

# AmyloGram: Analysis of proteins in R

Jarek Chilimoniuk

Department of Bioinformatics and Genomics, University of  
Wrocław

# PRESENTATION PLAN

- 1 Amino acids and proteins
- 2 n-grams and reduced alphabets
- 3 Amyloid prediction
- 4 Shiny application

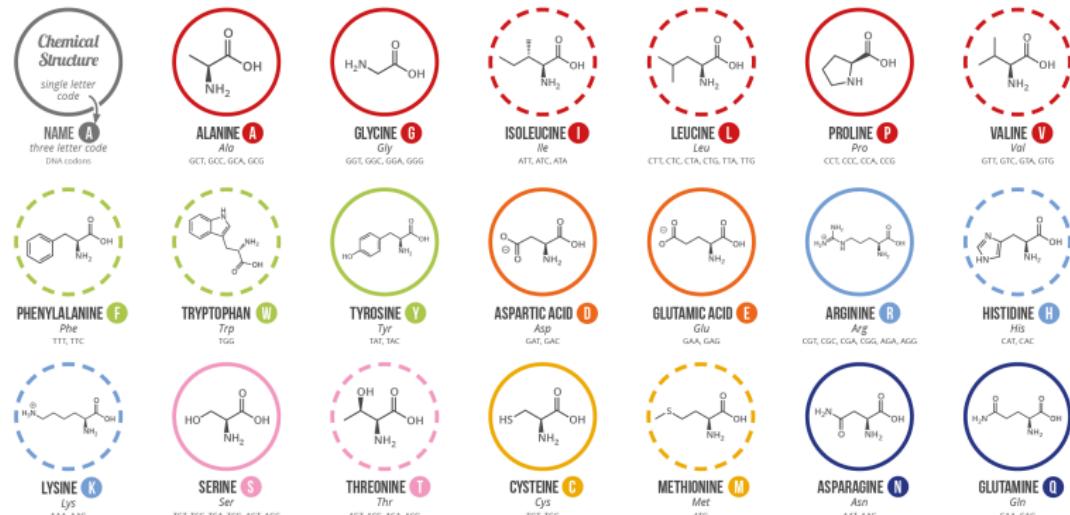
# Amino acids and proteins

# AMINO ACIDS

## A GUIDE TO THE TWENTY COMMON AMINO ACIDS

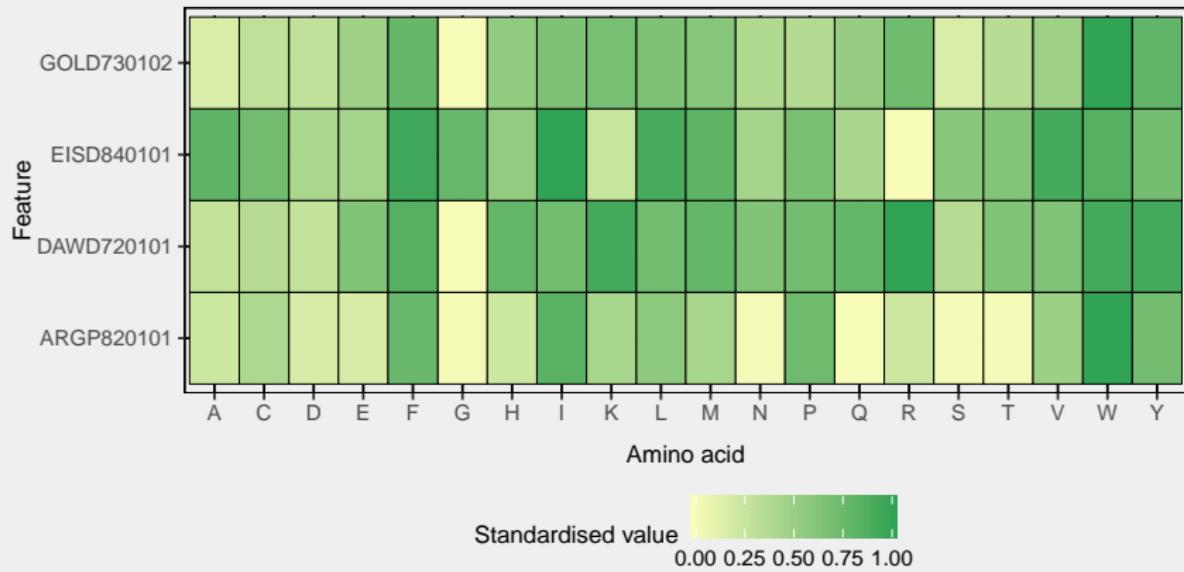
AMINO ACIDS ARE THE BUILDING BLOCKS OF PROTEINS IN LIVING ORGANISMS. THERE ARE OVER 500 AMINO ACIDS FOUND IN NATURE - HOWEVER, THE HUMAN GENETIC CODE ONLY DIRECTLY ENCODES 20. 'ESSENTIAL' AMINO ACIDS MUST BE OBTAINED FROM THE DIET, WHILST NON-ESSENTIAL AMINO ACIDS CAN BE SYNTHESISED IN THE BODY.

**Chart Key:** ● ALIPHATIC ● AROMATIC ● ACIDIC ● BASIC ● HYDROXYLIC ● SULFUR-CONTAINING ● AMIDIC ● ○ NON-ESSENTIAL ● ○ ESSENTIAL

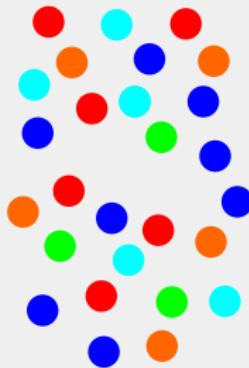


**Note:** This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult.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

# AMINO ACIDS



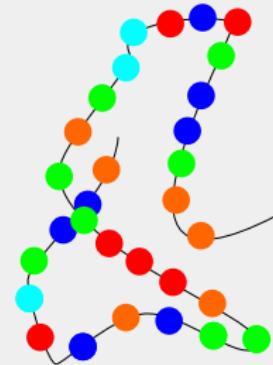
# PROTEINS



Aminoacids

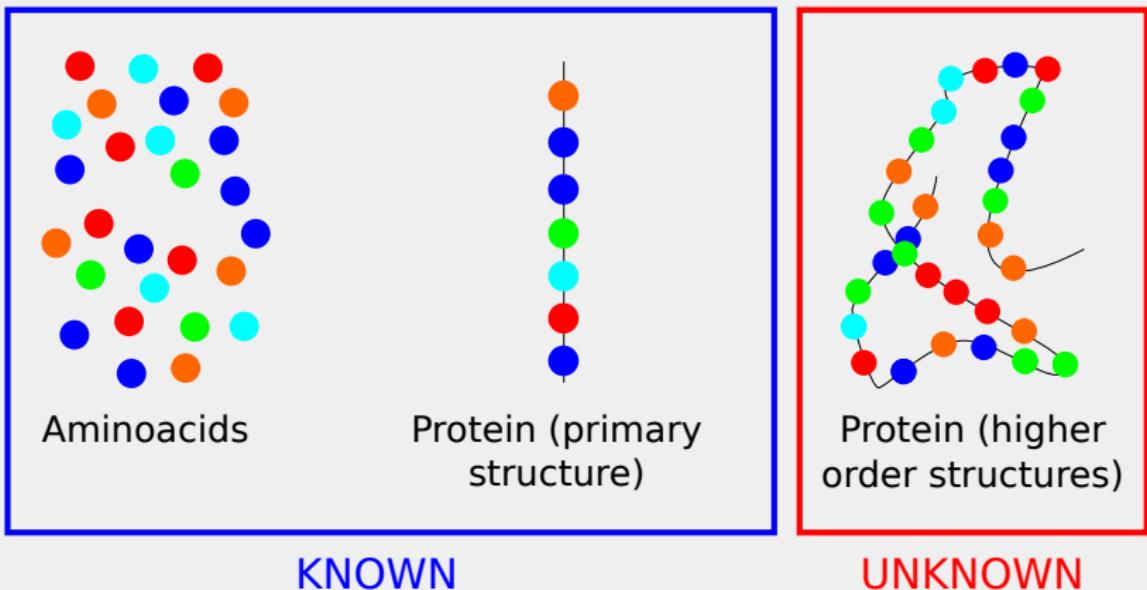


Protein (primary structure)

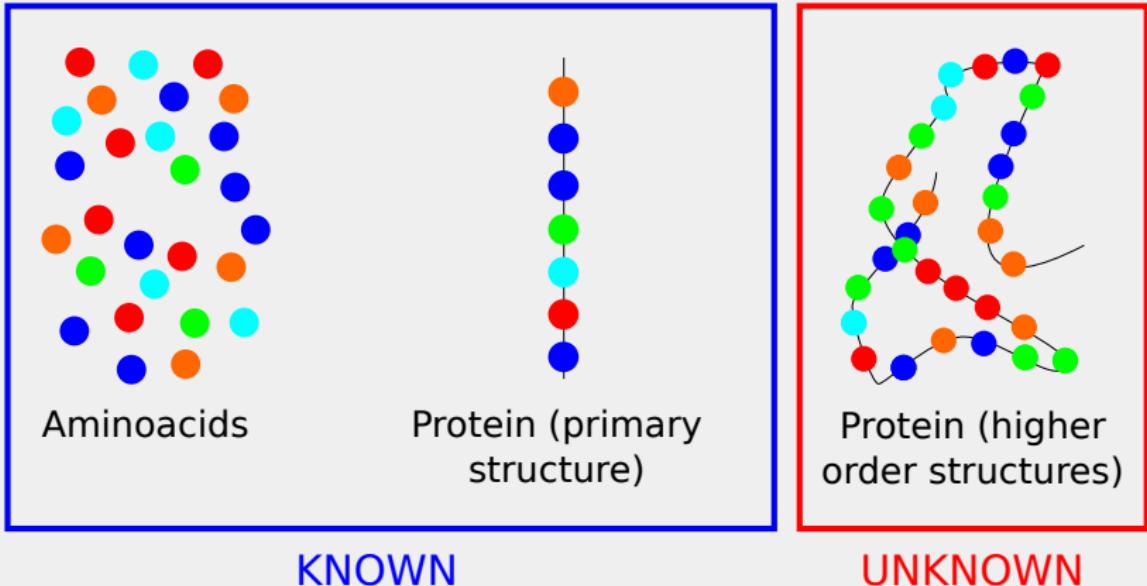


Protein (higher order structures)

# PROTEINS

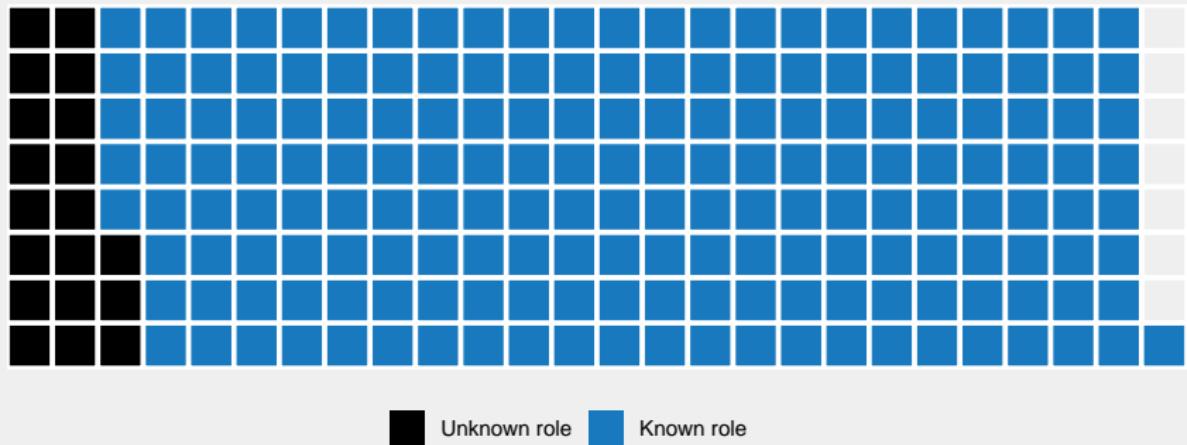


# PROTEINS



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 reduced alphabets

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

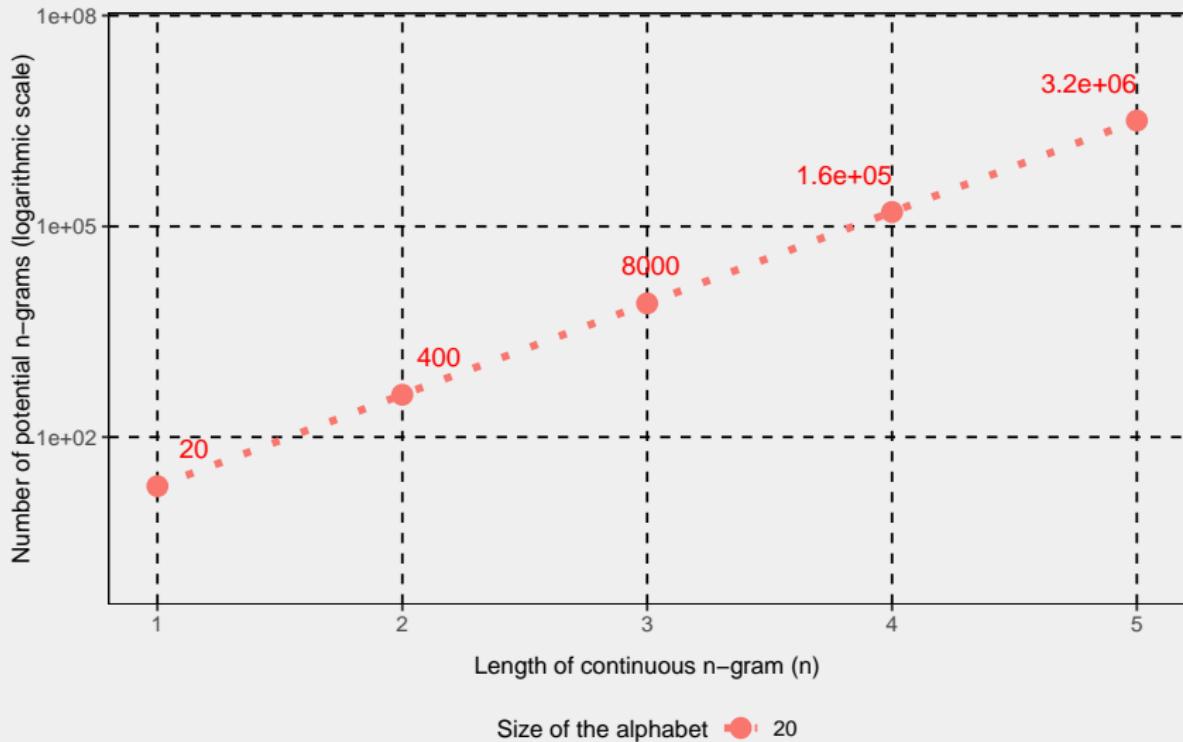
- subsequences (continuous or discontinuous)  $n$  amino acid or nucleotide residues,
- more informative than the individual residues.

Peptide I: **FKVWPDHGSG**

Peptide II: **YMCIYRAQTN**

n-gram examples from peptide I and II:

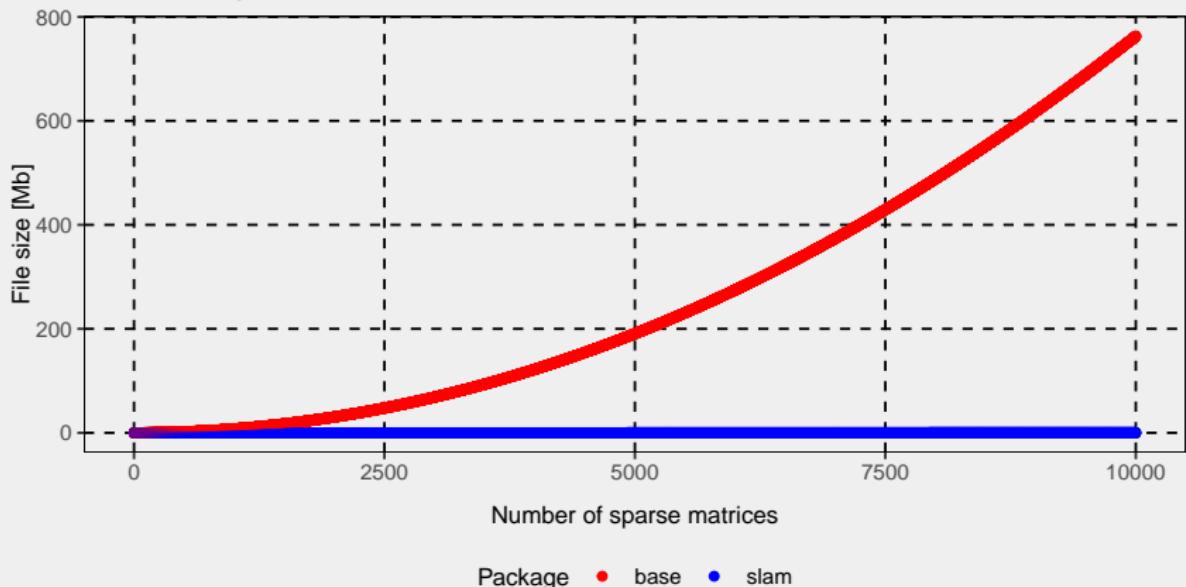
1. 1-gram: **F, Y, K, M,**
2. 2-gram: **FK, YM, KV, MC,**
3. 2-gram (discontinuous): **F-V, Y-C, K-W, M-I,**
4. 3-gram (discontinuous): **F-WP, Y-IY, K-PD, M-YR.**



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

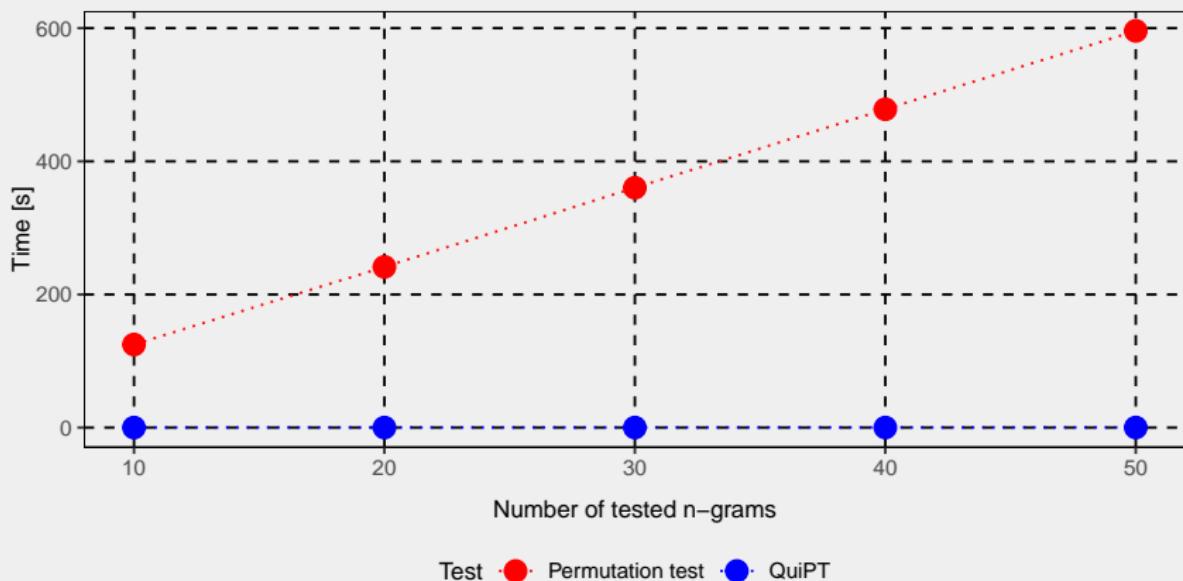
# SLAM: SPARSE LIGHTWEIGHT ARRAYS AND MATRICES

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



# SLAM: SPARSE LIGHTWEIGHT ARRAYS AND MATRICES

Number of sparse matrices		Package	File size [Mb]
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



Quick Permutation Test is a fast alternative to permutation tests for n-gram data. It also allows precise estimation of p-value. QuiPT is available as part of the biogram R package.

# REDUCED ALPHABETS

Reduced 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.

# REDUCED ALPHABETS

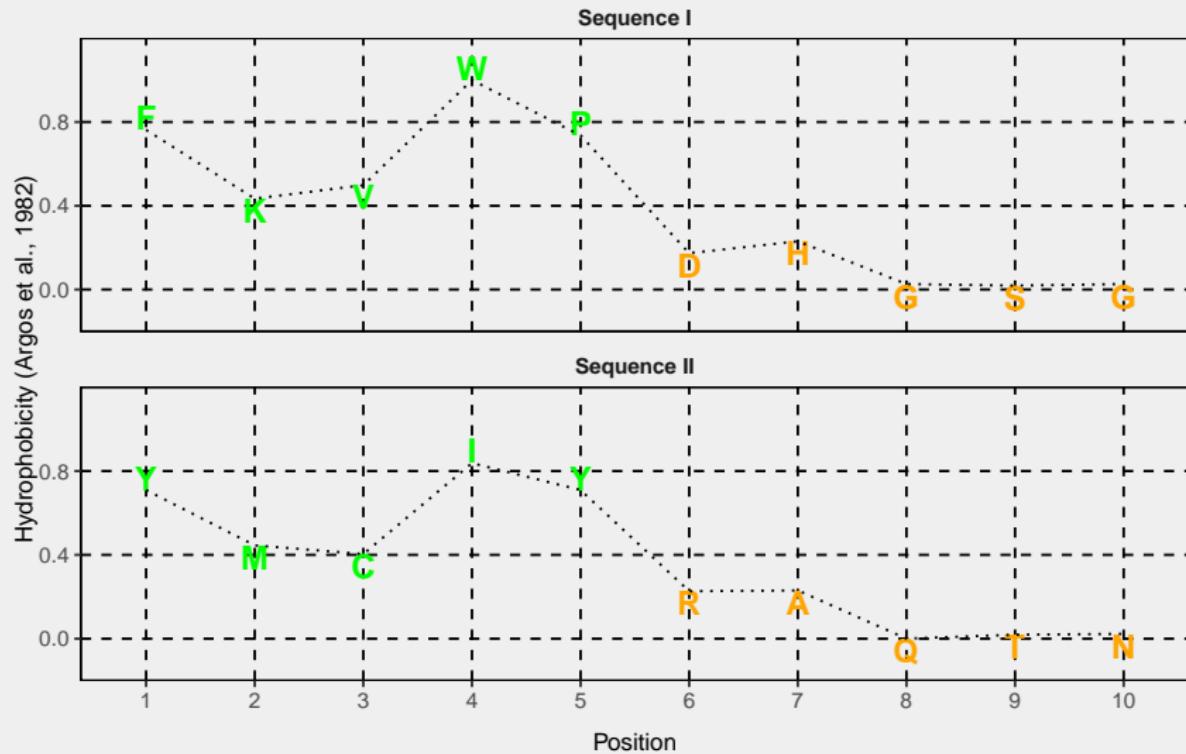
Following peptides appear to be completely different in terms of amino acid composition.

Peptide I:

FKVWPDHGSG

Peptide II:

YMCIYRAQTN



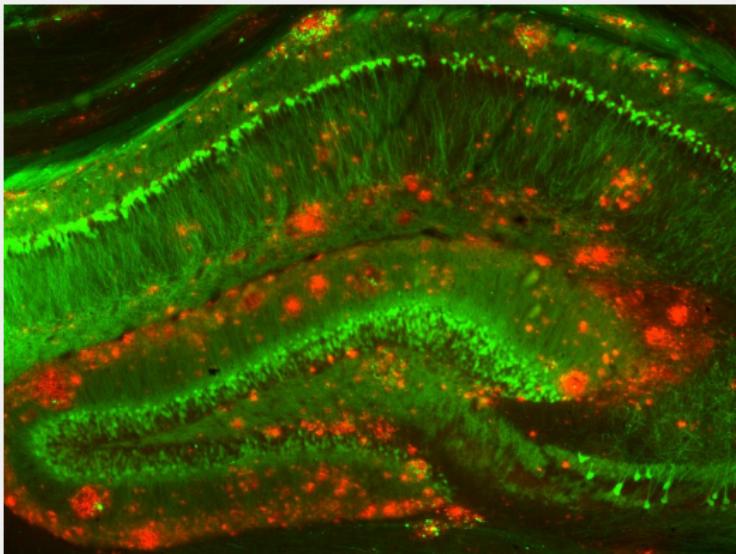
Group	Amino acids
1	C, I, L, K, M, F, P, W, Y, V
2	A, D, E, G, H, N, Q, R, S, T

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

# Amyloid prediction

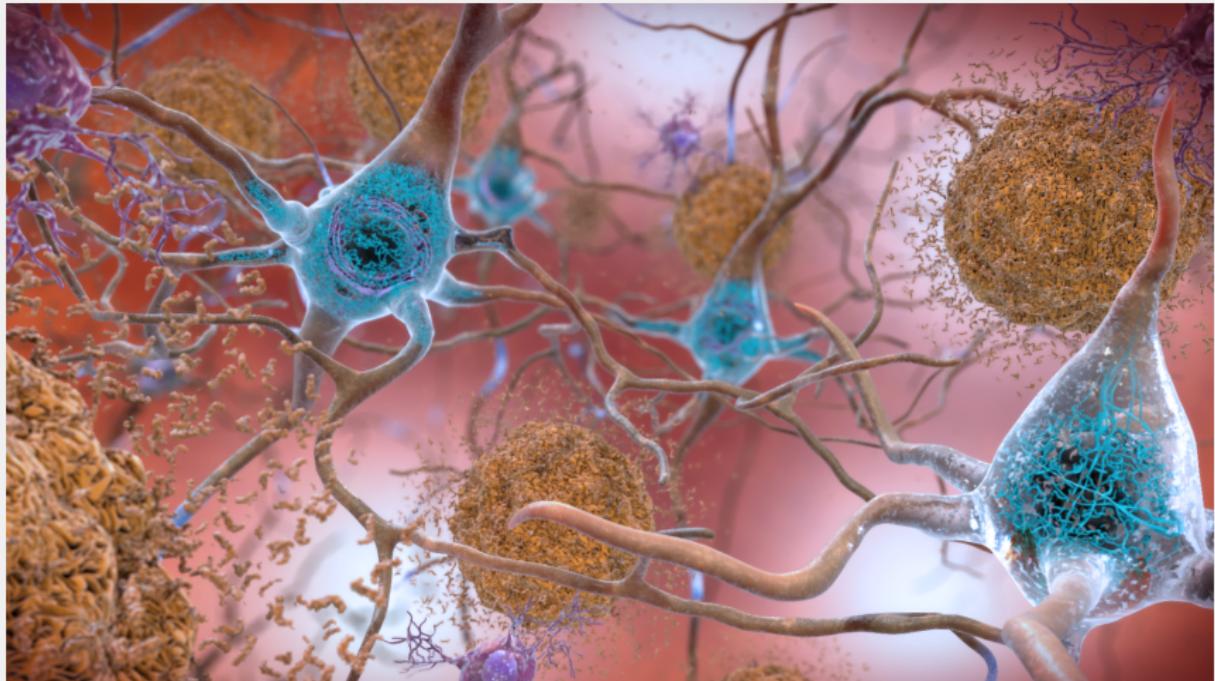
# AMYLOIDS

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

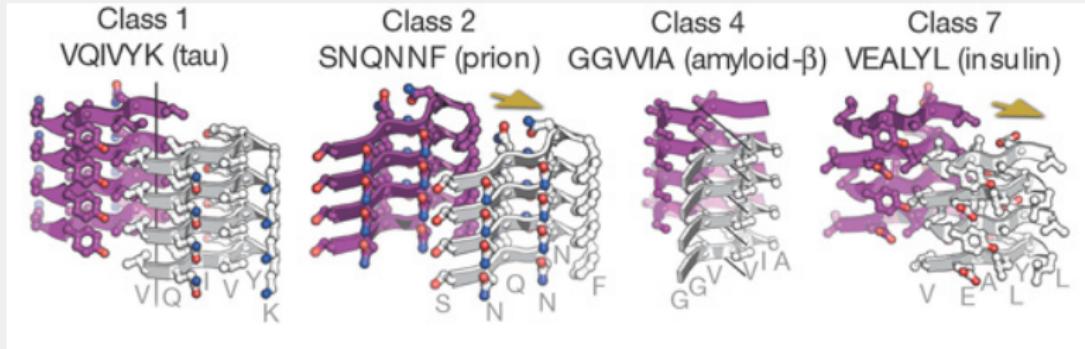


Source: National Institute on Aging (NIA) | National Institutes of Health (NIH)

# AMYLOID PROTEINS

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

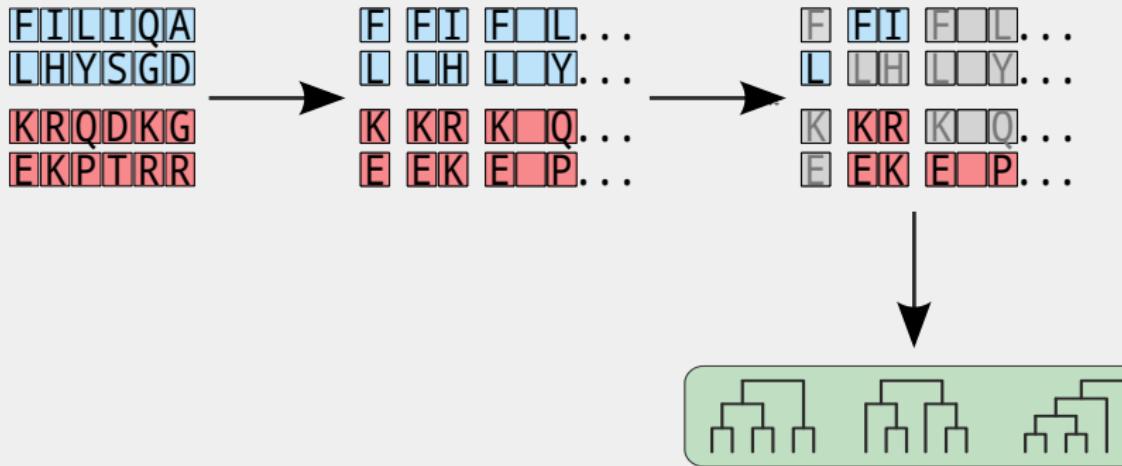
- short (6-15 amino acids),
- very variable, usually hydrophobic amino acid composition,
- create unique  $\beta$ -structures.



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

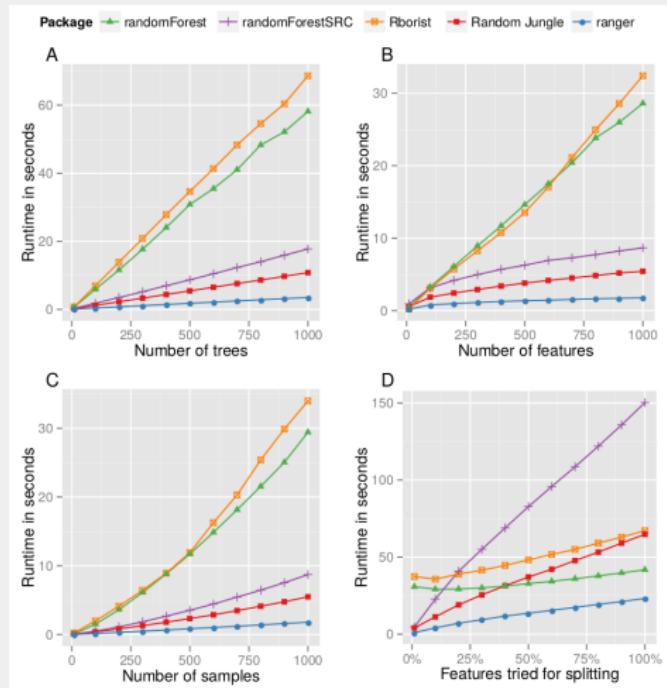
Package	Runtime [h]			Memory usage [GB]
	mtry=	5000	15,000	135,000
randomForest	101.24	116.15	248.60	39.05
randomForest (MC)	32.10	53.84	110.85	105.77
bigrf	NA	NA	NA	NA
randomForestSRC	1.27	3.16	14.55	46.82
Random Jungle	1.51	3.60	12.83	0.40
Rborist	NA	NA	NA	>128
ranger	0.56	1.05	4.58	11.26
ranger (save.memory)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

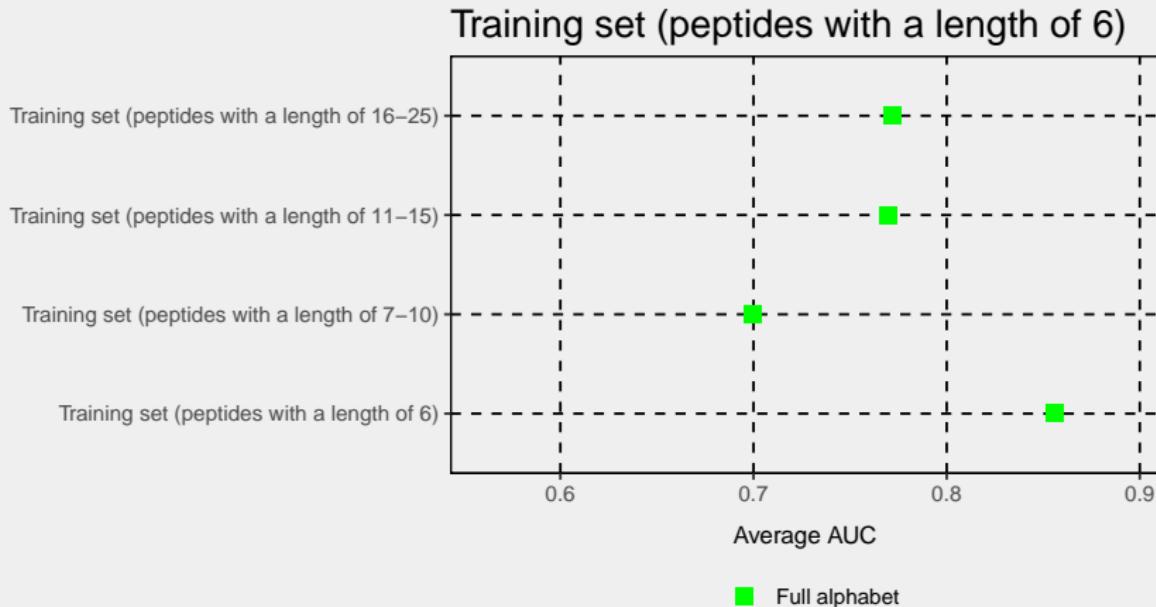
Marvin N. Wright and Andreas Ziegler. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software 1, 77

# 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

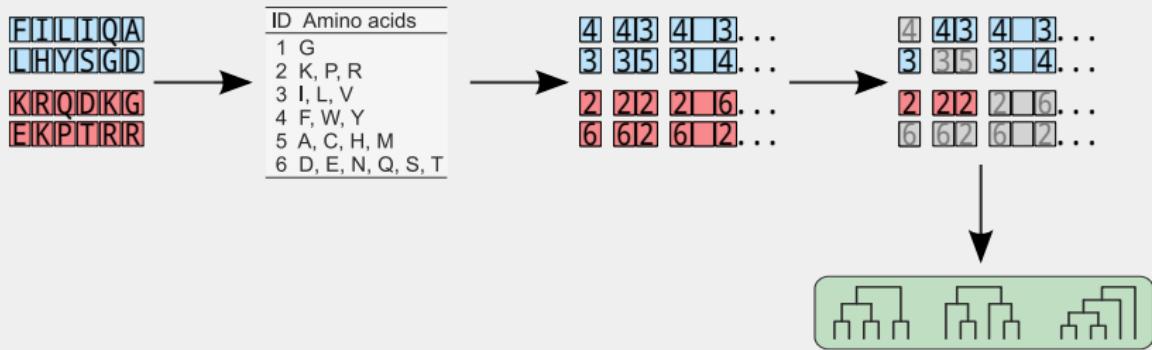


Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

## STANDARD REDUCED ALPHABETS

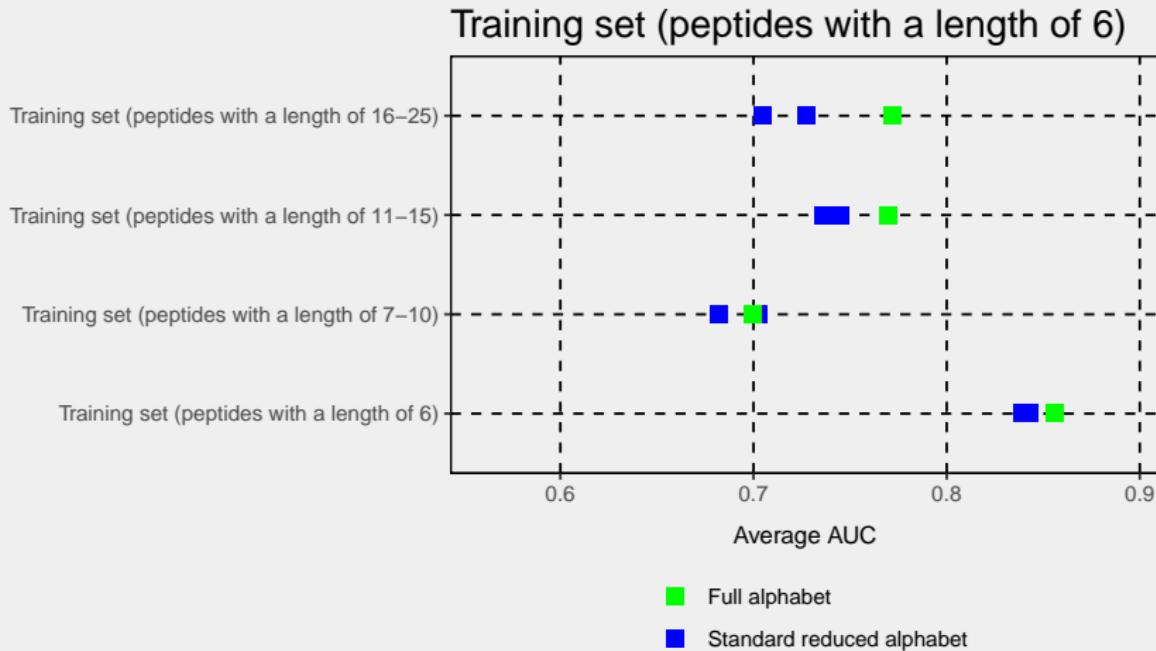
Do standard reduced alphabets developed for different biological issues help to improve amyloid prediction?

# STANDARD REDUCED 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 REDUCED ALPHABET



Standard amino acid alphabets do not improve the quality of amyloid prediction.

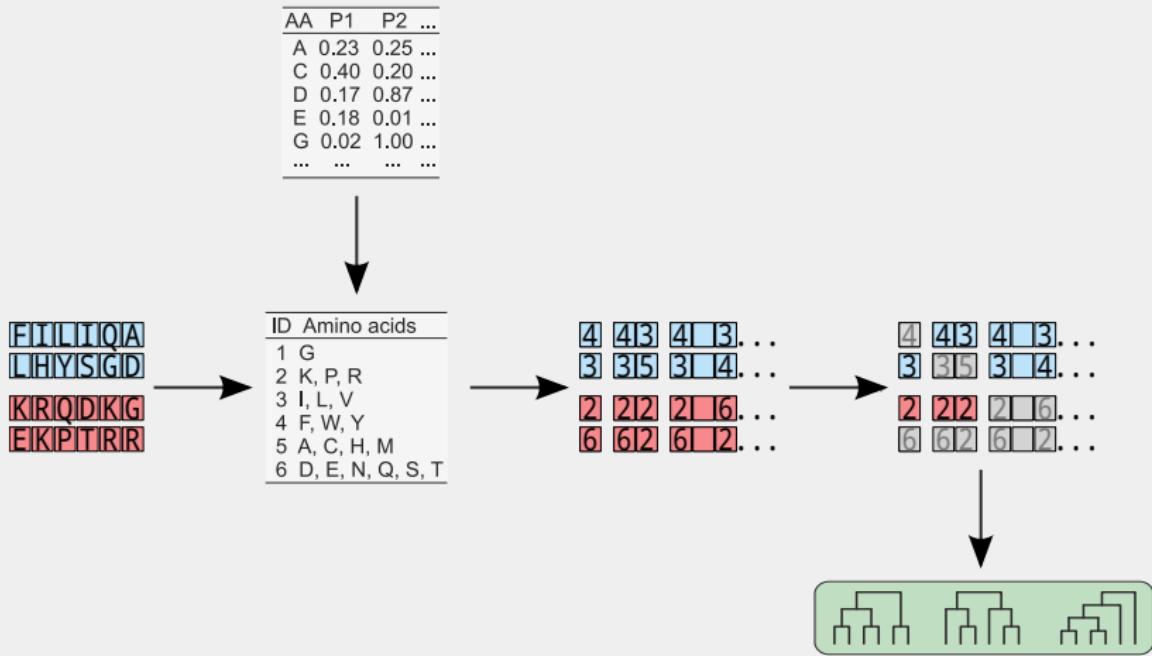
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

# NOVEL REDUCED AMINO ACID ALPHABETS

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

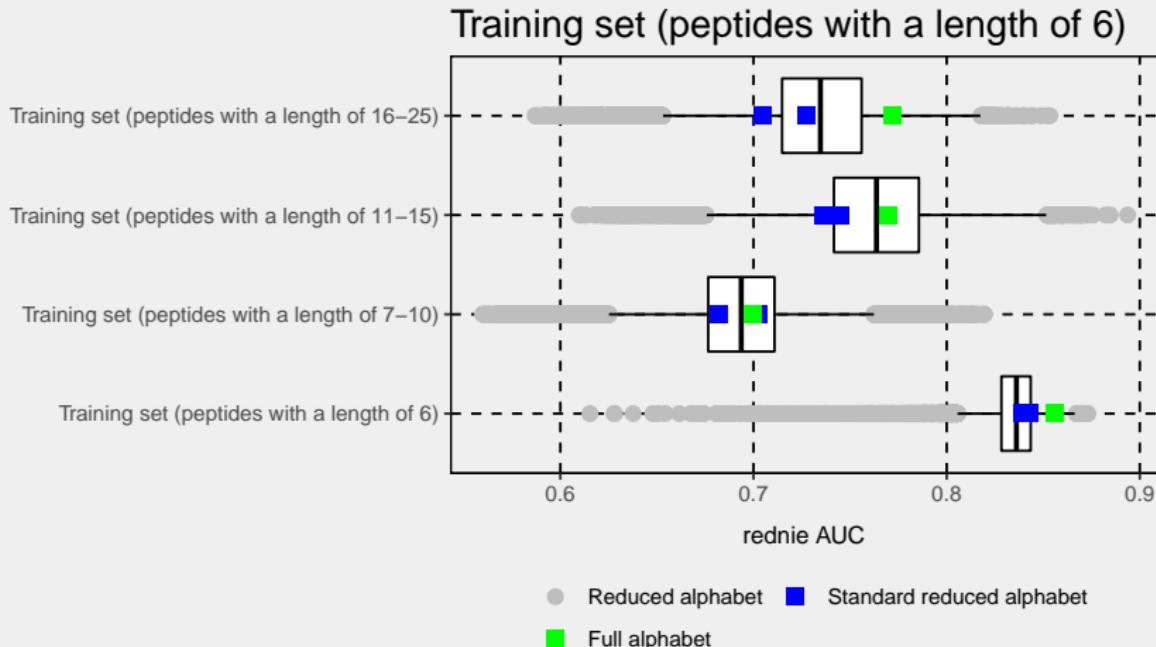
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

# NOVEL REDUCED AMINO ACID ALPHABETS



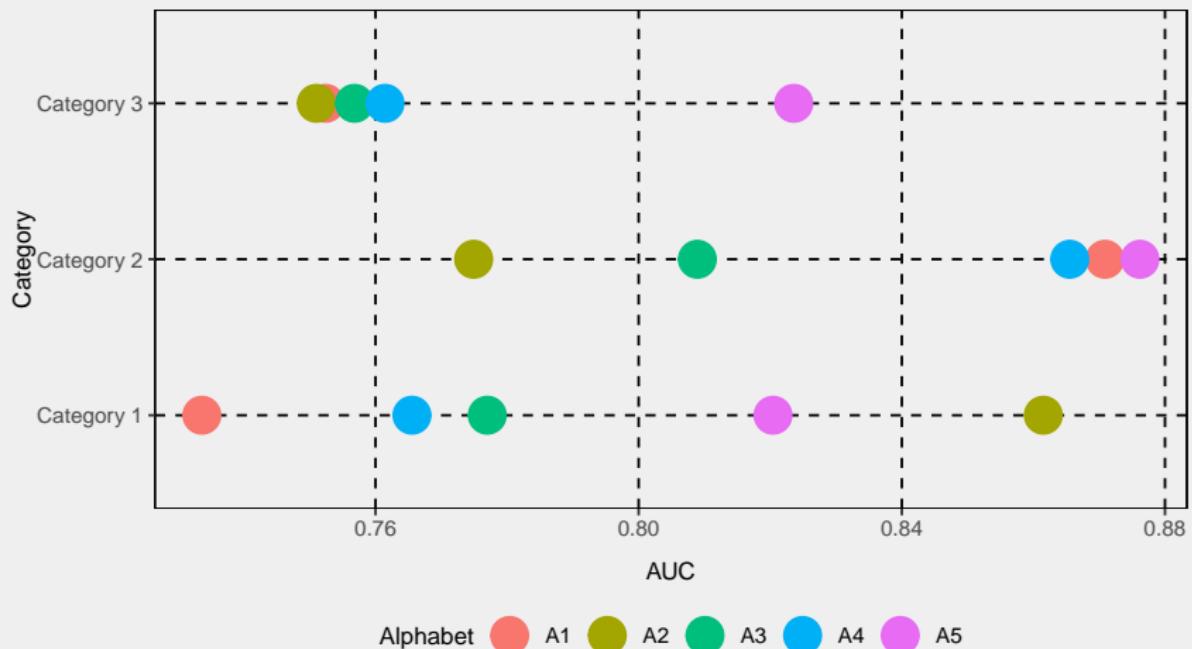
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

# STANDARD REDUCED ALPHABET

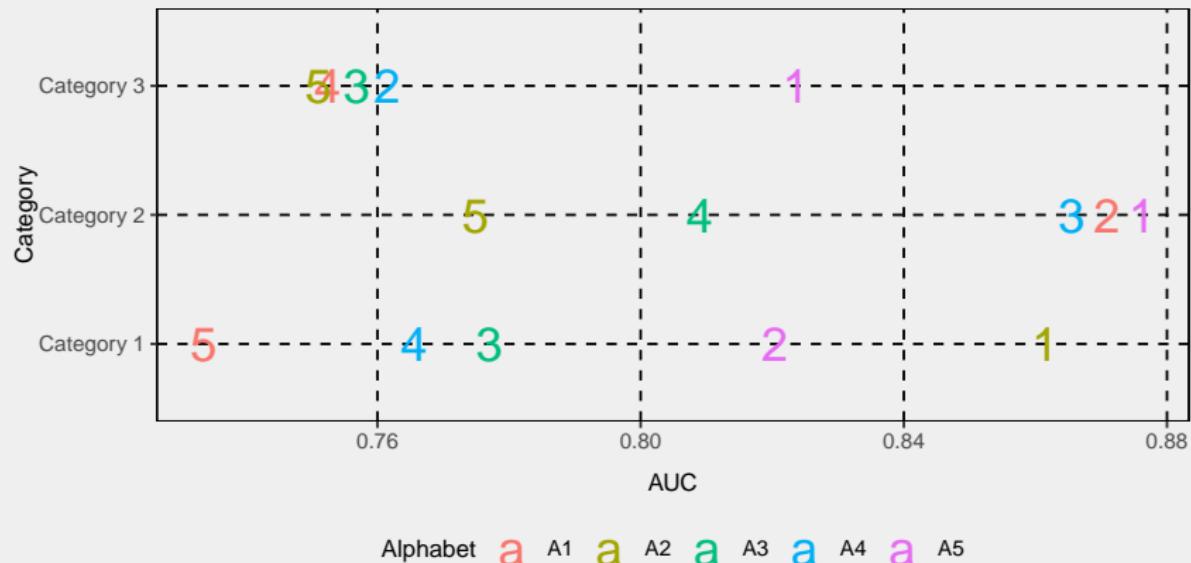


Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

# SELECTION OF BEST-PERFORMING REDUCED ALPHABET

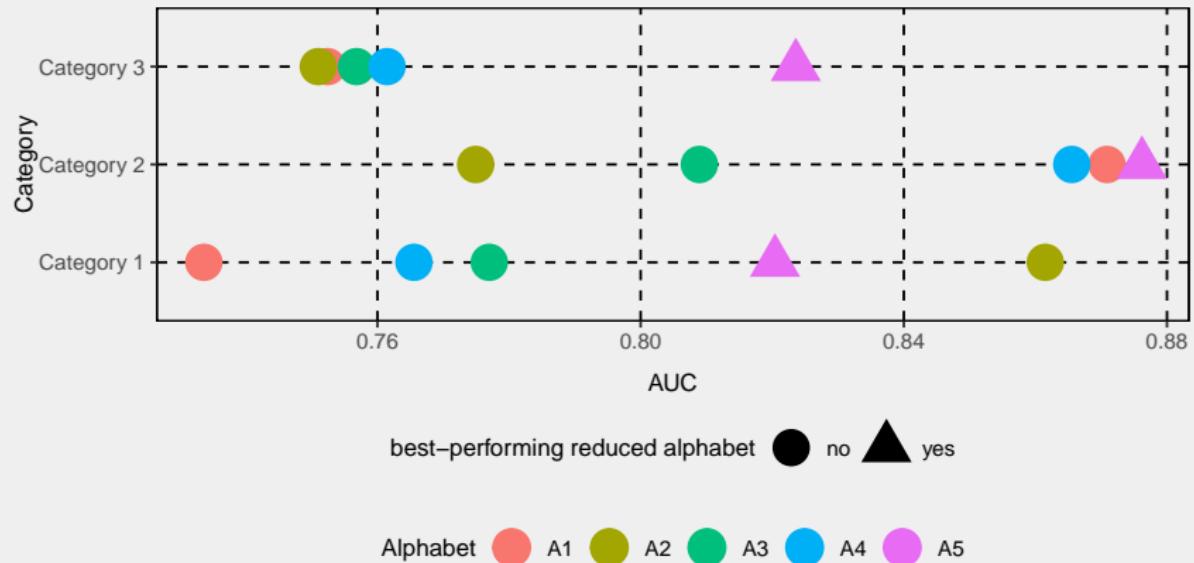


# SELECTION OF BEST-PERFORMING REDUCED ALPHABET



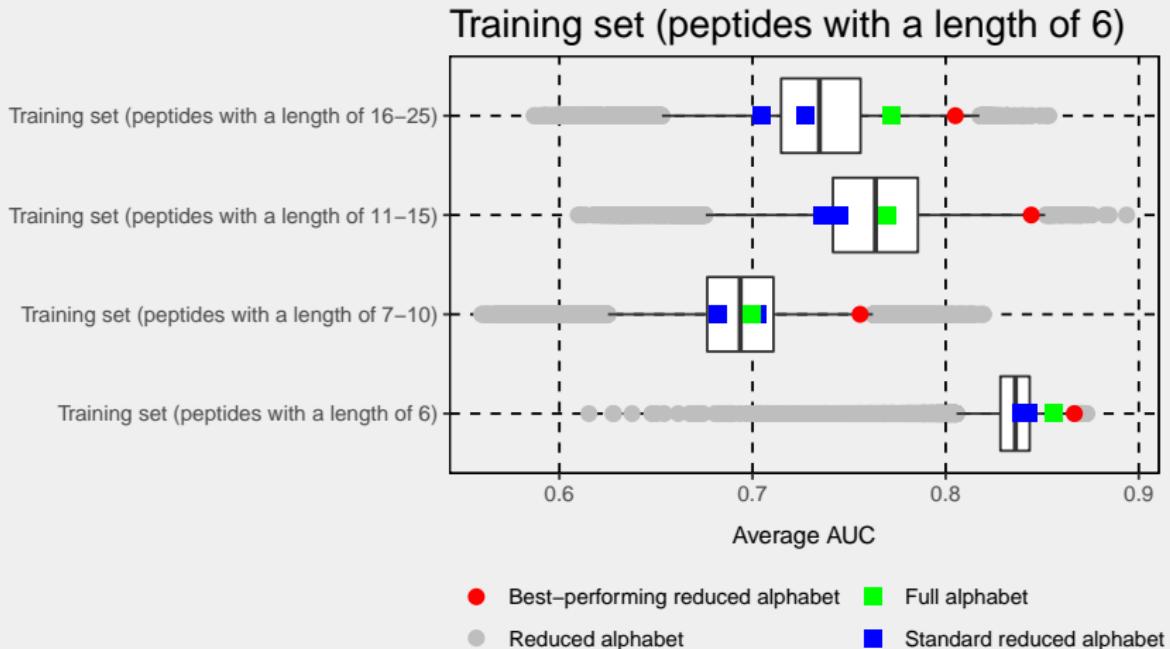
For each category the alphabets have been ranked (rank 1 for the best AUC, etc.).

# SELECTION OF BEST-PERFORMING REDUCED ALPHABET



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

# BEST-PERFORMING REDUCED 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 REDUCED ALPHABET

Group	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

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 REDUCED ALPHABET

Group	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

Grupy 3 i 4 - hydrophobic amino acids.

# BEST-PERFORMING REDUCED ALPHABET

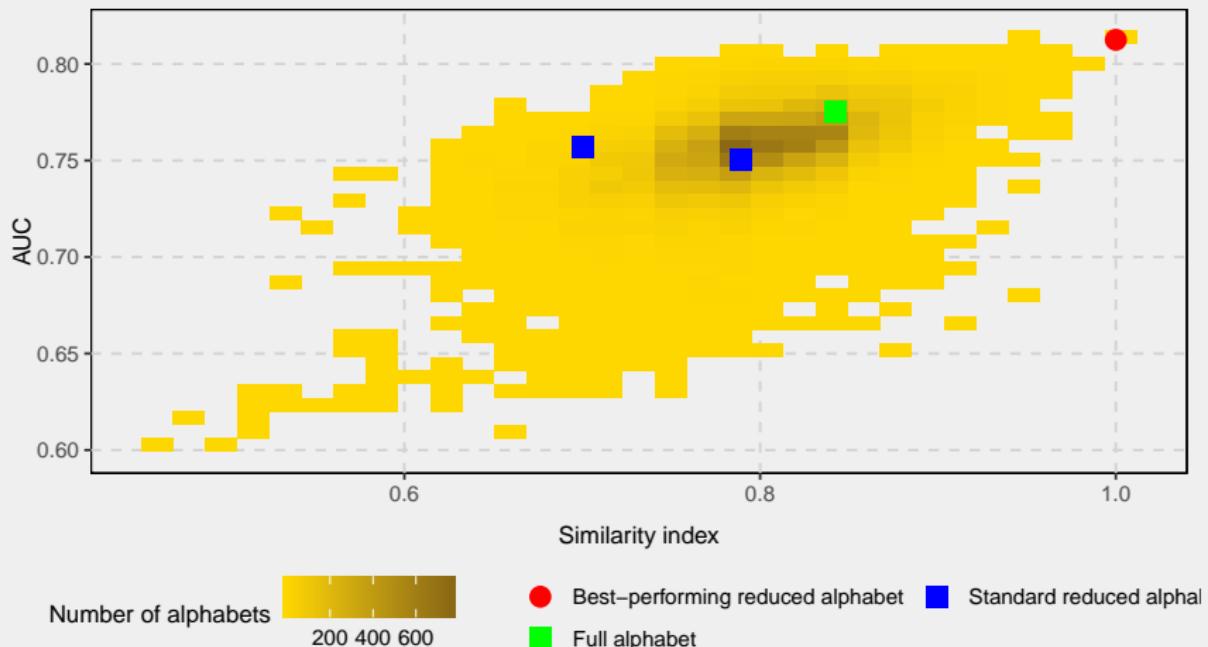
Group	Amino acids
1	G
2	K, P, R
3	I, L, V
4	F, W, Y
5	A, C, H, M
6	D, E, N, Q, S, T

Group 2 - amino acids disrupting the  $\beta$ -structure ( $\beta$ -breakers).

## ALPHABET SIMILARITY AND QUALITY OF PREDICTION

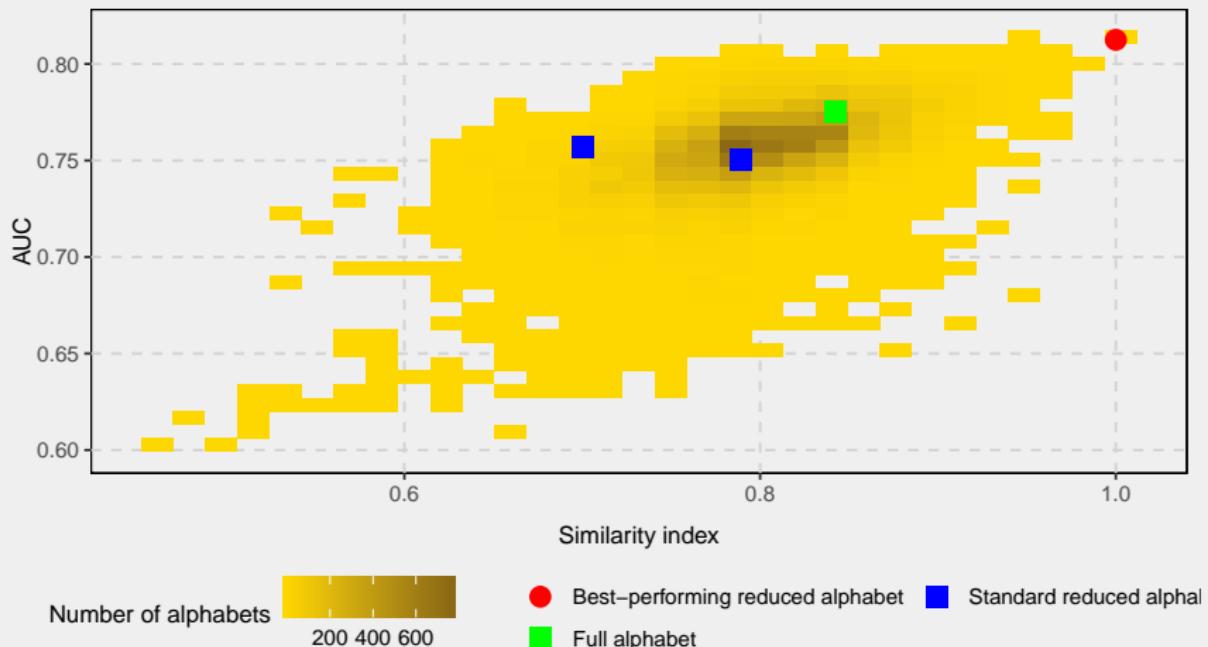
Is the best-performing reduced amino acid alphabet associated with amyloidogenicity?

# SIMILARITY INDEX



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

# SIMILARITY INDEX

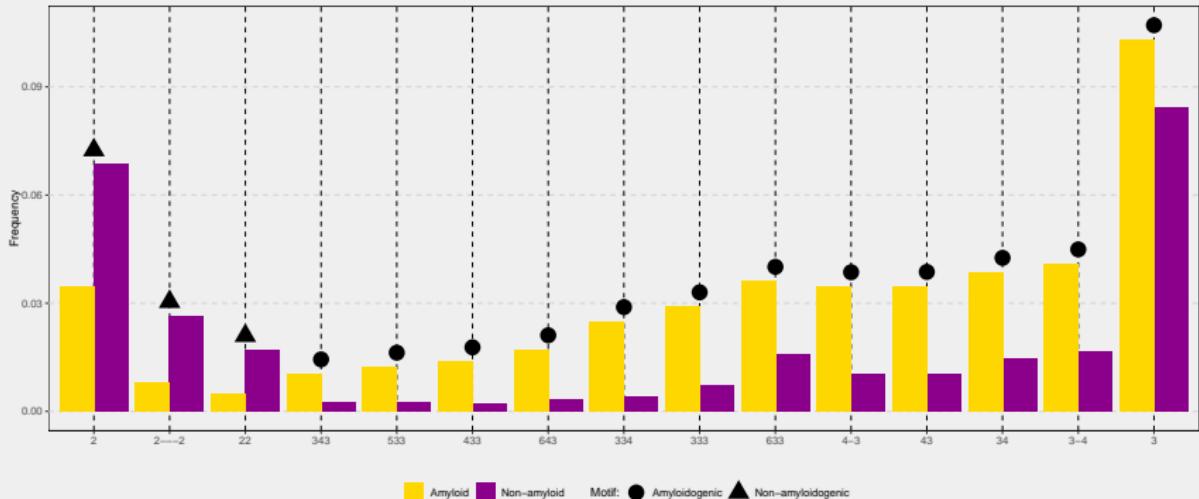


The correlation between the similarity index and the average AUC is important ( $p\text{-value} \leq 2.2^{-16}$ ;  $\rho = 0.51$ ).

Burdzkiekiewicz, 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

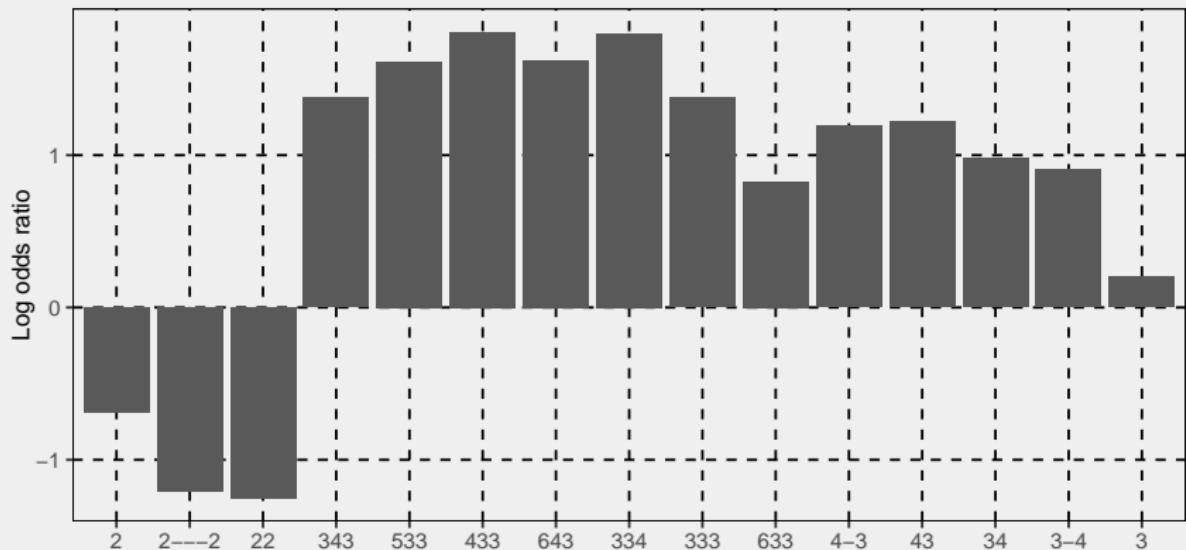
Are informative n-grams found by QuiPT associated with amyloidogenicity?

# 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).

# 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).

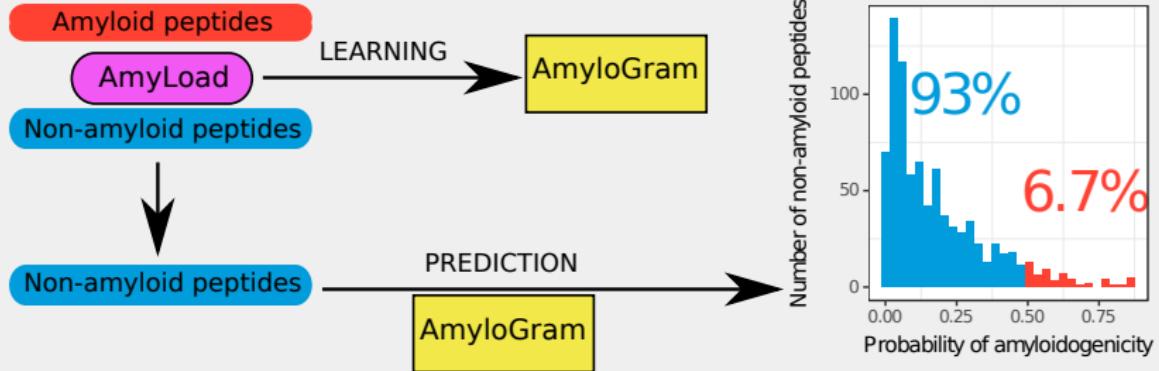
Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* 7, 12961

## BENCHMARK RESULTS

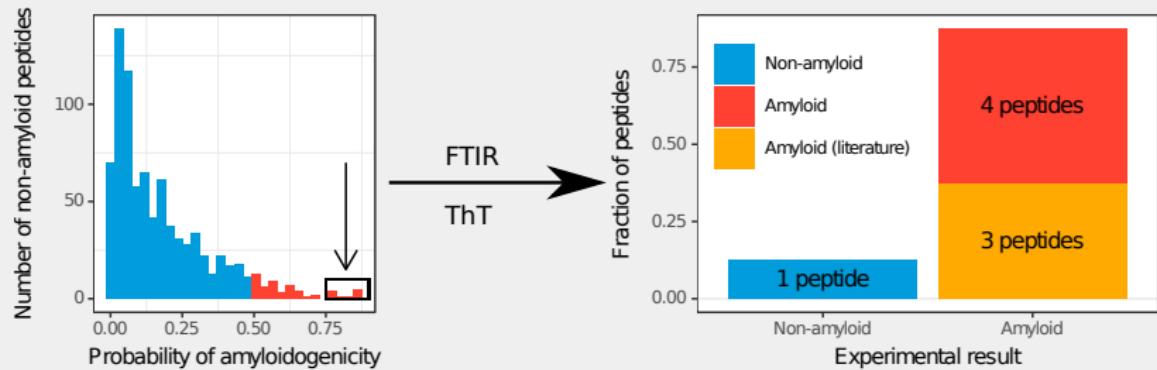
Program	AUC	MCC
AmyloGram	0.8972	0.6307
PASTA 2.0 (Walsh et al., 2014)	0.8550	0.4291
FoldAmyloid (Garbuzyntsiy et al., 2010)	0.7351	0.4526
APPNN (Família et al., 2015)	0.8343	0.5823

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

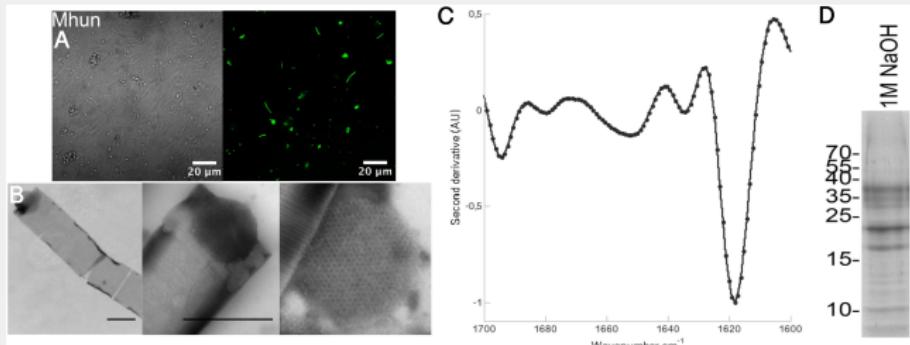
# EXPERIMENTAL VALIDATION



# 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 application

# 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: Michał Burdakiewicz, Piotr Sobczyk.

Citation: Burdakiewicz, 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  
>Ado-2h  
VPSNEEQIKNLLQLEAQEHLQY  
>AMY138|Alpha 6(Glutathione S Transeferase P domain  
QISFADVNLLDLRIHQVLN  
>AMY143|M8|Spectrin SH3  
DILTLNLNSTNKDWKKVEVND  
>CsgA|region 1  
SELNIYQYGGGN SALALQTDARN
```

Paste sequences (FASTA format required) here...

Submit data from field above

Submit .fasta or .txt file:

Browse... No file selected

# 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:** Michał Burdakiewicz, Piotr Sobczyk.

**Citation:** Burdakiewicz, 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).

<b>Cutoff</b>	Sensitivity: 0.8658 Specificity: 0.7852 MCC: 0.6268
---------------	---

[Start a new query](#)

Results (tabular) [Detailed results](#)

Copy CSV Excel Print

Input name	Amyloid probability	Is amyloid?
All	All	All
AMY133 Alpha	0.6725	yes
Ada-2h	0.4702	no
AMY138 Alpha	0.7515	yes
AMY143 M8 Spectrin	0.6488	yes
CsgA region	0.8216	yes

Showing 1 to 5 of 5 entries

Previous 1 Next

# 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: Michał Burdakiewicz, Piotr Solcyna.

Citation: Burdakiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Markiewicz, 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 residue).

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

All

II

Fraction of amyloid residues

All

AMY1331Alpha

0.6364

Ada-2h

0.0000

AMY138Alpha

0.7500

AMY141MBISpectrum

0.3500

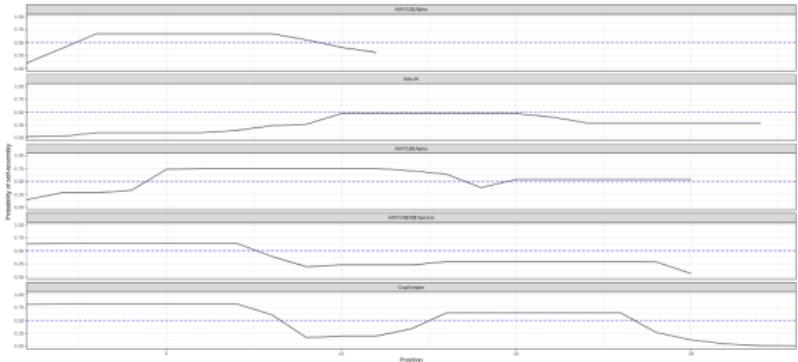
CspARegion

0.6087

Showing 1 to 5 of 5 entries

Previous 1 Next

### Amyloid regions



# SUMMARY

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>.

# SUMMARY

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 |

- 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

- Garbuzyanskiy, 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. 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. , Riek, 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.