

## Prediction of amyloidogenicity based on the n-gram analysis

<sup>1</sup>University of Wrocław, Department of Genomics, <sup>2</sup>Wrocław University of Technology, Department of Mathematics and <sup>3</sup>Wrocław University of Technology, Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology

Associate Editor: XXXXXXXX

## Abstract

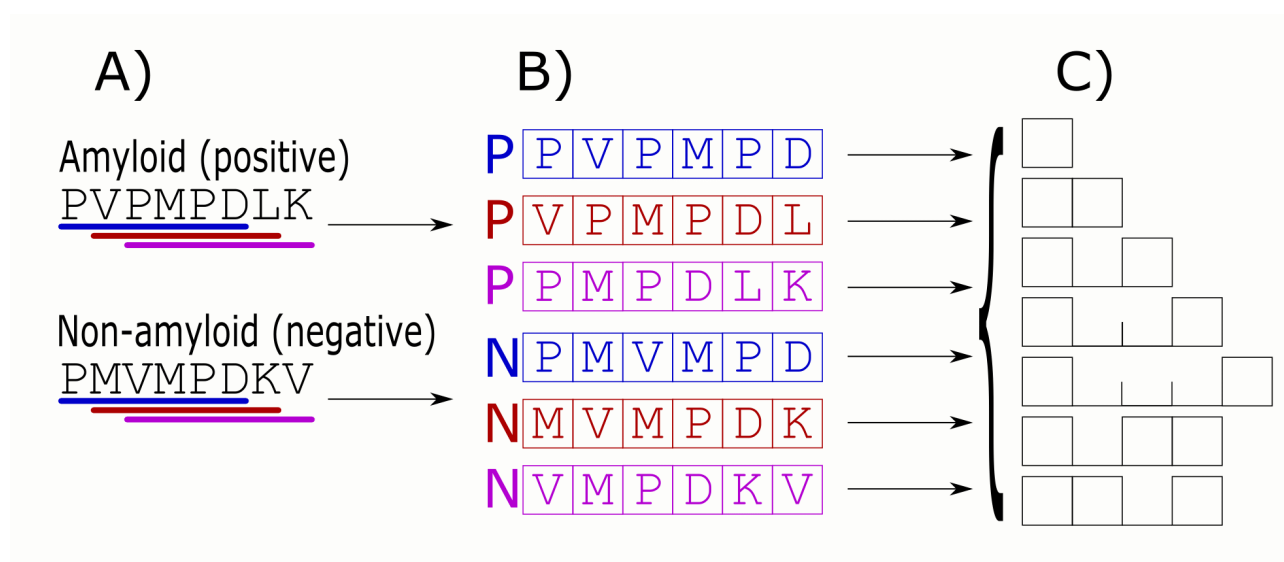
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

### 3.1 Data set

During the training phase, we extracted overlapping hexamers from each sequence. Each hexamer was tagged with the same etiquette (amyloid/nonamyloid) as the original peptide. For example, the sequence of length 6 residues yields only one hexamer and the sequence of 8 residues

Category	Property
Contactivity	Average flexibility indices (Bhaskaran and Ponnuswamy, 1988)
Contactivity	14 Å contact number (Nishikawa and Ooi, 1986)
Contactivity	Accessible surface area (Radzicka and Wolfenden, 1988)
Contactivity	Buriability (Zhou and Zhou, 2004)
Contactivity	Values of Wc in proteins from class Beta, cutoff 12 Å, separation 5 (Wozniak and Kotulska, 2014)
Contactivity	Values of Wc in proteins from class Beta, cutoff 12 Å, separation 15 (Wozniak and Kotulska, 2014)
$\beta$ -frequency	Average relative probability of inner beta-sheet (Kanehisa and Tsong, 1980)
$\beta$ -frequency	Relative frequency in beta-sheet (Prabhakaran, 1990)
$\beta$ -frequency	Thermodynamic beta sheet propensity (Kim and Berg, 1993)
Hydrophobicity	Hydrophobicity index (Argos <i>et al.</i> , 1982)
Hydrophobicity	Optimal matching hydrophobicity (Sweet and Eisenberg, 1983)
Hydrophobicity	Hydrophobicity-related index (Kidera <i>et al.</i> , 1985)
Hydrophobicity	Scaled side chain hydrophobicity values (Black and Mould, 1991)
Polarity	Polarizability parameter (Charton and Charton, 1982)
Polarity	Mean polarity (Radzicka and Wolfenden, 1988)
Size	Average volumes of residues (Pontius <i>et al.</i> , 1996)
Stability	Side-chain contribution to protein stability (kJ/mol) (Takano and Yutani, 2001)

Table 1. Physicochemical properties used in the creation of reduced amino acid alphabets.



**Fig. 1.** The scheme of n-gram extraction. A) Input data - peptides with a known amyloidogenicity status. B) Each peptide sequence was divided into overlapping hexamers. The amyloidogenicity status of a source sequence was used as the amyloidogenicity status of a derived hexamer. C) From each hexamer we extracted continuous 1-, 2- and 3-grams. We selected also gapped 2-grams with the length of gap equal from 1 to 3 residues and gapped 3-grams with a single gap between the first and the second or the second and the third element of the n-gram.

yields 3 hexamers. Assuming that in longer amyloids only a short part of the sequence is responsible for amyloidogenicity, our method might results in many false positives in the training data set. To evade this problem, we restricted the maximum length of peptides in training data set to fifteen amino acids to easy the extraction of probable hot-spots.

From each hexamer we extracted n-grams of the following length: 1, 2 and 3. In the case of 2- and 3-grams, we separately counted both gapped and continous n-grams. For 2-grams we counted n-grams with gaps of lengths from 1 to 3 and for 3-grams with a single gap between the first and the second or the second and the third element (see Fig. 1).

The inquire further the length of amyloidogenicity signal, we trained three classifiers for each encoding on the sequences of different length. We considered separately six-residue sequences, shorter of equal to ten residues and shorter or equal to fifteen residues.

All n-grams extracted from the hexamers in the training data set were filtered using QuiPT, our own implementation of permutation test with the information gain (mutual information) as the criterion of the importance of a specific n-gram. In the next step, we used n-grams with the p-value smaller than 0.05 to build a random forest (Breiman, 2001) classifier using ranger R package (Wright and Ziegler, 2015).

Furthermore, we repeated the procedure described above on two typical reduced alphabets of amino acids derived from the literature to check if the process of amyloidogenicity does require nonstandard groupings of amino acids. We also added the full amino acid alphabet to assess the advantages of the amino acid encoding.

### 3.4 Cross-validation and selection of the best performing encoding

The ability to correctly predict amyloidogenicity was assessed during the five-fold cross-validation. Since the data set was very heterogenous, we repeated the cross-validation fifteen times for each classifier to obtain more precise estimates of performance measures for each classifier.

To evaluate if our classifiers are able to use decision rules extracted from sequences of given length to correctly classify longer or shorter sequences, we calculate performance measures separately for four ranges of lengths of sequences:

- 6;
- 7-10;
- 11-15;
- 16-25.

The number of sequences from the given length range was roughly comparable between folds of cross-validation.

To choose the most adequate reduced amino acid alphabet, we ranked the values of AUC (with rank 1 for the best AUC, rank 2 for the second best AUC and so on) for each range of sequence length in the testing data set. The encoding having the lowest sum of ranks from all sequence length categories was selected as the best one and further used to develop AmyloGram, a predictor of amyloidogenicity.

### 3.5 Benchmark

We used pep424 data set Walsh *et al.* (2014) to compare the performance of AmyloGram and other predictors of amyloidogenicity. Since the model of AmyloGram does not covers peptides shorter than five amino acids, we removed them from the total benchmark data set. It should not affect the comparison as only five sequences were eliminated (around 1% of the original data set).

MCC to choose the best classifier. How sensitivity/specificity depends on the lengths of sequences in the positive and negative training set. Which the simplest (shortest) alphabet gives the best prediction. Correlation matrix of Hamming distance of best n-grams. Encoding distance.

## 4 Results

## 5 Discussion

## 6 Conclusion

## Acknowledgments

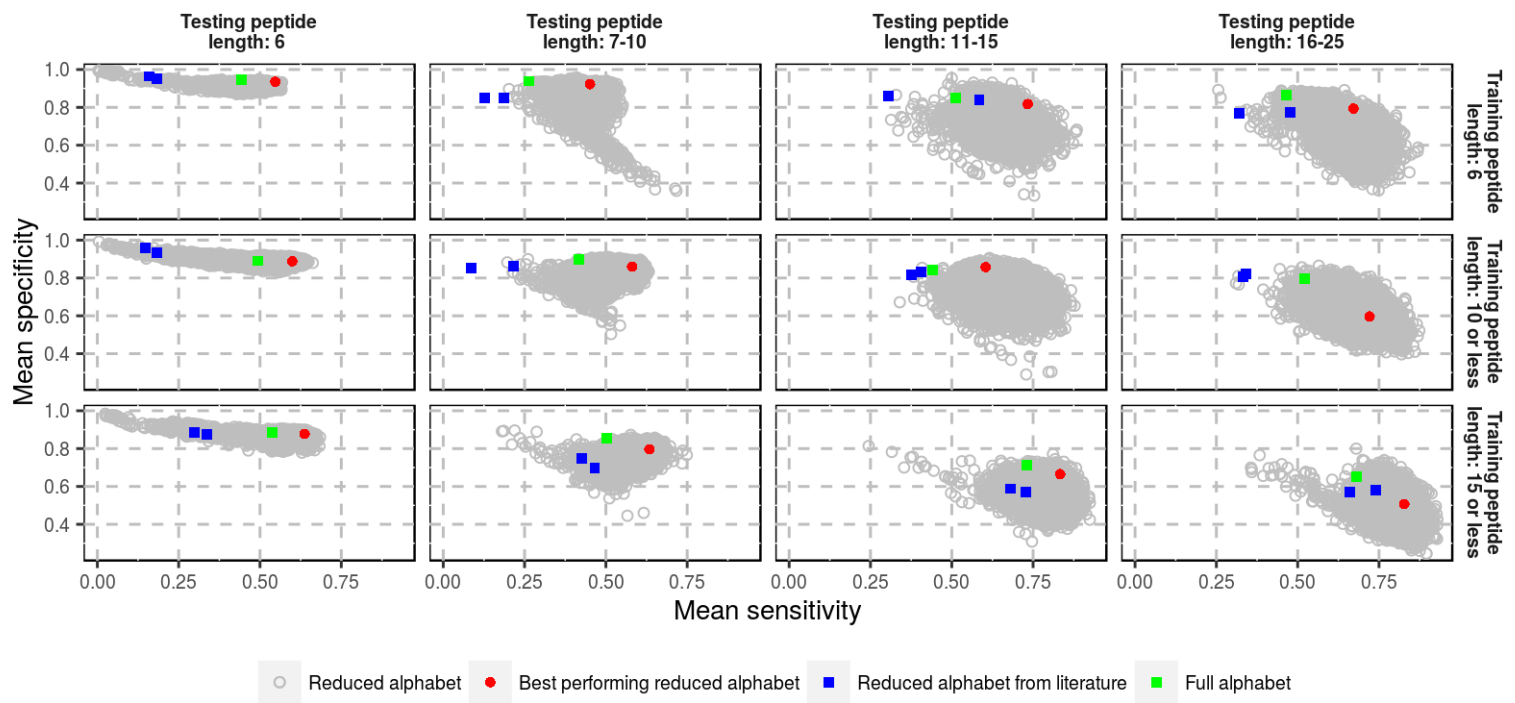
Text Text Text Text Text Text Text Text.

## Funding

This research was partially funded by the KNOW Consortium.

## References

- Argos, P., Rao, J. K., and Hargrave, P. A. (1982). Structural prediction of membrane-bound proteins. *European journal of biochemistry / FEBS*, **128**(2-3), 565–575.
- Bhaskaran, R. and Ponnuswamy, P. (1988). Positional flexibilities of amino acid residues in globular proteins. *International Journal of Peptide and Protein Research*, **32**(4), 241–255.
- Black, S. D. and Mould, D. R. (1991). Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Analytical Biochemistry*, **193**(1), 72–82.
- Breiman, L. (2001). Random Forests. *Machine Learning*, **45**(1), 5–32.
- Charton, M. and Charton, B. I. (1982). The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology*, **99**(4), 629–644.
- Jr, J. H. W. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, **58**(301), 236–244.
- Kanehisa, M. I. and Tsong, T. Y. (1980). Local hydrophobicity stabilizes secondary structures in proteins. *Biopolymers*, **19**(9), 1617–1628.
- Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). Aaindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, **36**(suppl 1), D202–D205.
- Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, **4**(1), 23–55.
- Kim, C. A. and Berg, J. M. (1993). Thermodynamic beta-sheet propensities measured using a zinc-finger host peptide. *Nature*, **362**(6417), 267–270.
- Nishikawa, K. and Ooi, T. (1986). Radial locations of amino acid residues in a globular protein: correlation with the sequence. *Journal of Biochemistry*, **100**(4), 1043–1047.
- Pontius, J., Richelle, J., and Wodak, S. J. (1996). Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *Journal of Molecular Biology*, **264**(1), 121–136.
- Prabhakaran, M. (1990). The distribution of physical, chemical and conformational properties in signal and nascent peptides. *The Biochemical Journal*, **269**(3), 691–696.
- Radzicka, A. and Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, **27**(5), 1664–1670.
- Sweet, R. M. and Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *Journal of Molecular Biology*, **171**(4), 479–488.
- Takano, K. and Yutani, K. (2001). A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins. *Protein Engineering*, **14**(8), 525–528.
- 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.
- Wozniak, P. P. and Kotulska, M. (2014). Characteristics of protein residue-residue contacts and their application in contact prediction. *Journal of Molecular Modeling*, **20**(11).
- Wozniak, P. P. and Kotulska, M. (2015). AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics (Oxford, England)*.
- Wright, M. N. and Ziegler, A. (2015). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv:1508.04409 [stat]*. arXiv: 1508.04409.
- Zhou, H. and Zhou, Y. (2004). Quantifying the effect of burial of amino acid residues on protein stability. *Proteins*, **54**(2), 315–322.



**Fig. 2.** Sensitivity and specificity of classifiers in cross-validation. Red circle: classifier employing best encoding of amino acid. Green square: classifier using full amino acid alphabet. Red square: classifiers employing encodings from literature. The classifier based on the best encoding always have good specificity and sensitivity.

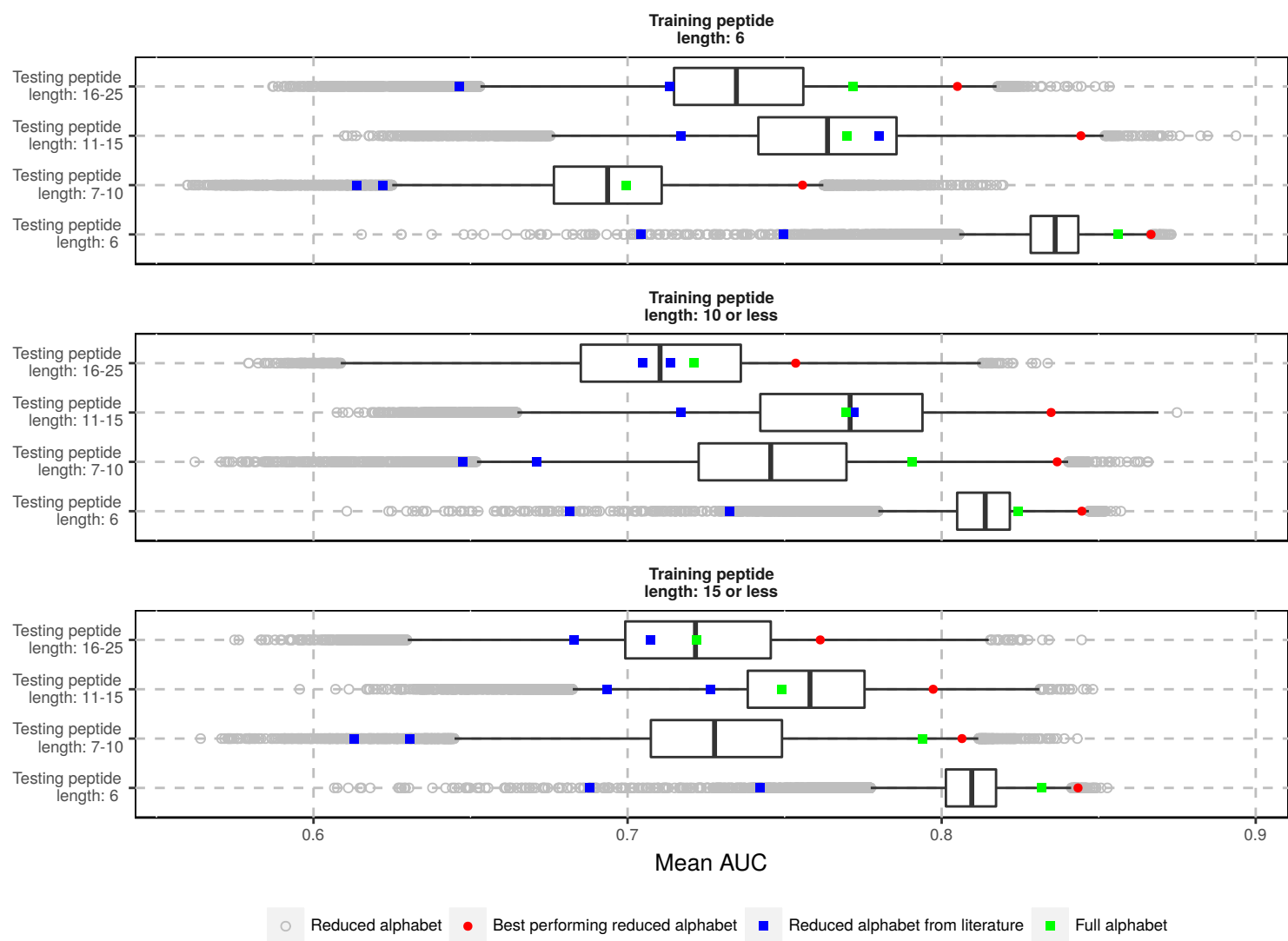


Fig. 3. Caption, caption.