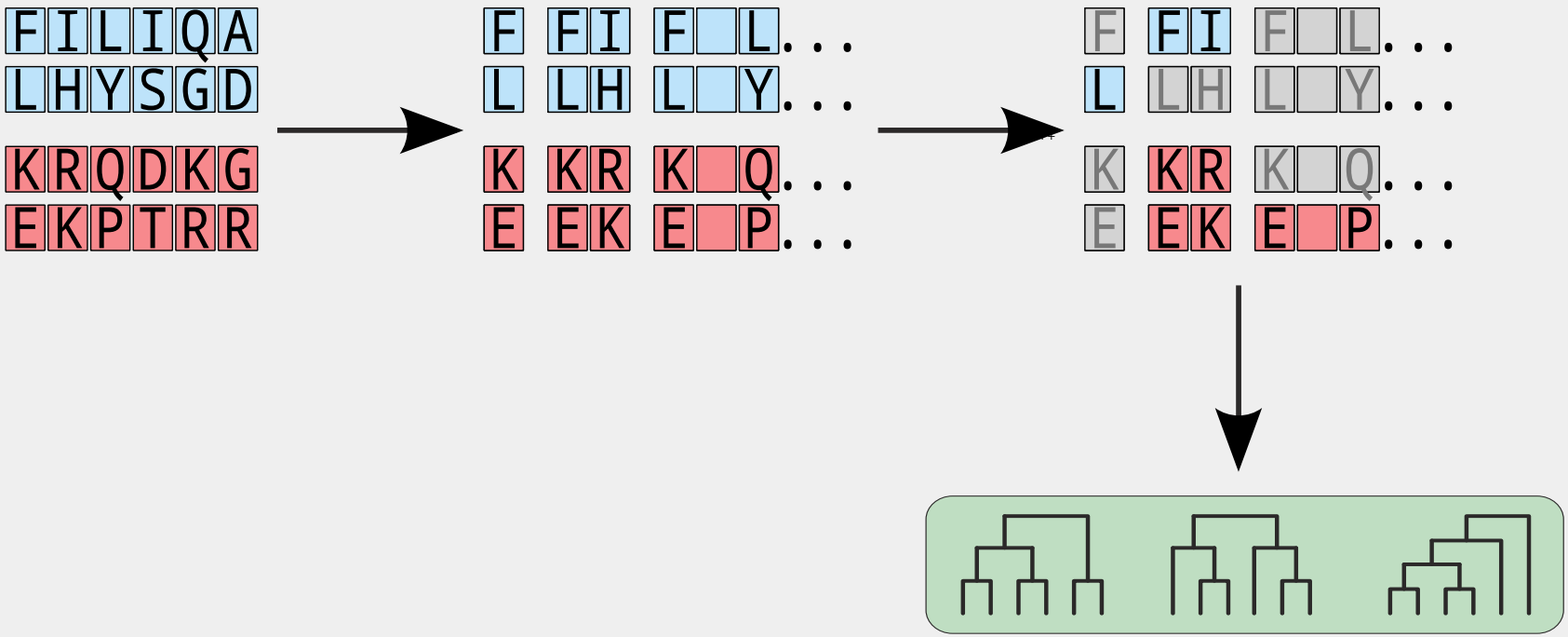
- short (6-15 aminoacids),
- very variable, usually hydrophobic, aminoacid composition,
- create unique β-structures.



Sawaya et al. (2007)

AmyloGram

AmyloGram: n-gram-based amyloid prediction tool (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

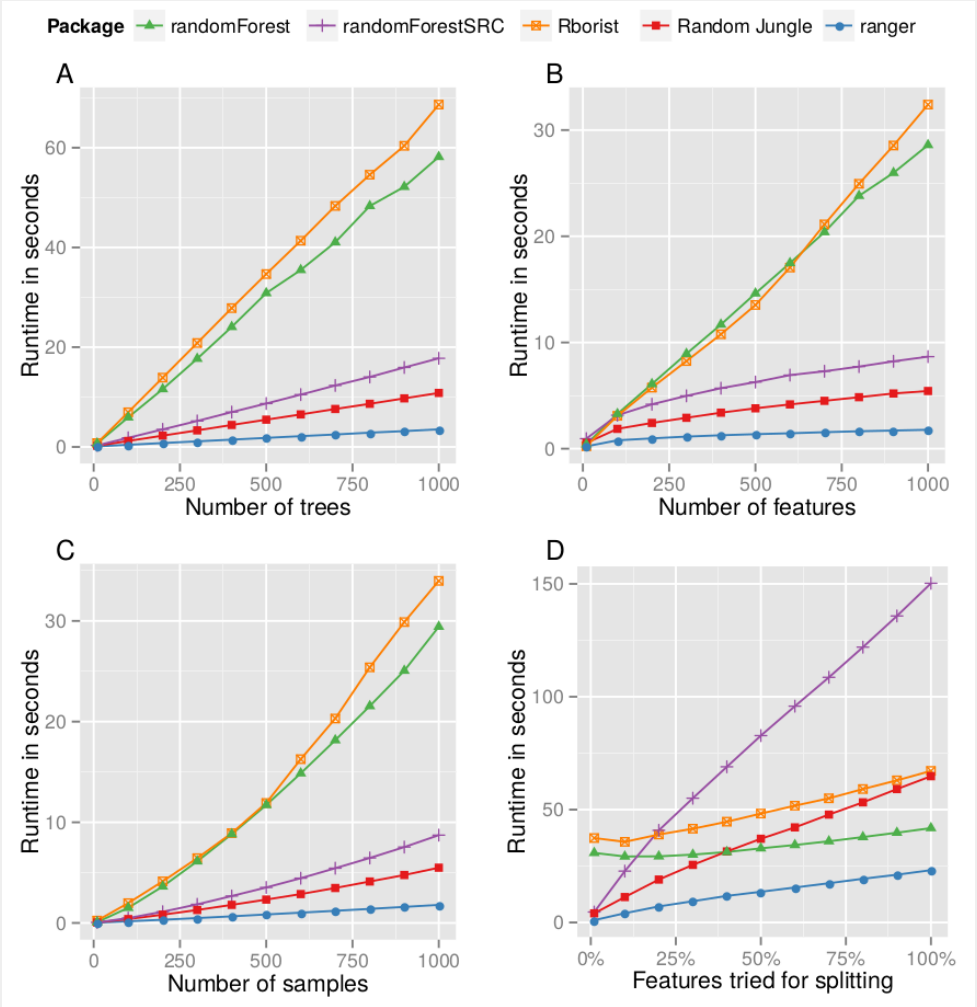
ranger: A Fast Implementation of Random Forests

Package	Runtime [h]			Memory usage [GB]
	mtry=			
	5000	15,000	135,000	
randomForest	101.24	116.15	248.60	39.05
randomForest (MC)	32.10	53.84	110.85	105.77
bigrf	NA	NA	NA	NA
randomForestSRC	1.27	3.16	14.55	46.82
Random Jungle	1.51	3.60	12.83	0.40
Rborist	NA	NA	NA	>128
ranger	0.56	1.05	4.58	11.26
ranger (save.memory)	0.93	2.39	11.15	0.24
ranger (GWAS mode)	0.23	0.51	2.32	0.23

Runtime and memory usage for the analysis of a simulated dataset mimicking a genome-wide association study (GWAS). NA values indicate unsuccessful analyses:
without disk caching failed because of memory shortage for all mtry values and number of CPU cores.
With disk caching, we stopped bigrf after 16 days of computation.

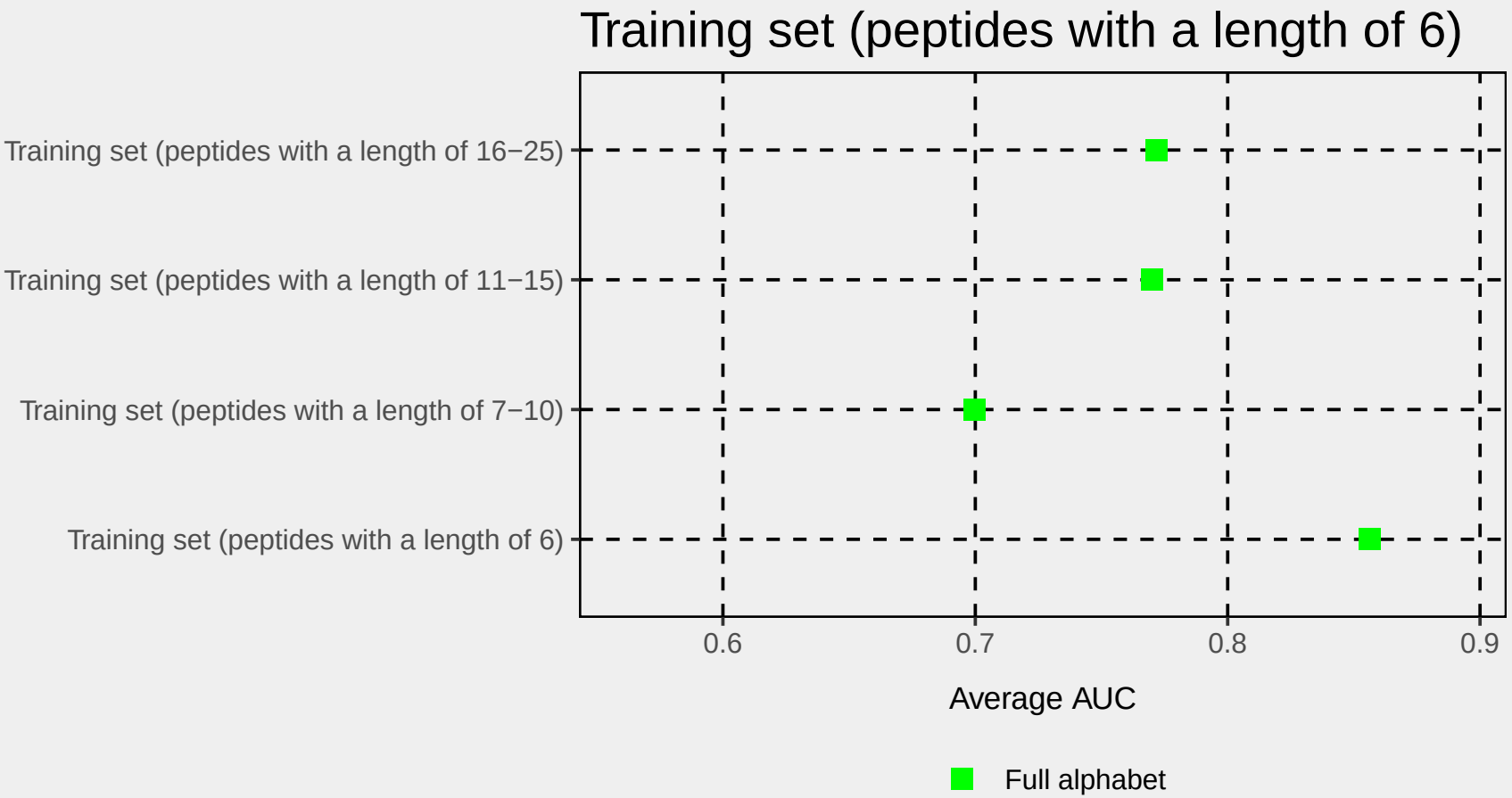
Marvin N. Wright and Andreas Ziegler. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software 1, 77

ranger: A Fast Implementation of Random Forests



Marvin N. Wright and Andreas Ziegler. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software 1, 77

Cross-validation

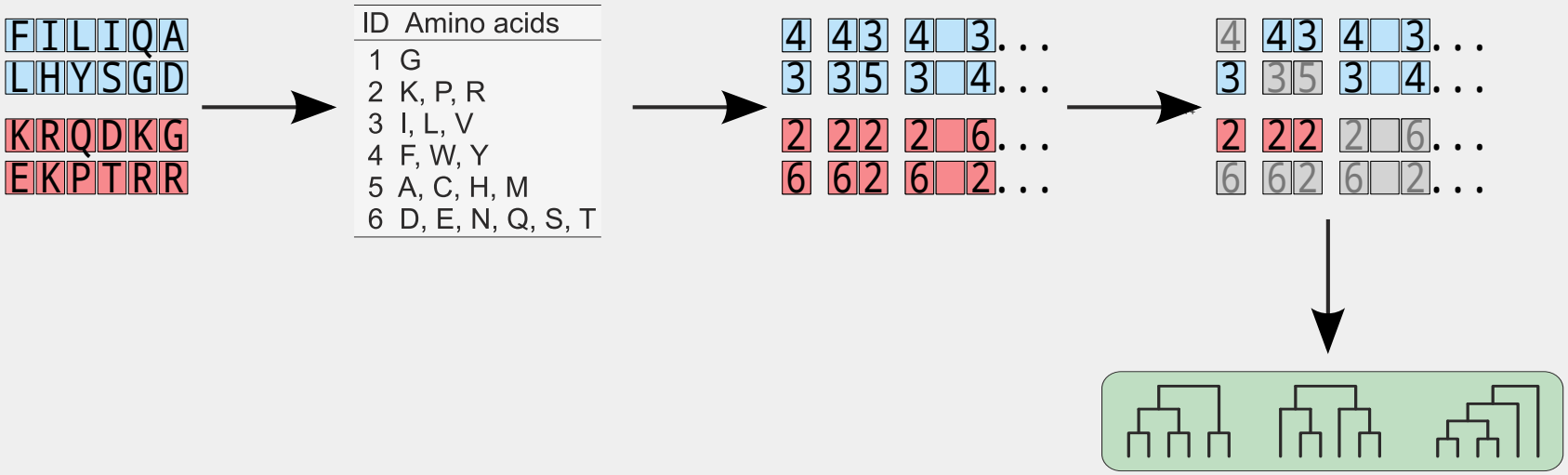


Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

Standard simplified alphabets

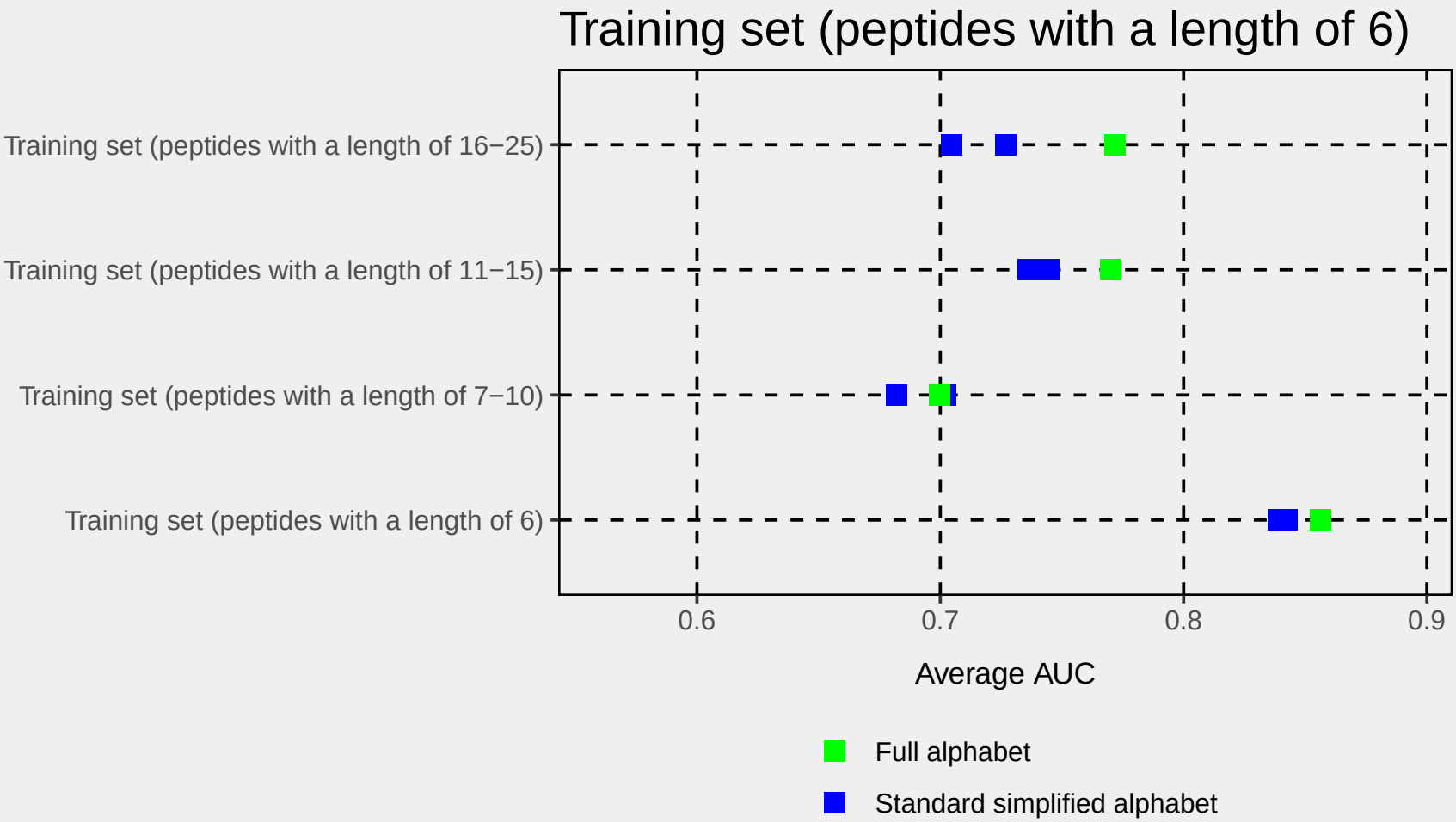
Do standard simplified alphabets developed for different biological issues help to improve amyloid prediction?

Standard simplified alphabets



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

Standard simplified alphabet



Standard aminoacid alphabets do not improve the quality of amyloid prediction.

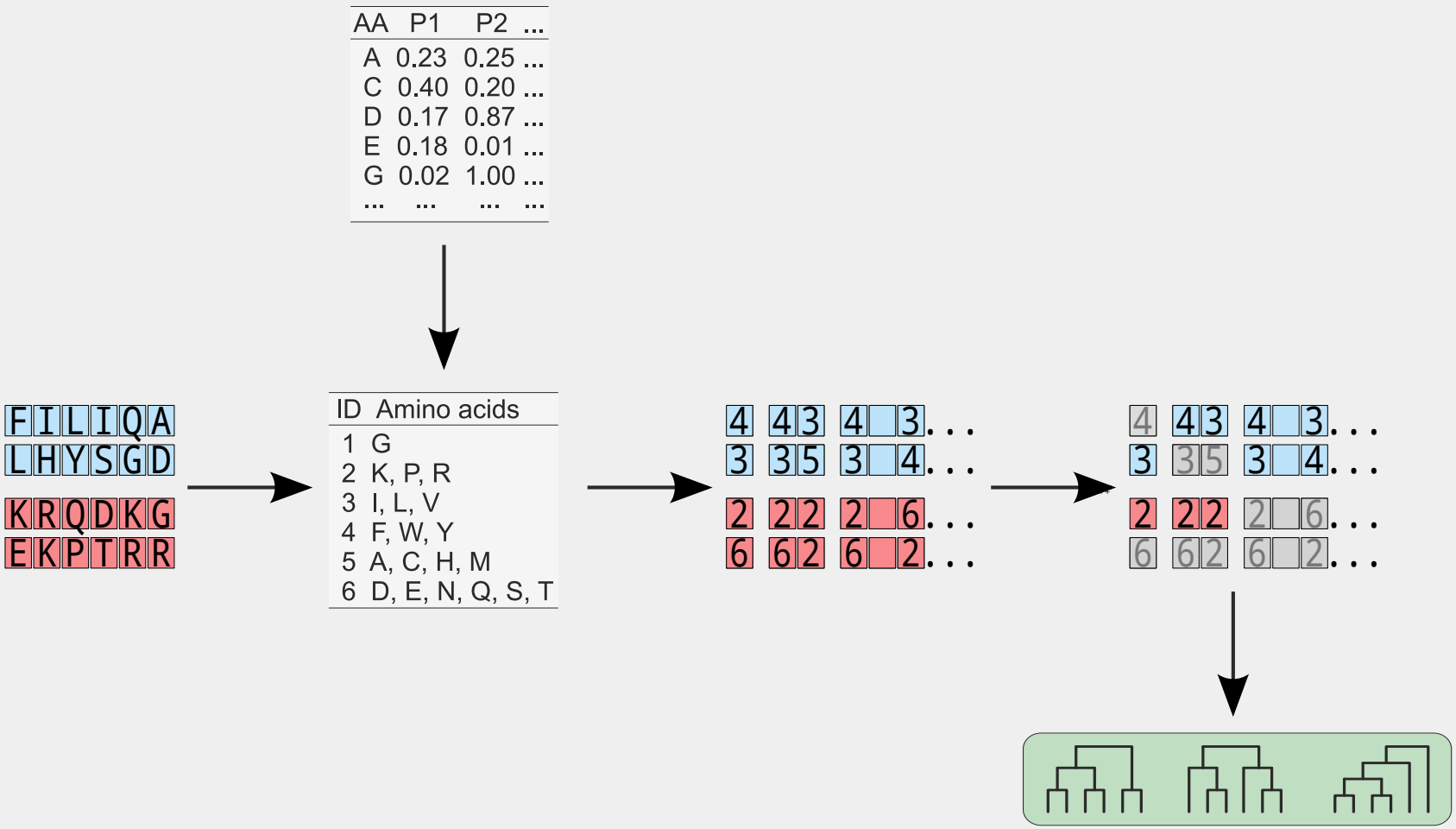
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

New simplified alphabets

- 17 physicochemical parameters selected from AAindex database:
  - size,
  - hydrophobicity,
  - frequency in β-sheets,
  - ability to make contact.
- 524 284 simplified aminoacid alphabets of various sizes (from 3 to 6 groups)

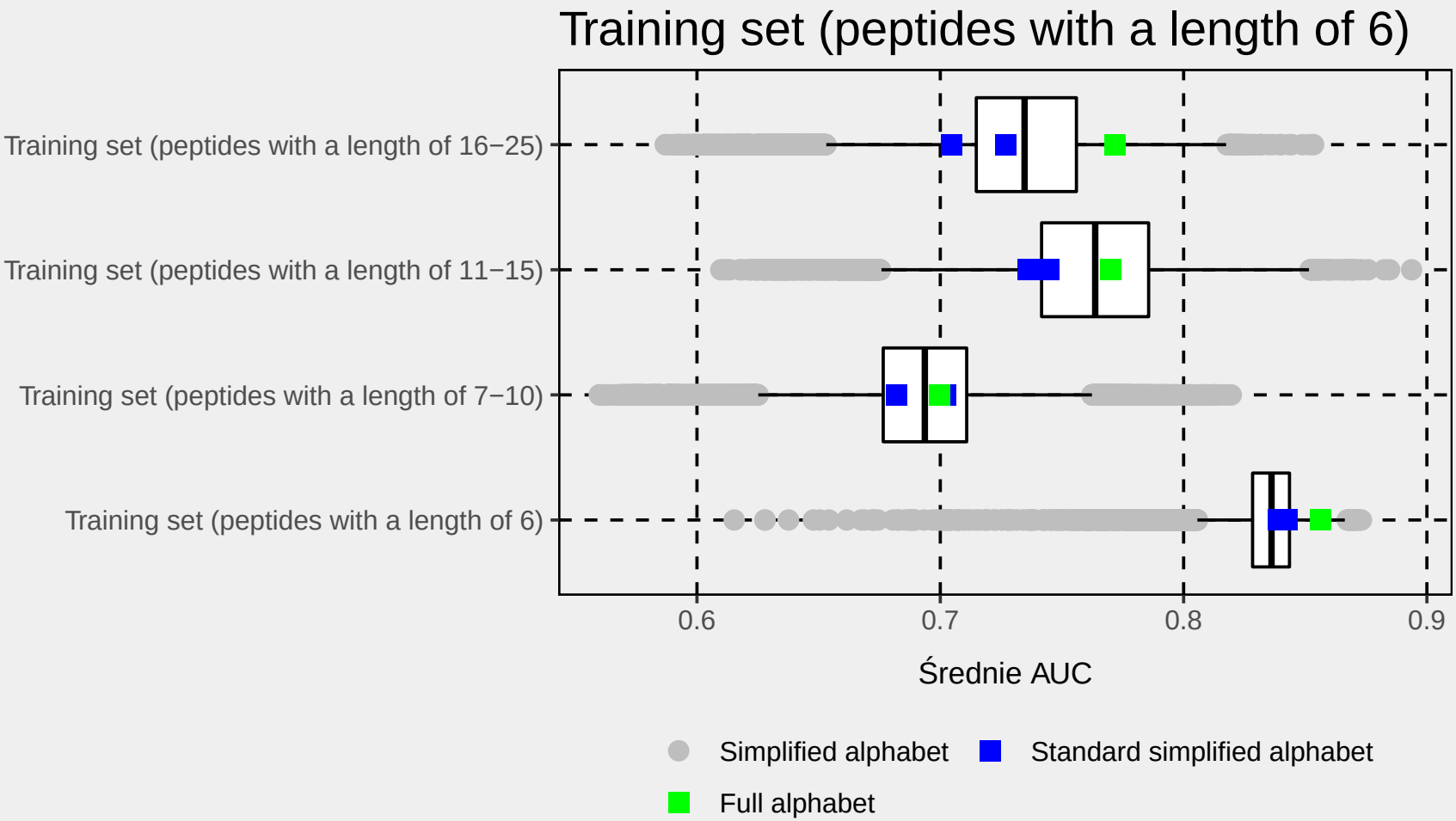
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

New simplified alphabets



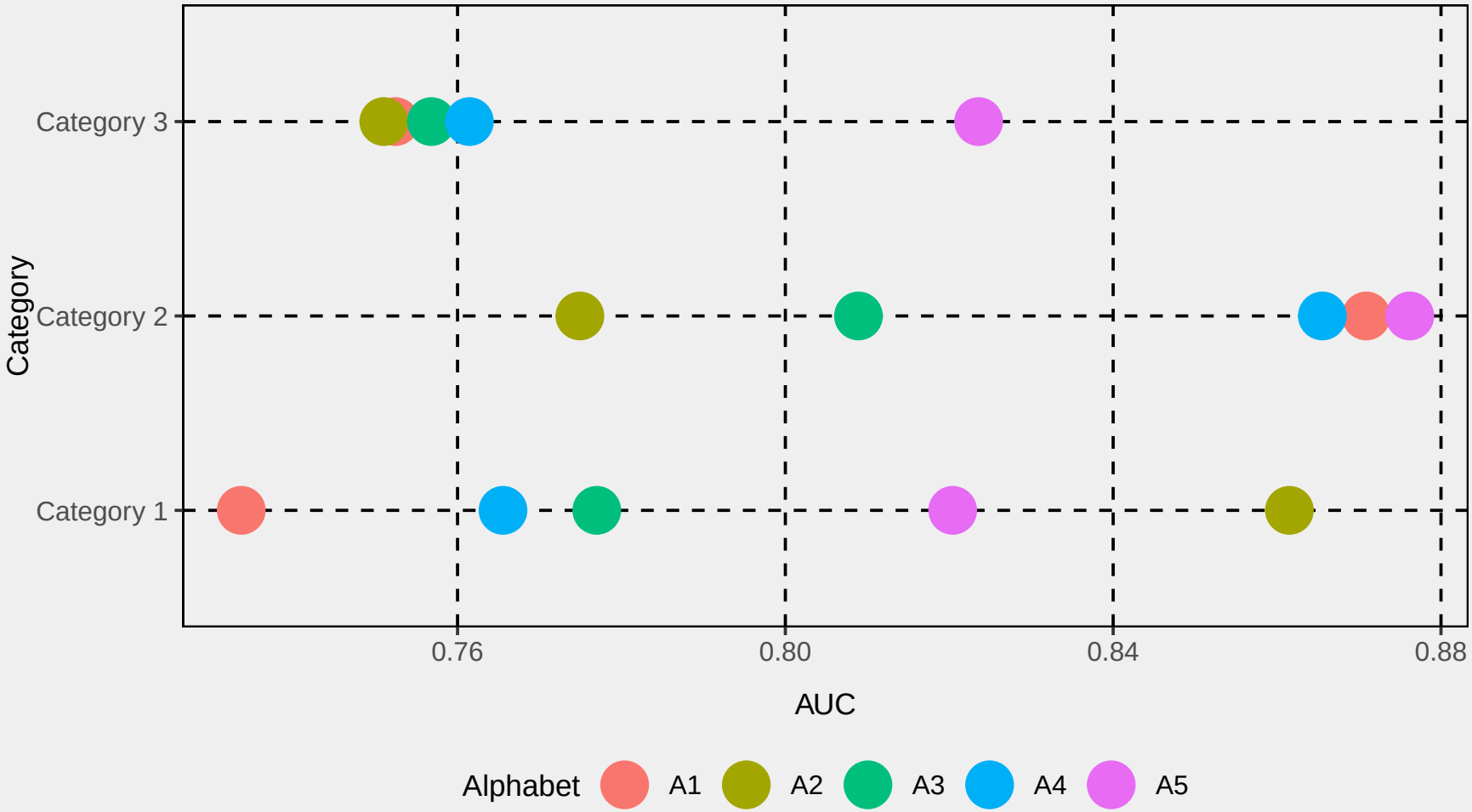
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

Standard simplified alphabet

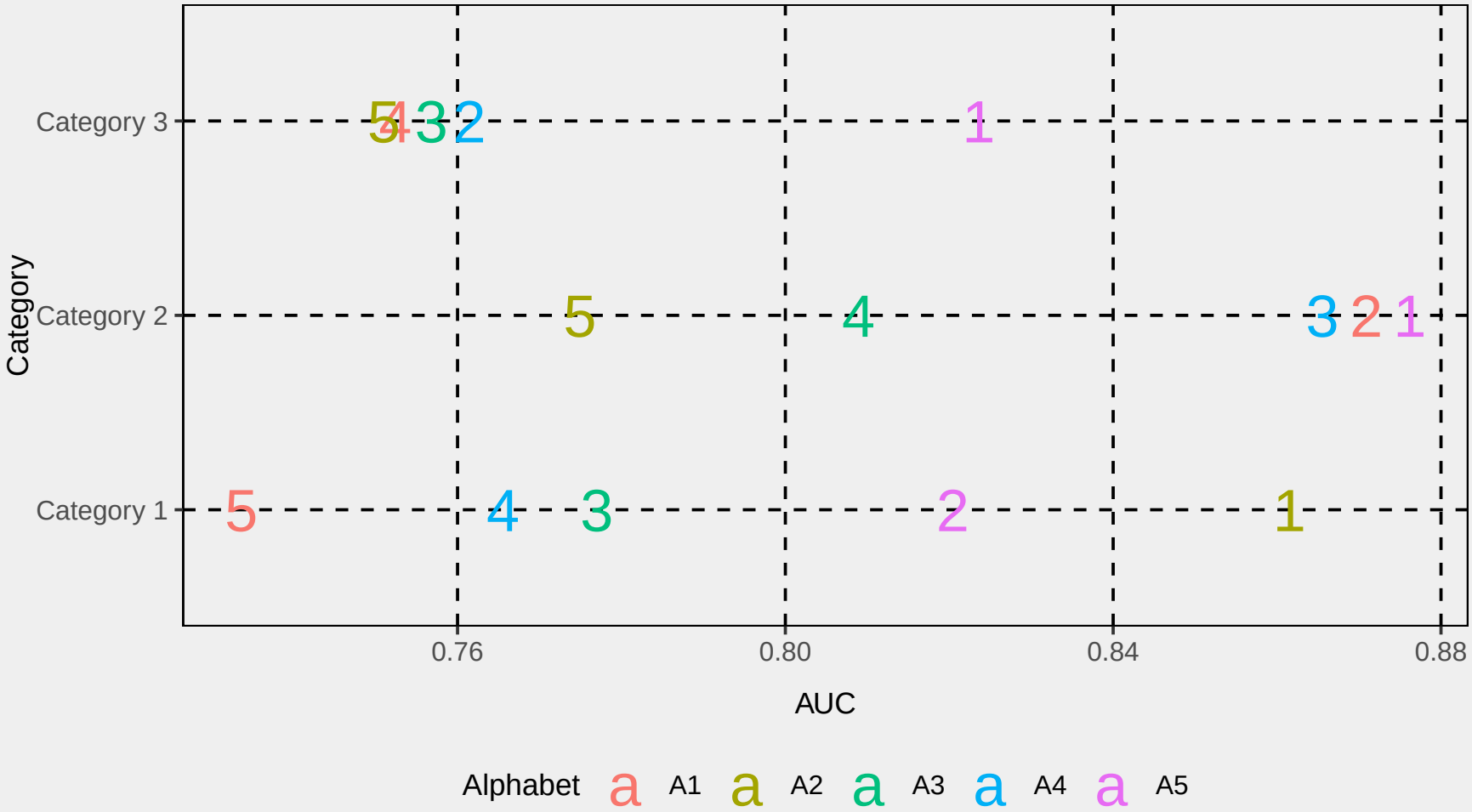


Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961

Selection of best-performing simplified alphabet

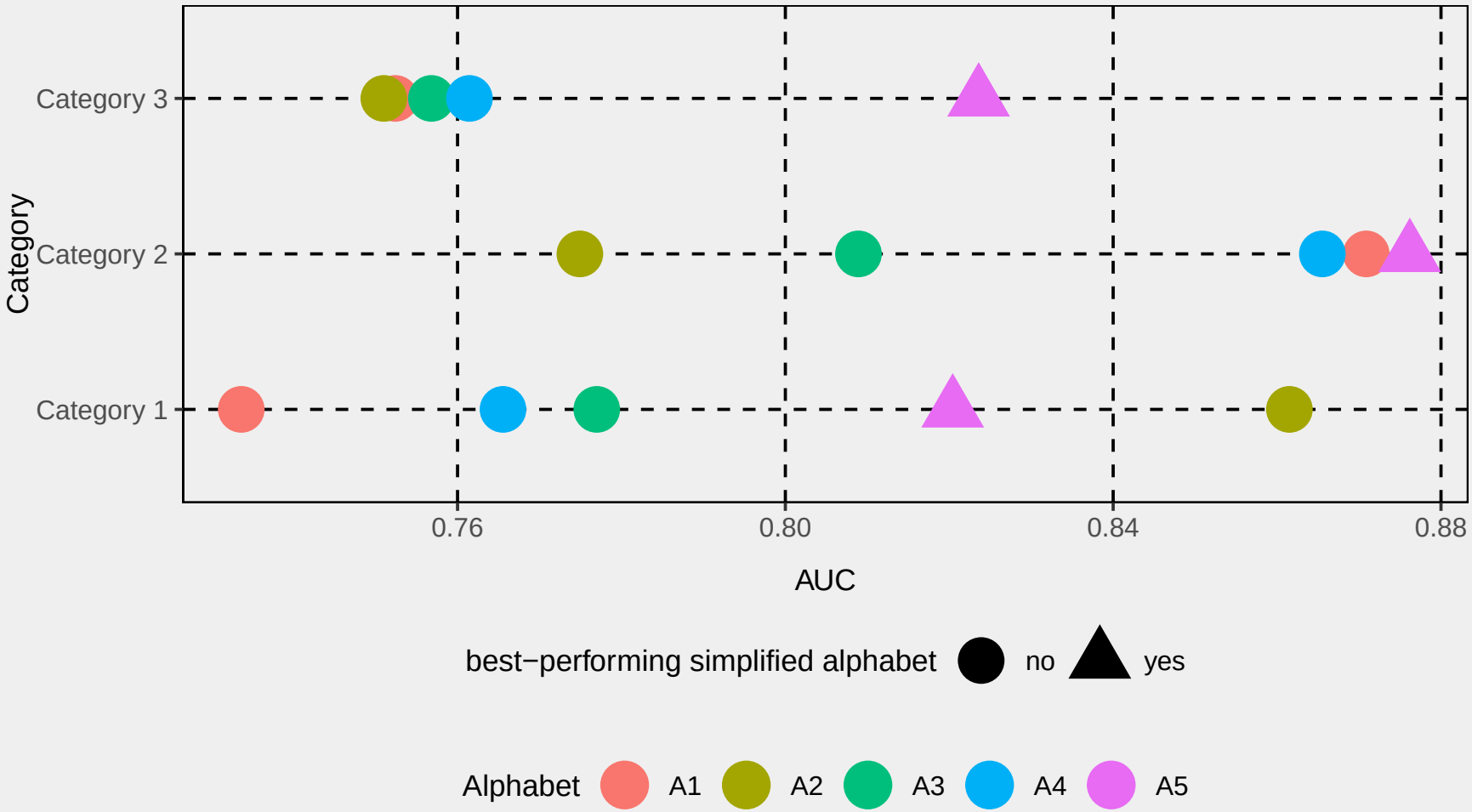


Selection of best-performing simplified alphabet



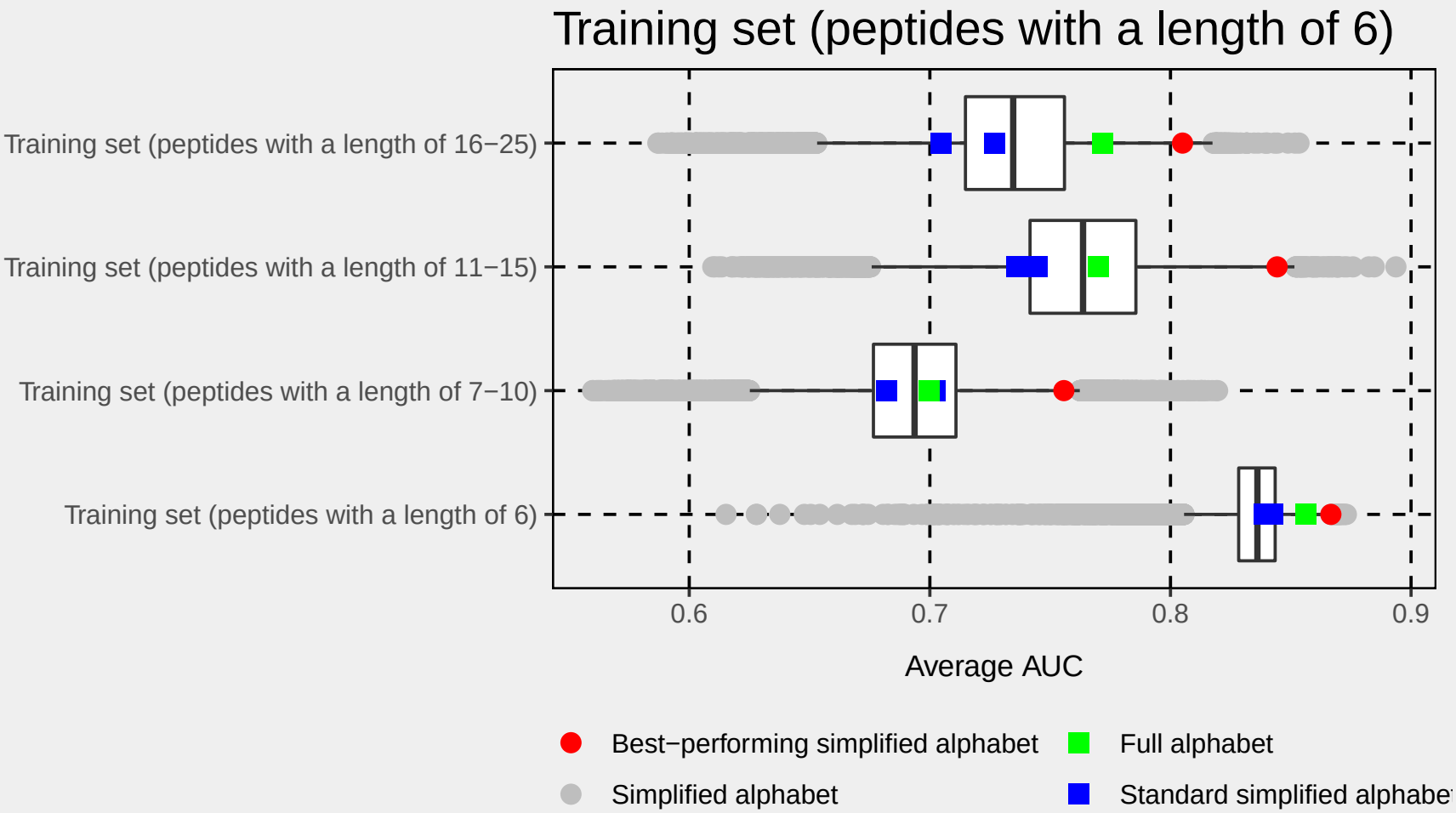
For each category the alphabets have been ranked (rank 1 for the best AUC, etc.)

Selection of best-performing simplified alphabet



The best alphabet was the one with the lowest rank sum.

Best-performing simplified alphabet



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

Best-performing simplified alphabet

Group	Aminoacids
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

Best-performing simplified alphabet

Group	Aminoacids
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 - hydrophobic aminoacids.

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

Best-performing simplified alphabet

Group	Aminoacids
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

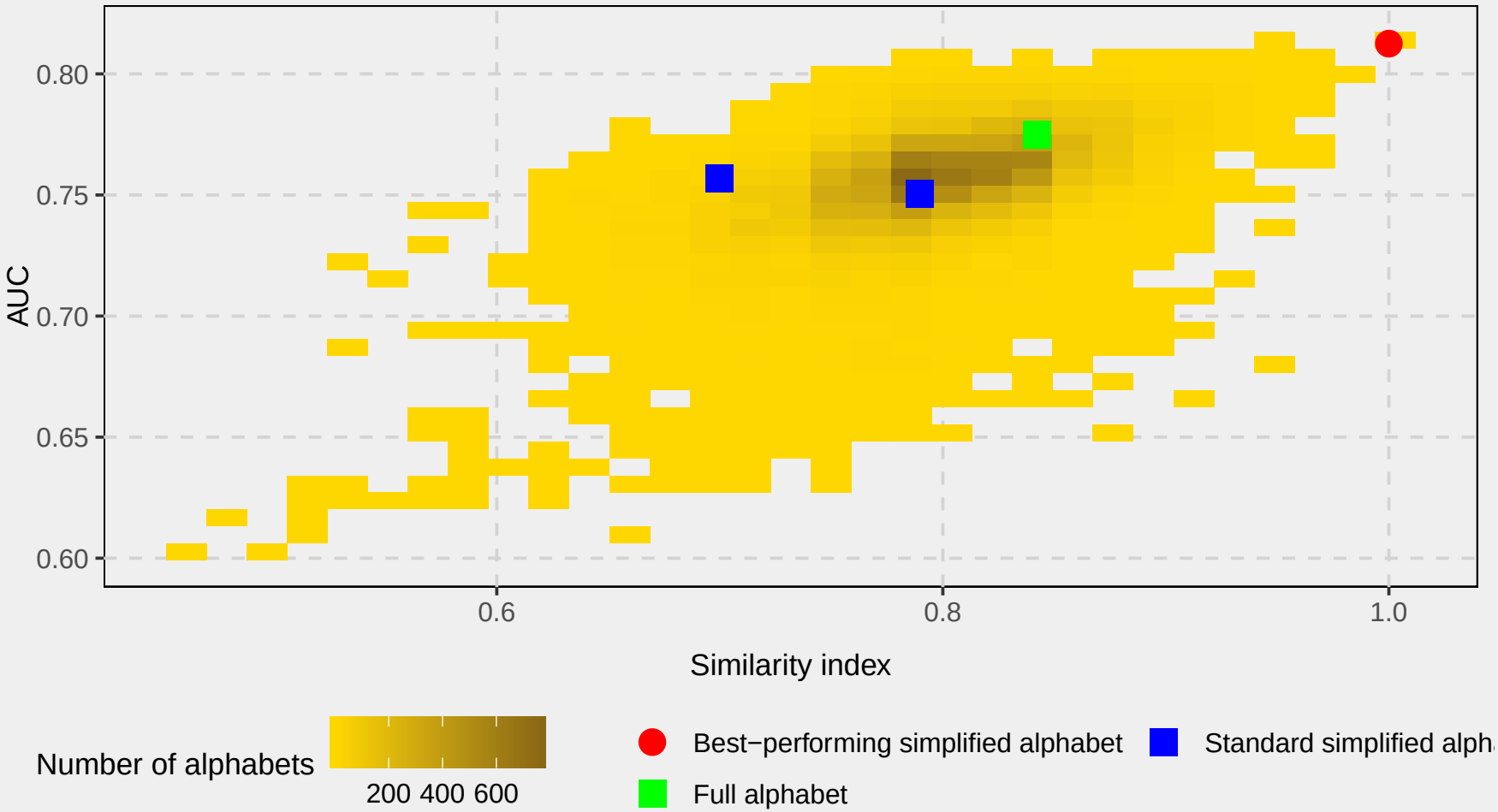
Group 2 - aminoacids disrupting the β-structure.

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

Alphabet similarity and quality of prediction

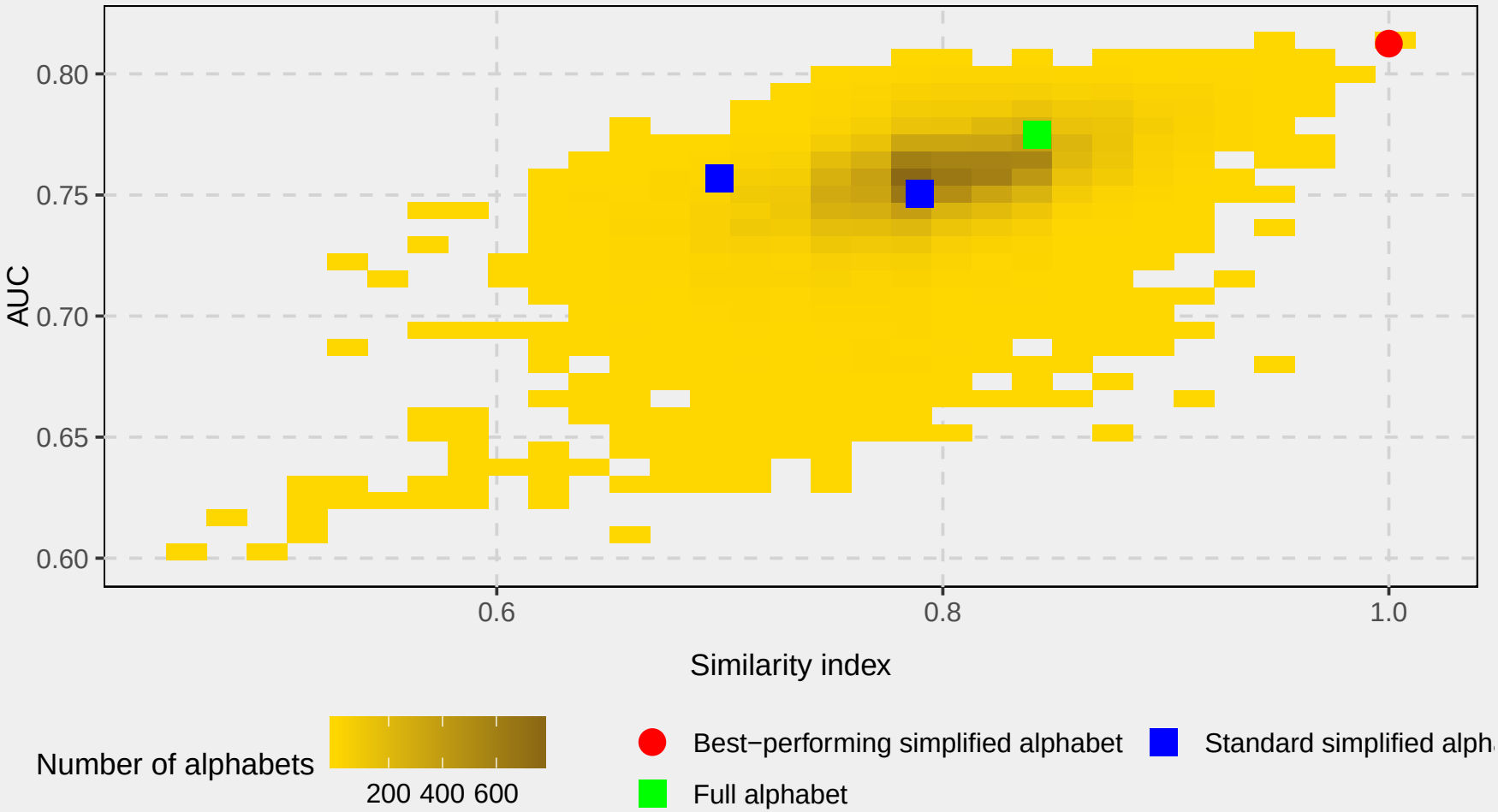
Do alphabets similar to the best simplified alphabet also support amyloid predictions?

Similarity index



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

Similarity index

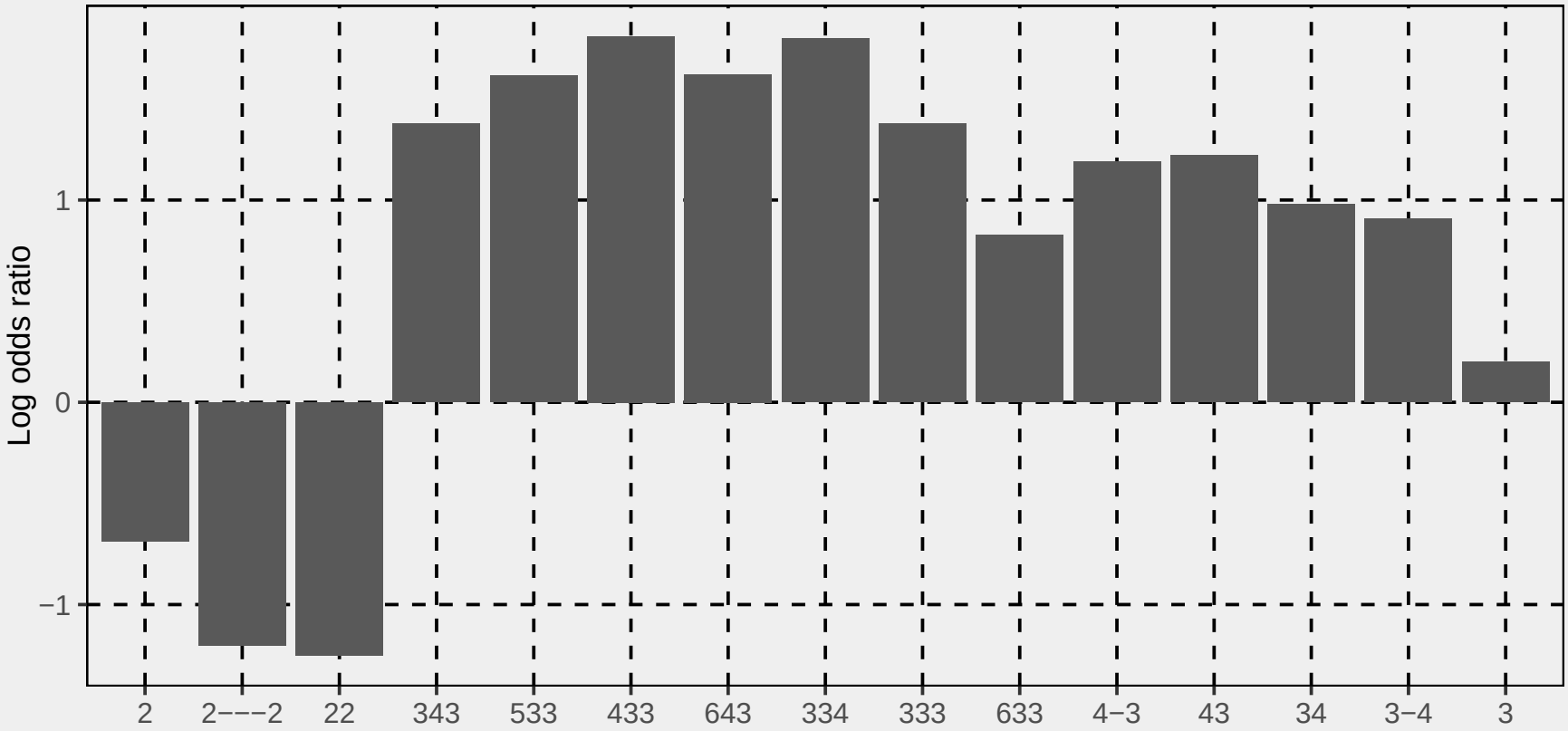


The correlation between the similarity index and the average AUC is important ($\text{p-value} \leq 2.2^{-16}$; $\rho = 0.51$).

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

Are the informative n-grams found by QuiPT are connected with amyloidogenicity?

Informative n-grams



Of the 65 most informative n-grams, 15 (23%) are also present in amino acid motifs found experimentally (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

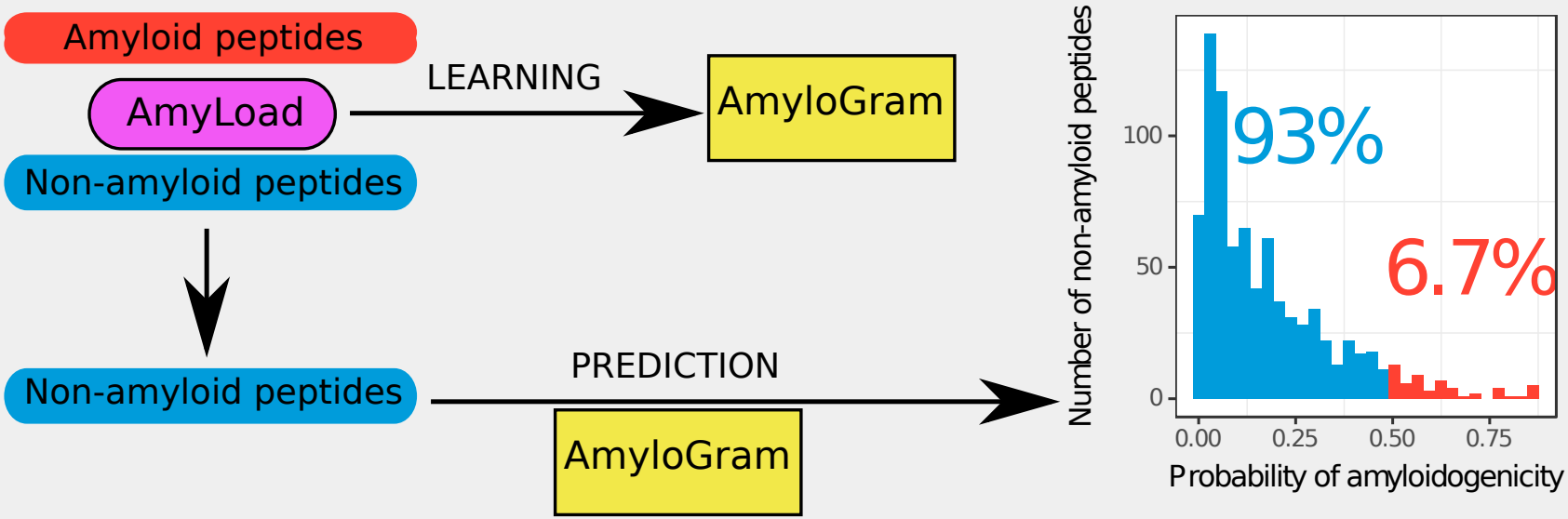
Benchmark with other software

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

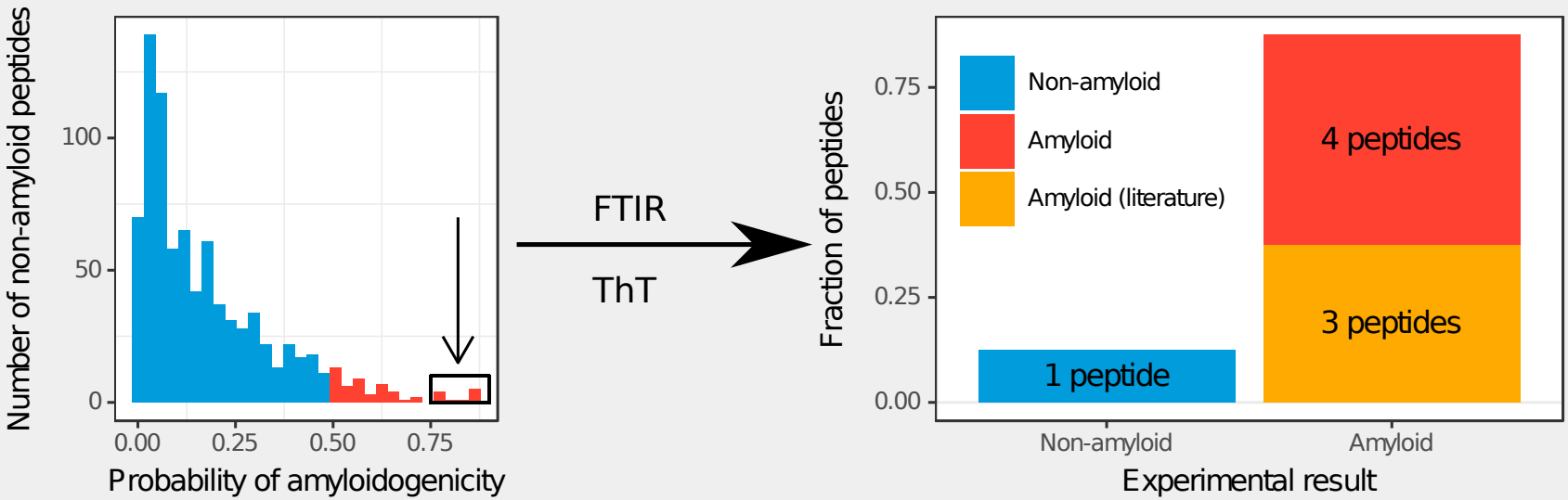
The classifier trained using the best simplified alphabet, AmyloGram, has been compared with other amyloid prediction tools using an external dataset pep424.

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

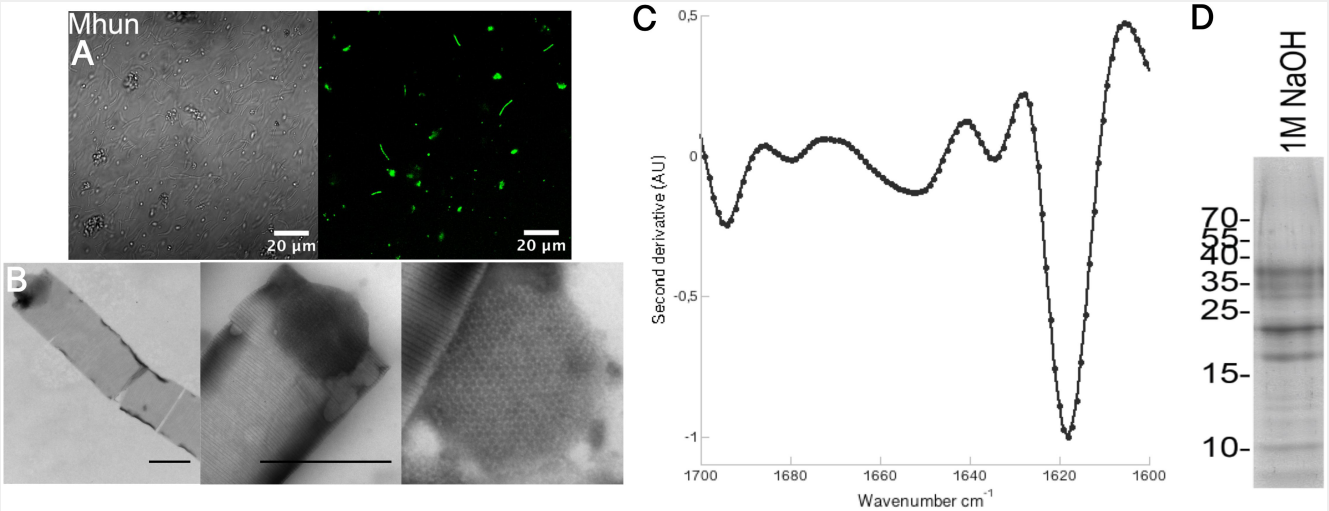
Experimental validation



Experimental validation



New amyloid



A new functional amyloid produced by Methanospirillum sp. (Christensen et al., 2018) was selected for in vitro analysis by AmyloGram.

Shiny aplication

AmyloGram web server

AmyloGram

AmyloGram predicts amyloidogenic sequences (hot spots) in eukaryotic proteins using random forests and n-gram analysis.

Restrictions:

- Be patient - calculations can take up to few minutes.
- Up to 50 sequences may be analyzed at the same time using web server. If you need larger query, please use the AmyloGram package for **R**.

Authors: Michal Burdukiewicz, Piotr Sobczyk.

Citation: 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.

Exemplary sequences

```
>AMY133|Alpha 1|N-terminal domain of Ribosomal prot
GYANNFLFKQG
>Ada-2h
VPSNEEQIKNLLQLEAQEHLQY
>AMY138|Alpha 6|Glutathione S Transeferase P domain
QISFADYNLLDLLRIHQVLN
>AMY143|M8|Spectrin SH3
DILTLLNSTNKDWWKVEVND
>CsgA|region 1
SELNIYQYGGGNSALALQTDARN
```

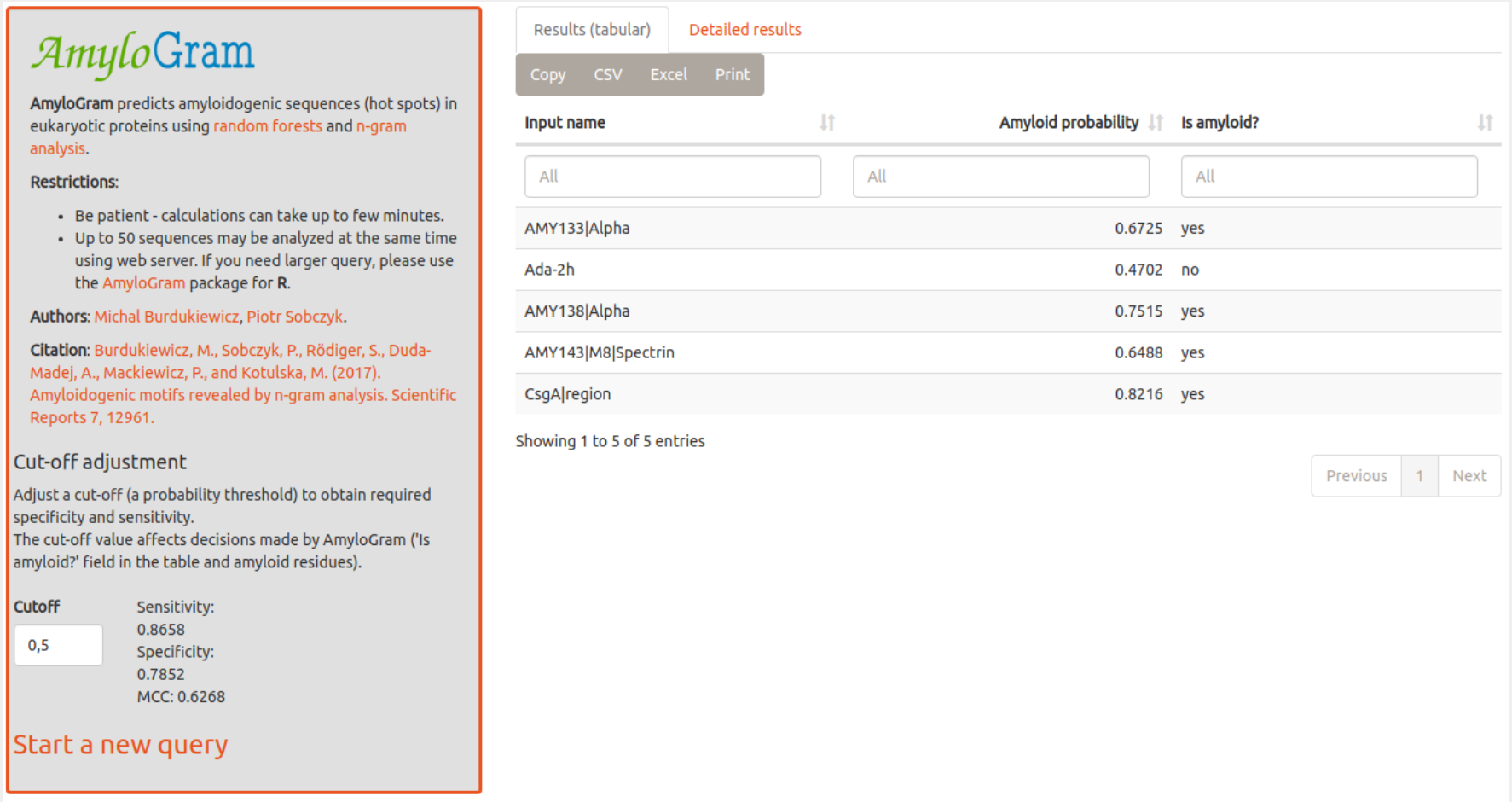
Paste sequences (FASTA format required) here...

Submit data from field above

Submit .fasta or .txt file:

Browse... No file selected

AmyloGram web server



AmyloGram web server

AmyloGram

AmyloGram predicts amyloidogenic sequences (hot spots) in eukaryotic proteins using random forests and n-gram analysis.

Restrictions:

- Be patient - calculations can take up to few minutes.
- Up to 50 sequences may be analyzed at the same time using web server. If you need larger query, please use the AmyloGram package for **R**.

Authors: Michal Burdukiewicz, Piotr Sobczyk.

Citation: 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.

Cut-off adjustment

Adjust a cut-off (a probability threshold) to obtain required specificity and sensitivity.
The cut-off value affects decisions made by AmyloGram ('Is amyloid?' field in the table and amyloid residues).

Cutoff: 0,5

Sensitivity: 0.8658
Specificity: 0.7852
MCC: 0.6268

Start a new query

Results (tabular) | Detailed results

Amyloid residues

Residues are defined as belonging to the amyloid part of a protein, if their amyloid probability is higher than the cut-off

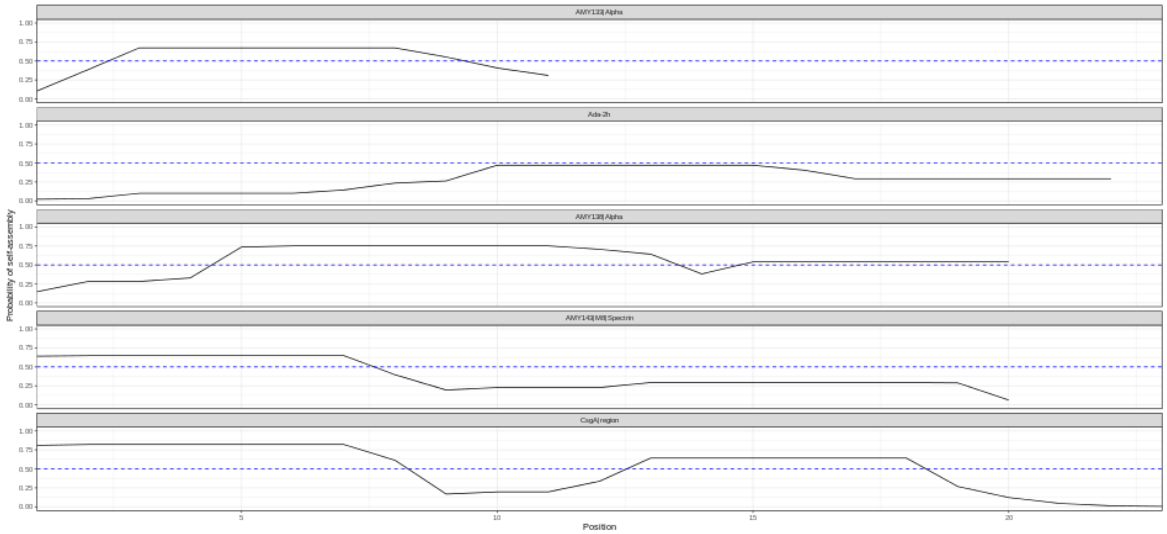
Copy CSV Excel Print

Protein	Fraction of amyloid residues
All	All
AMY133|Alpha	0.6364
Ada-2h	0.0000
AMY138|Alpha	0.7500
AMY143|M8|Spectrin	0.3500
CsgA|region	0.6087

Showing 1 to 5 of 5 entries

Previous 1 Next

Amyloid regions



Summarise()

Web servers:

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

R packages:

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

Summarise()

Models predicting the properties of proteins may be based on precise rules that are understandable to biologists and experimentally verifiable without losing their effectiveness.

Acknowledgements

- Michał Burdukiewicz (Politechnika Warszawska).
- Małgorzata Kotulska (Politechnika Wrocławska).
- Stefan Rödiger (Brandenburg University of Technology Cottbus-Senftenberg).
- Paweł Mackiewicz (Uniwersytet Wrocławski).
- Piotr Sobczyk (Politechnika Wrocławska).

Acknowledgements

Funding:

- Polish National Science Centre (2015/17/N/NZ2/01845 i 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.
- German Federal Ministry of Education and Research (InnoProfile-Transfer-Projekt 03IPT611X).

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.

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.

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 III

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.