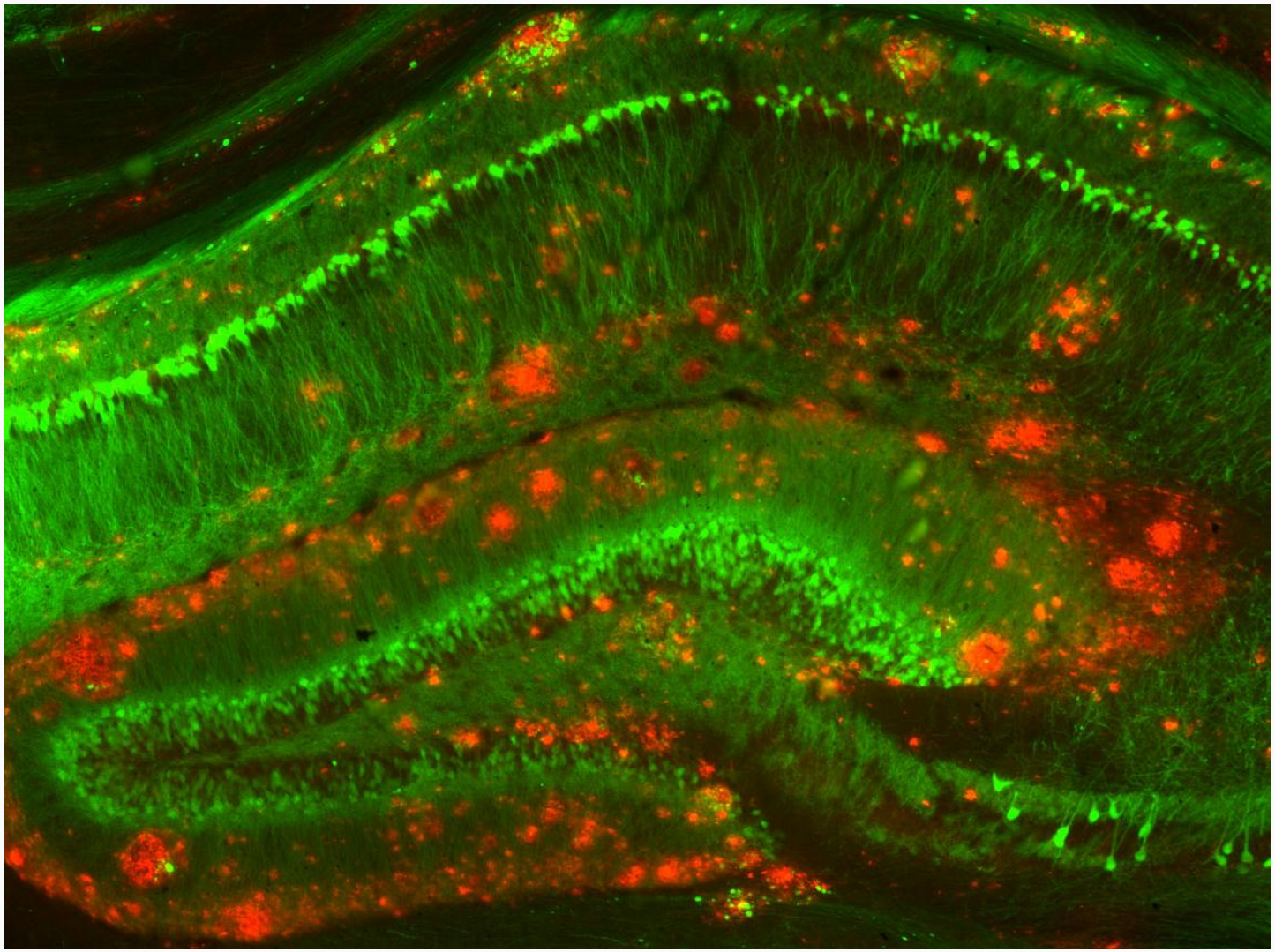
AmyloGram:a novel predictor of amyloidogenicity

Michał Burdukiewicz[1], Piotr Sobczyk[2], Stefan Rödiger[3], Paweł Mackiewicz[1] and Małgorzata Kotulska[4]

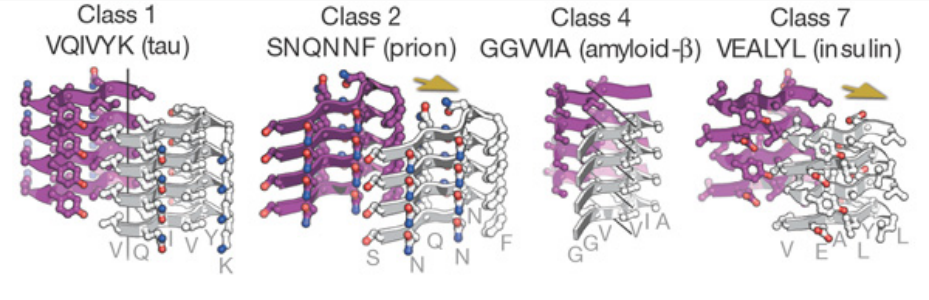
[1]University of Wrocław, Department of Genomics,
[2]Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics,
[3]Brandenburg University of Technology Cottbus-Senftenberg, Institute of Biotechnology,
[4]Wrocław University of Science and Technology, Department of Biomedical Engineering

Amyloids

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



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



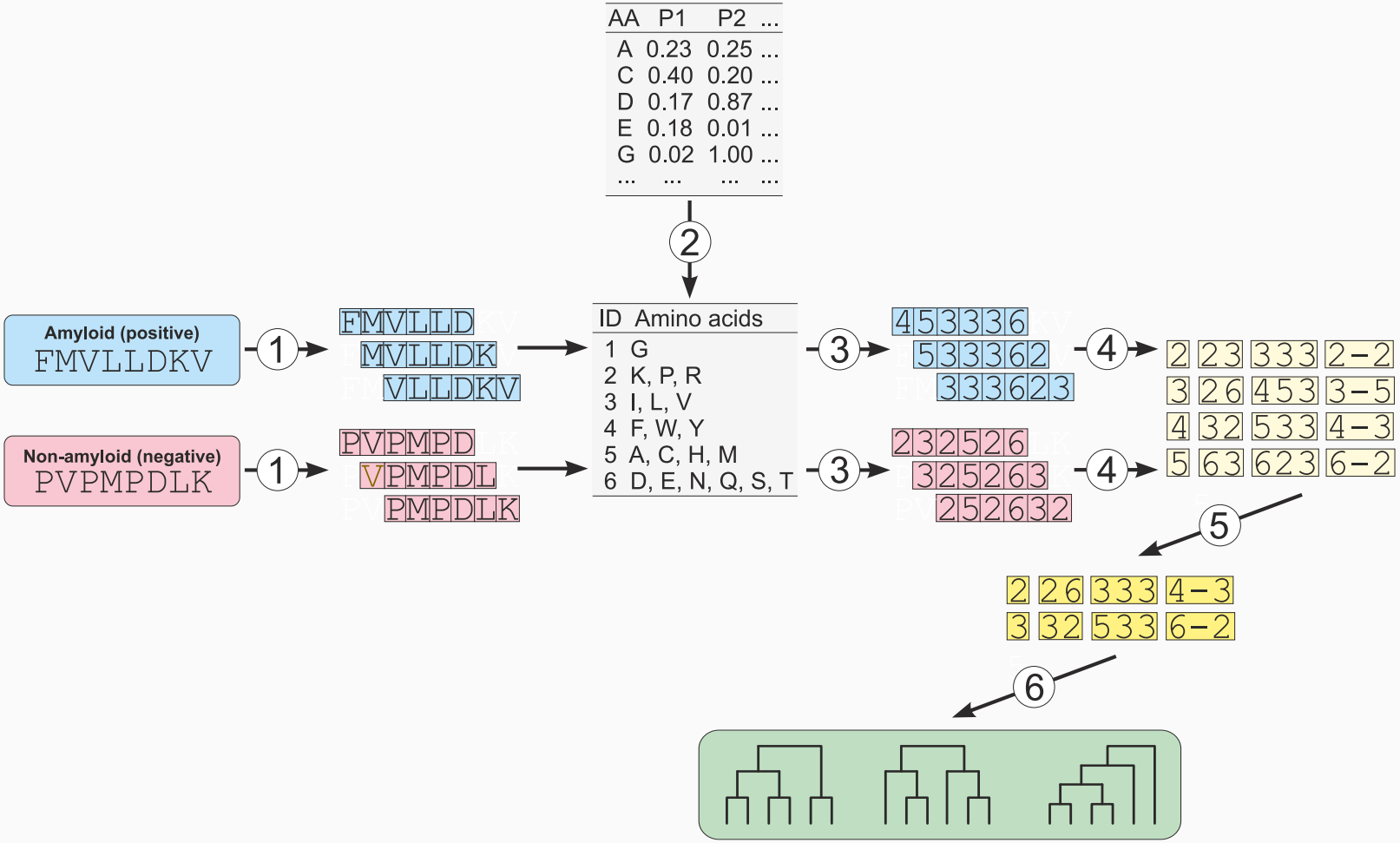
Sawaya et al. (2007)

Aim

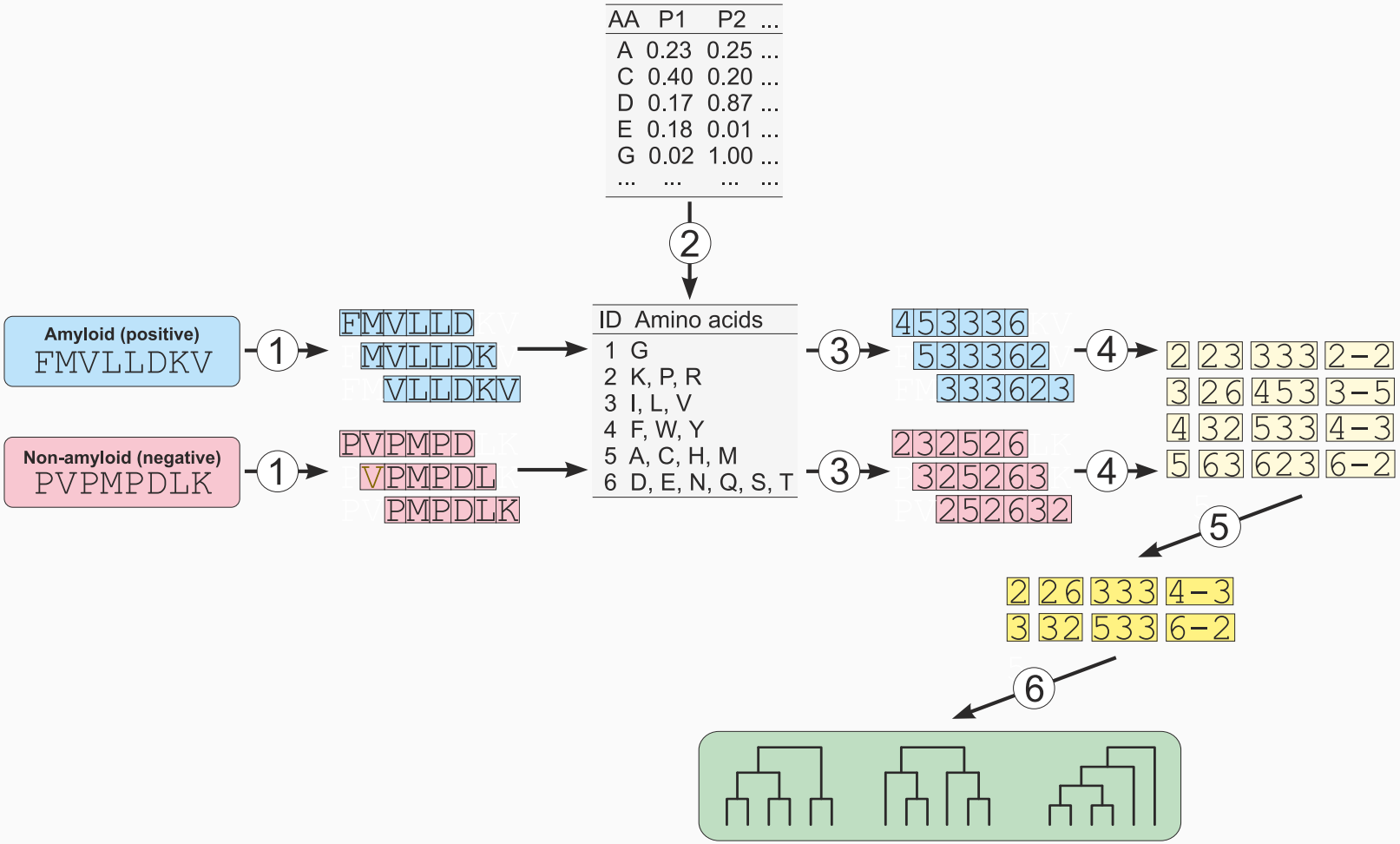
Analize structure of hot spots and create a novel predictor of amyloids.

- Does amyloidogenicity depend on the exact sequence of amino acids?
- Which motifs are associated with amyloidogenicity?

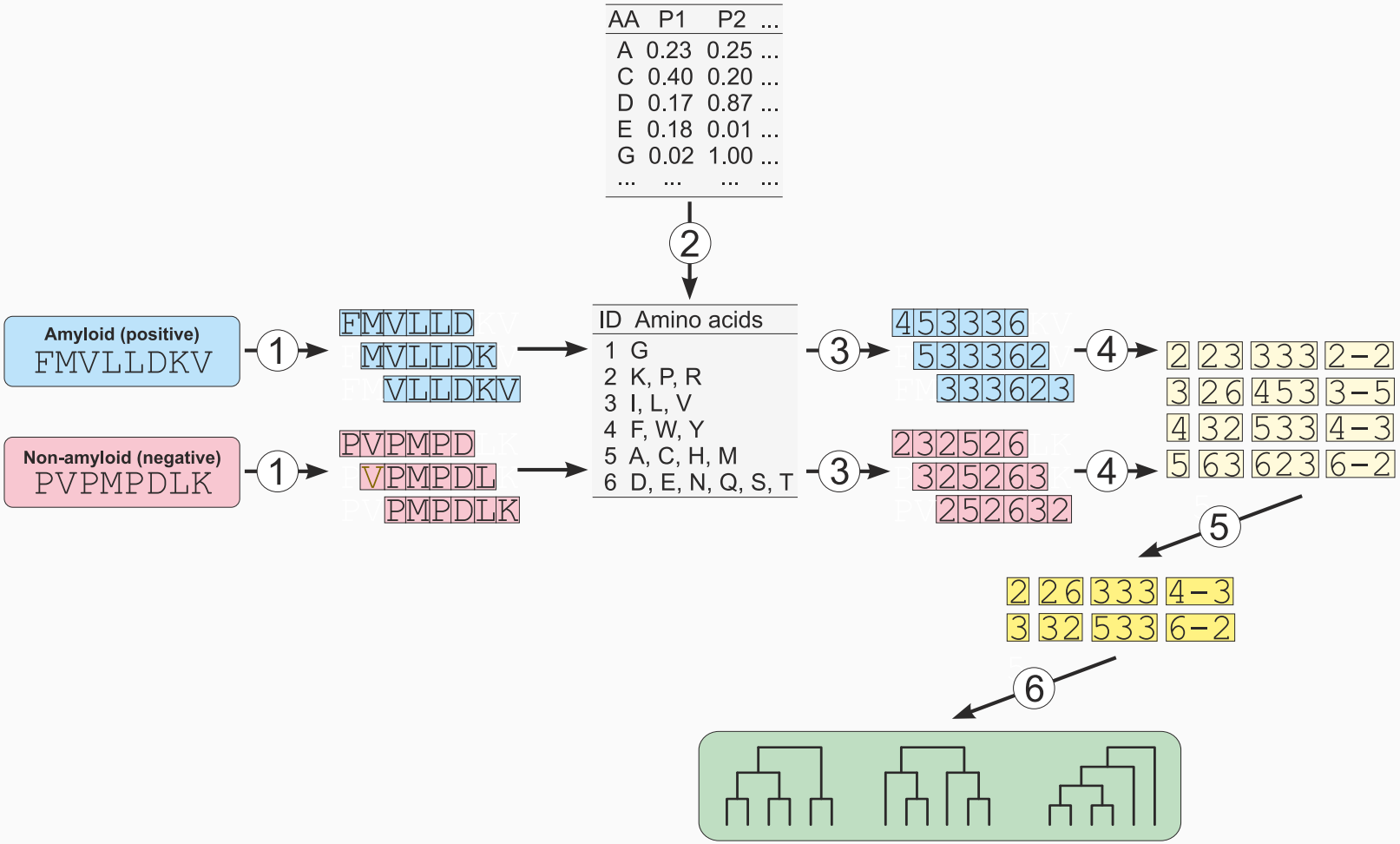
Learning framework



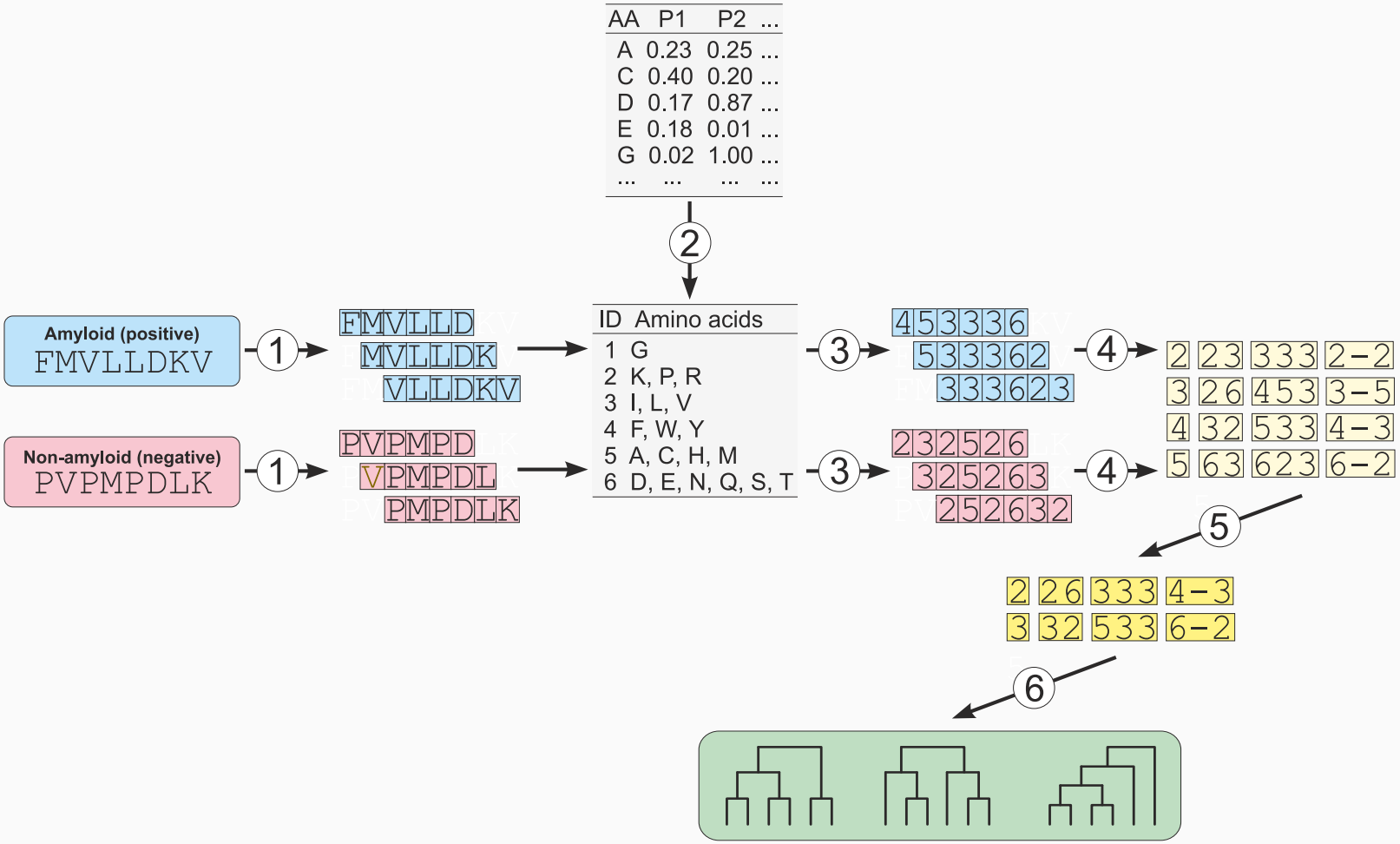
1. Extraction of overlapping hexamers from peptides with known amyloidicity status.



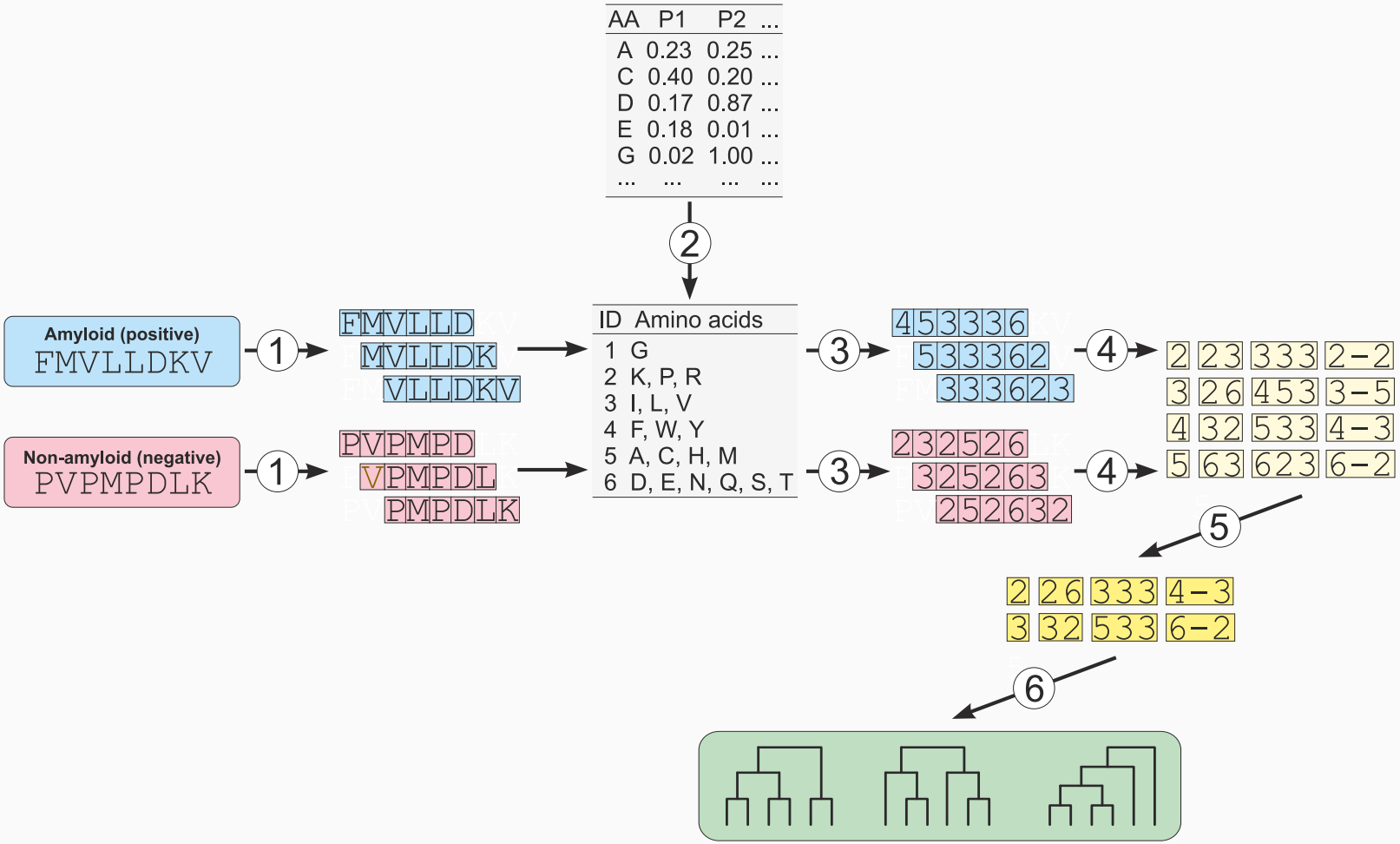
2. Clusterization of amino acids (AA) into groups (ID) using a combination of various physicochemical properties (P1, P2, ...).



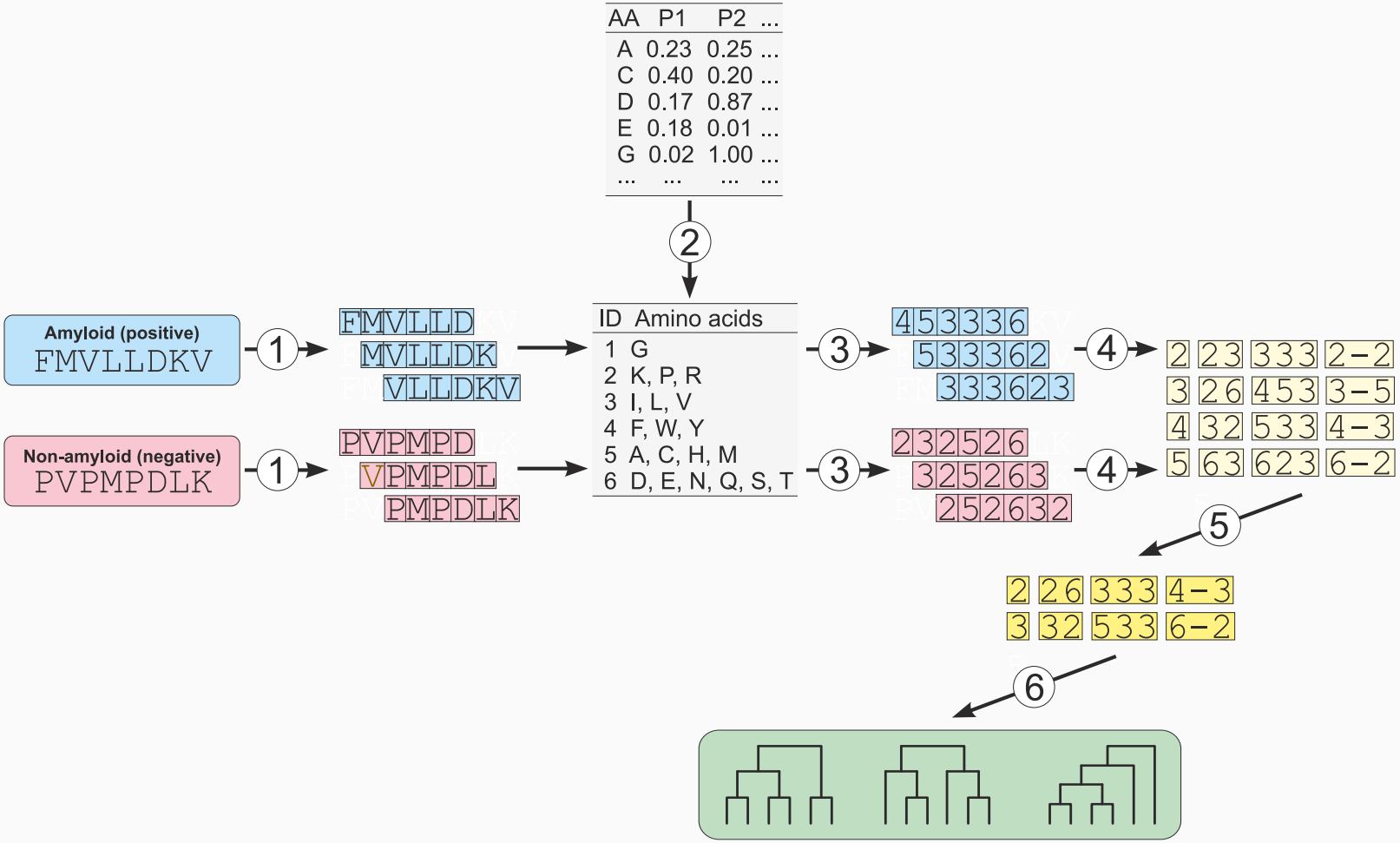
3. Encoding amino acids of hexamers into corresponding groups (reduced alphabet).



4. Extraction of encoded n-grams of different types.



5. Selection of informative n-grams using Quick Permutation Test (QuiPT).



6. Cross-validation of encodings using random forest classifier, which is trained on the informative n-grams.

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).

Quick Permutation Test

Informative n-grams are usually selected using permutation tests.

During a permutation test we shuffle randomly class labels and compute a defined statistic (e.g. information gain). Values of statistic for permuted data are compared with the value of statistic for original data.

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

$N_{T_P > T_R}$: number of cases, where T_P (permuted test statistic) has more extreme values than T_R (test statistic for original data).

N: number of permutations.

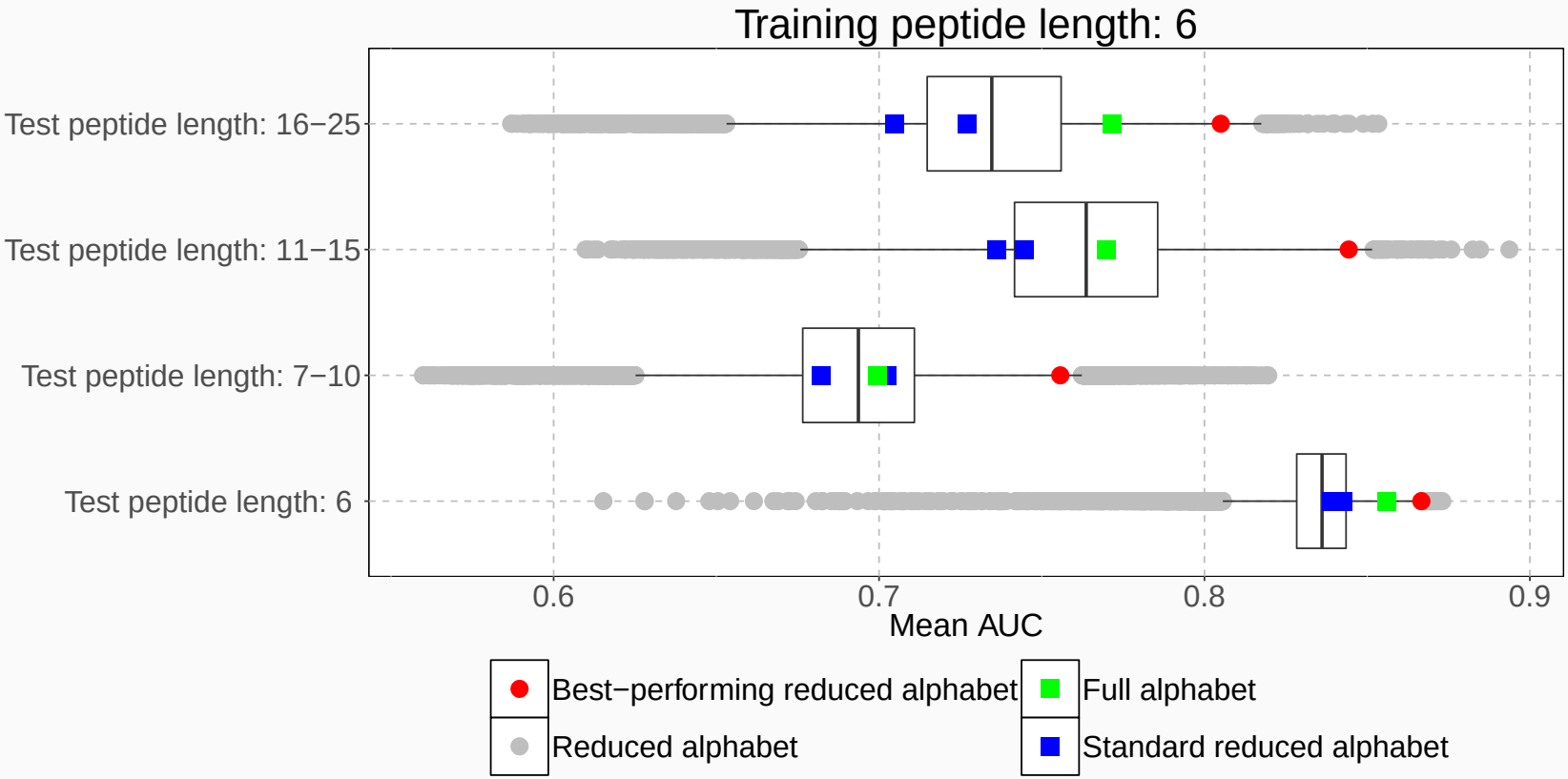
QuiPT

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

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

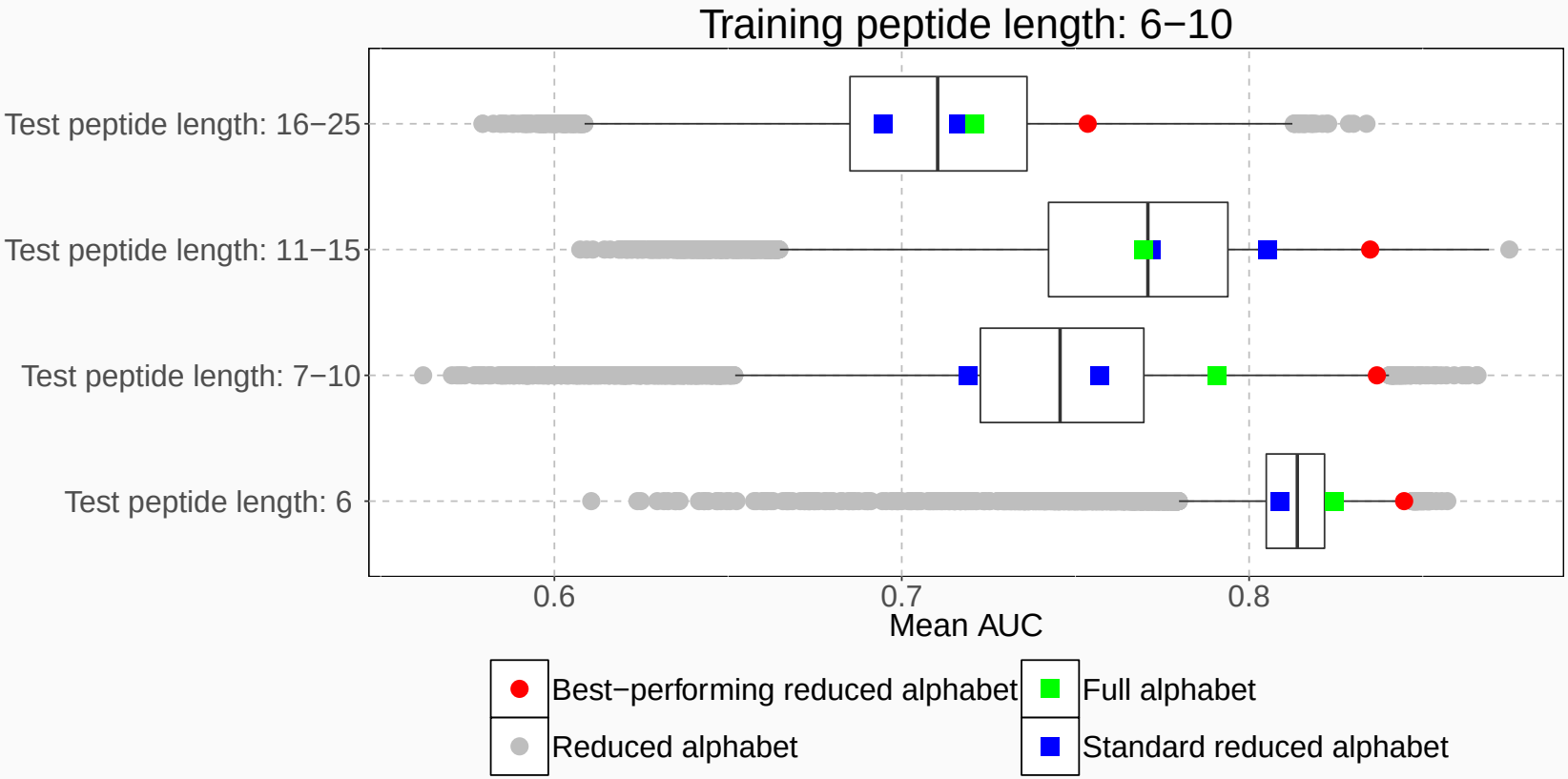
Results

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.

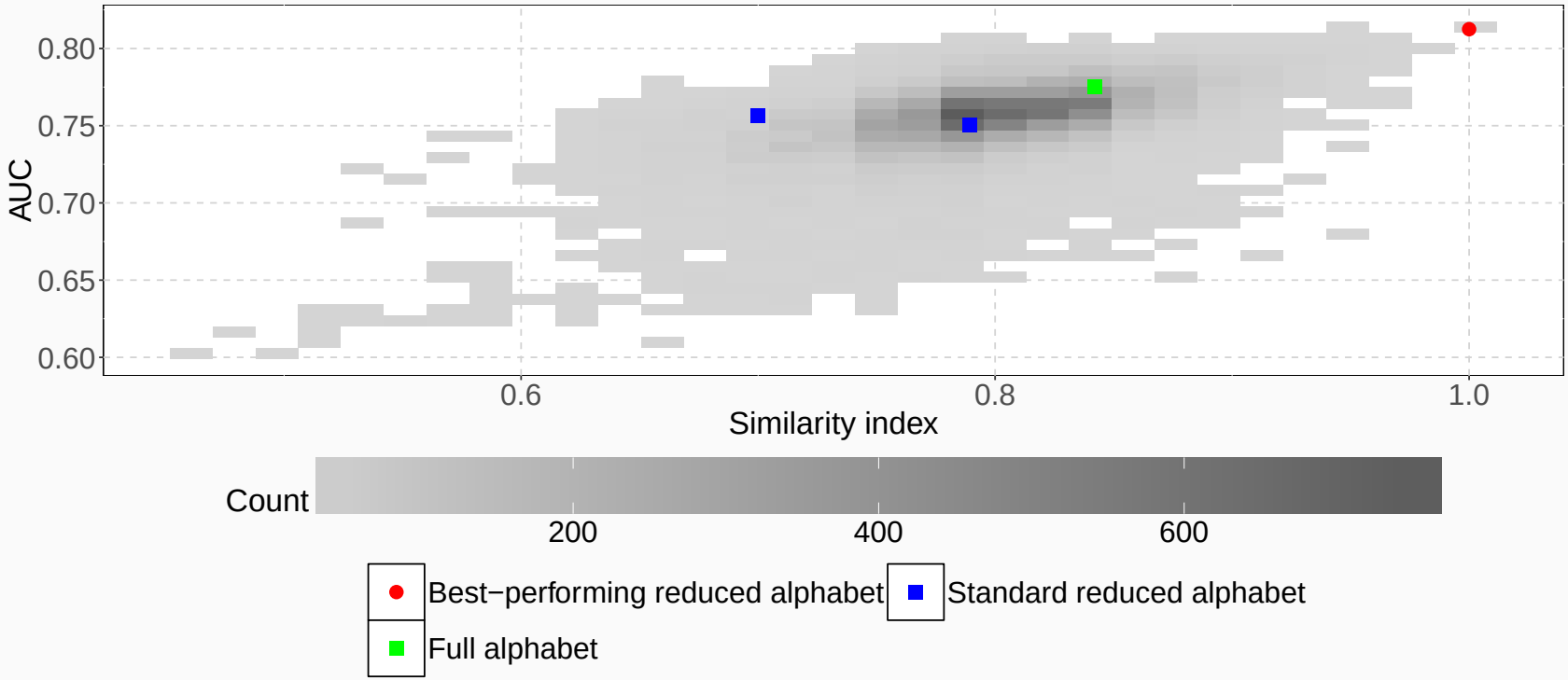
Cross-validation



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.

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

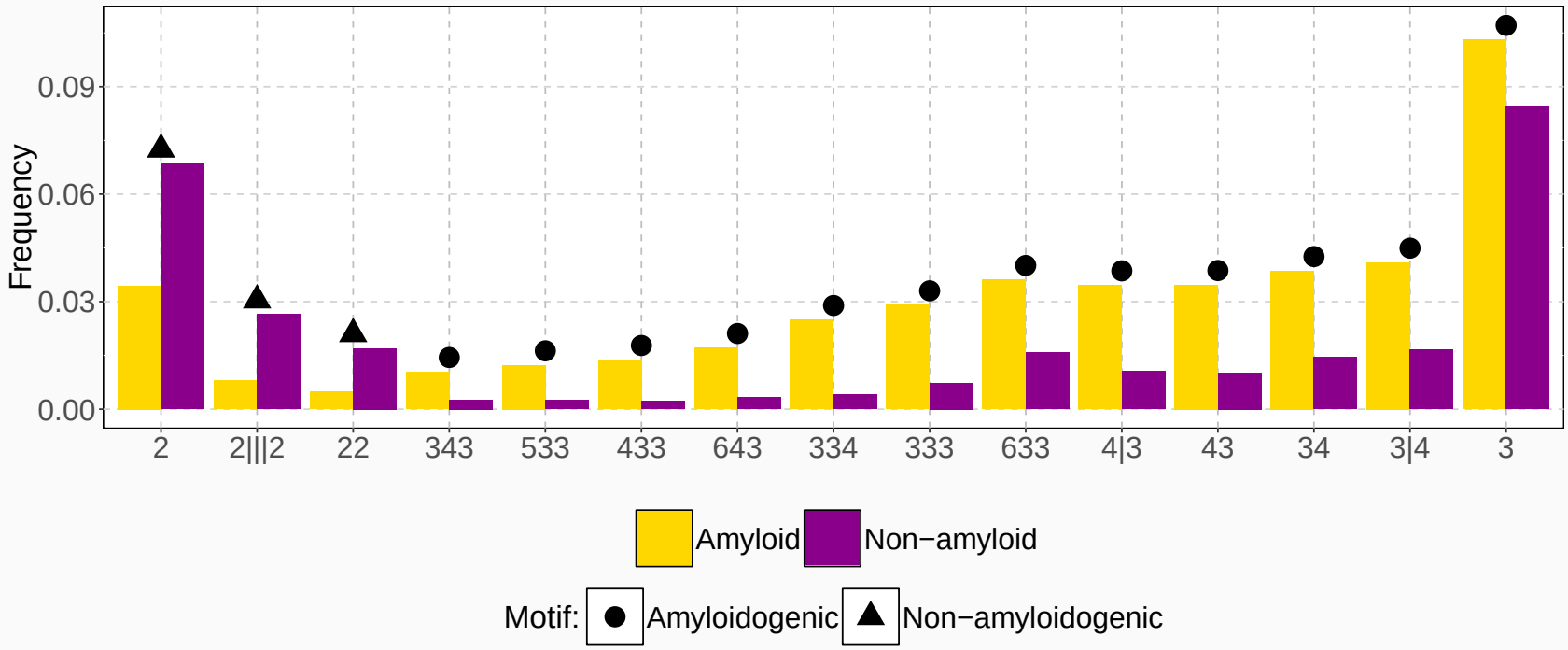
Similarity index



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

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).

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: `smorfland.uni.wroc.pl/amylogram`.

Funding

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

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

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 III

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.