Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Prediction of amyloidogenicity based on the n-gram analysis

Michał Burdukiewicz[1,*], Piotr Sobczyk[2], Paweł Mackiewicz[1] and Małgorzata Kotulska[3,*]

[1]University of Wrocław, Department of Genomics, [2]Wrocław University of Technology, Department of Mathematics and [3]Wrocław University of Technology, Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.
Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text vvv Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Contact: name@bio.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

2 Approach

3 Methods

3.1 Data set

The data used in the study was extracted from AmyLoad data base (Wozniak and Kotulska, 2015). Aside from eight sequences shorter than five residues that were removed from the final data set, we obtained 418 amyloidogenic sequences and 1039 non-amyloidogenic sequences (1457 peptides in total).

Sequences shorter than 6 amino acids and longer than 25 amino acids were removed from data set. The former were too short to be processed in the devised analysis framework and the latter were too diversified and rare, preventing the proper analysis.

The final data set contains 397 amyloidogenic and 1033 non-amyloidogenic sequences (1430 peptides in total).

3.2 Encodings of amino acids

The amyloidogenicity of given peptide may not depend on the exact sequence of amino acids, but on its more general properties. To verify this hypothesis, we created 524 284 reduced amino acid alphabets (encodings) with different lengths (from three to six letters) using Ward's clusterization Jr (1963) on the selected physicochemical properties from AAIndex database (Kawashima *et al.*, 2008). We handpicked several measures belonging to more general categories important in process of amyloidogenicity as size, hydrophobicity, solvent surface area, frequency in β-sheets and contactivity. As the rule of thumb, we limited ourselves to properties introduced after 1980 when, thanks to the technological advancements, the measurements were more accurate.

Some of the encodings created on different combinations of physicochemical properties were identical. In such case, only a single representative was included in the benchmark. After filtering duplicates, we obtained 18 535 unique encodings.

Since highly correlated or, contrarily, uncorrelated measures create very similar encodings, we further reduced the number of properties to 17 by selecting measures with the absolute value of Pearson's correlation coefficient for normalized values larger than 0.95.

3.3 Training of learners

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 A 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 A, separation 5 (Wozniak and Kotulska, 2014)
Contactivity	Values of Wc in proteins from class Beta, cutoff 12 A, separation 15 (Wozniak and Kotulska, 2014)
β-frequency	Average relative probability of inner beta-sheet (Kanehisa and Tsong, 1980)
β-frequency	Relative frequency in beta-sheet (Prabhakaran, 1990)
β-frequency	Thermodynamic beta sheet propensity (Kim and Berg, 1993)
Hydrophobicity	Hydrophobicity index (Argos *et al.*, 1982)
Hydrophobicity	Optimal matching hydrophobicity (Sweet and Eisenberg, 1983)
Hydrophobicity	Hydrophobicity-related index (Kidera *et al.*, 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 *et al.*, 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.

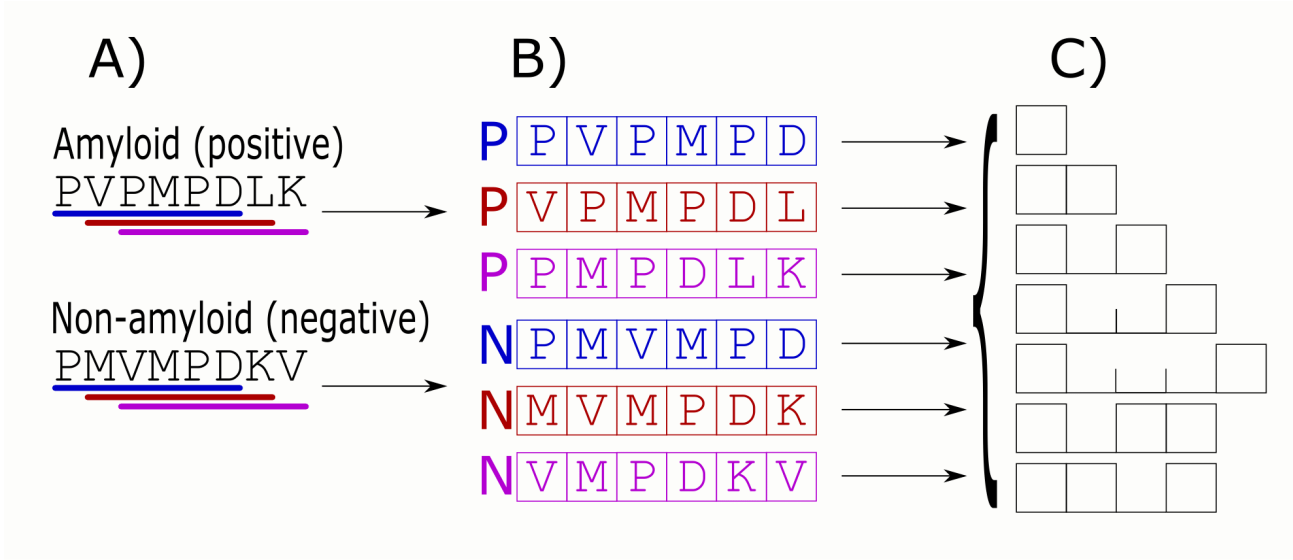

Fig. 1. The scheme of n-gram extraction. A) Input data - peptides with a known amyloidogenicity status. B) Each peptide sequence was divided into overlapping hexamers. The amyloidogenicity status of a source sequence was used as the amyloidogenicity status of a derived hexamer. C) From each hexamer we 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).

Furthemore, 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 peformance 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.

3.5 Benchmark

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

4.1 Selection of the best encoding

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 nd further used to develop AmyloGram, a predictor of amyloidogenicity.

4.2 Benchmark of classifiers

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).

The benchmark covered Amylogram as well as two best-performing peer-reviewed predictors of amyloidogenicity (PASTA2 (Walsh *et al.*, 2014) and FoldAmyloid (Garbuzynskiy *et al.*, 2010)). Additionally, we also benchmarked three predictors learned on the n-grams extracted from sequences of different length ranges without any amino acid encoding. We analyzed Area Under the Curve (AUC), Matthew's Correlation Coefficient (MCC), Sensitivity and Specificity (see Tab. 3).

Interesingly, the n-gram extraction method was efficient enough to produce classifiers able to outperformed published methods. Two of three n-gram based classifiers trained on the full alphabet have AUC higher than PASTA2 and all three were more successful than FoldAmyloid.

Although the proposed n-gram extraction creates efficient classifiers, the encoding of amino acids further increases the prediction of amyloidogenicity. AmyloGram has the highest AUC, MCC and Sensitivity among all tested classifiers. Is has lower specificity than two classifiers trained on the full alphabet, but still outperforms other published method in this category. It is important to highlight that AmyloGram is the most balanced of analyzed classifiers, having the best Specificity/Sensitivity balance, as indicated by the value of MCC.

Table 3. Results of benchmark on pep424 data set for AmyloGram, Pasta2, FoldAmyloid and random forest predictior learned on n-grams extracted without any amino acid encoding from the sequences of the length specified in the brackets.

Classifier	AUC	MCC	Sensitivity	Specificity
AmyloGram	**0.8972**	**0.6307**	**0.8658**	0.7889
PASTA2	0.8550	0.5227	0.7987	0.7444
FoldAmyloid	0.7351	0.4526	0.7517	0.7185
full alphabet (6)	0.8411	0.5427	0.4966	**0.9593**
full alphabet (6-10)	0.8581	0.5698	0.7517	0.8259
full alphabet (6-15)	0.8610	0.5490	0.8188	0.7519

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.

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.

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.

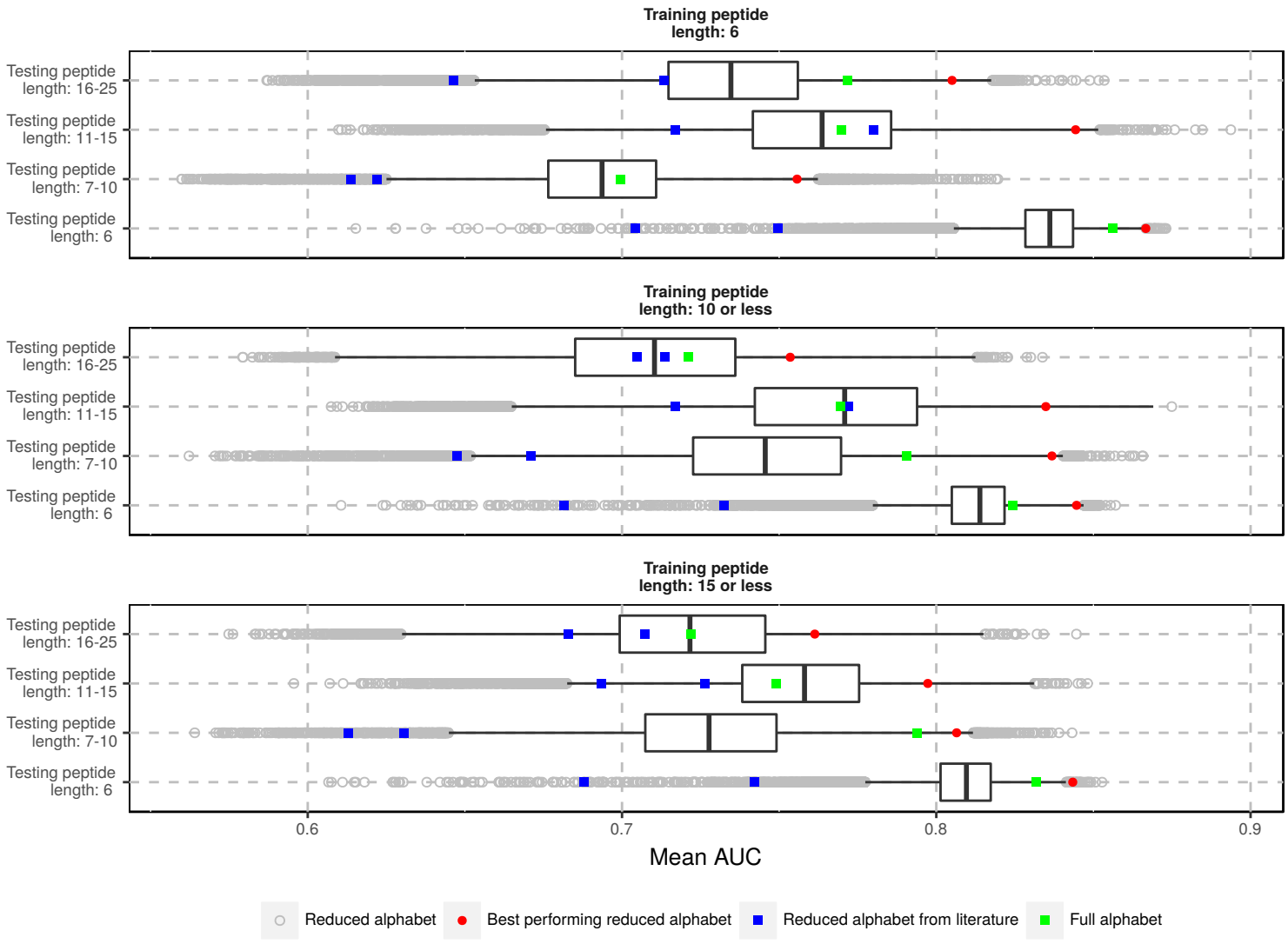

Fig. 2. Distribution of AUC values of different reduced amino acid alphabets. Red circle: classifier employing best encoding of amino acid. Green square: classifier using full amino acid alphabet. Blue square: classifiers employing encodings from literature.

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.

Table 2. Summarized results of 15 repeats of cross-validation. The results of all reduced alphabets (without the best performing encoding, the best encoding, the full alphabet and the reduced alphabet are presented separately.

Classifier	Length of peptides in testing set	Mean AUC	Mean MCC	Mean sensitivity	Mean specificity
All reduced alphabets	[5,6]	0.8176	0.4356	0.5261	0.8908
All reduced alphabets	(6,10]	0.7218	0.3332	0.4839	0.8234
All reduced alphabets	(10,15]	0.7611	0.3983	0.7045	0.6754
All reduced alphabets	(15,25]	0.7216	0.3141	0.7005	0.6015
Best reduced alphabet	[5,6]	**0.8516**	0.5111	0.5946	0.9004
Best reduced alphabet	(6,10]	0.7997	0.4484	0.5552	0.8597
Best reduced alphabet	(10,15]	0.8255	**0.5266**	0.7237	0.7801
Best reduced alphabet	(15,25]	0.7733	0.3884	**0.7402**	0.6326
Full alphabet	[5,6]	0.8375	0.4378	0.4910	0.9081
Full alphabet	(6,10]	0.7614	0.3508	0.3942	0.8971
Full alphabet	(10,15]	0.7628	0.3873	0.5607	0.8019
Full alphabet	(15,25]	0.7383	0.3438	0.5556	0.7706
Reduced alphabet from literature	[5,6]	0.7163	0.2063	0.2178	**0.9286**
Reduced alphabet from literature	(6,10]	0.6330	0.0550	0.2513	0.8097
Reduced alphabet from literature	(10,15]	0.7343	0.3038	0.5130	0.7518
Reduced alphabet from literature	(15,25]	0.6948	0.2167	0.4790	0.7205

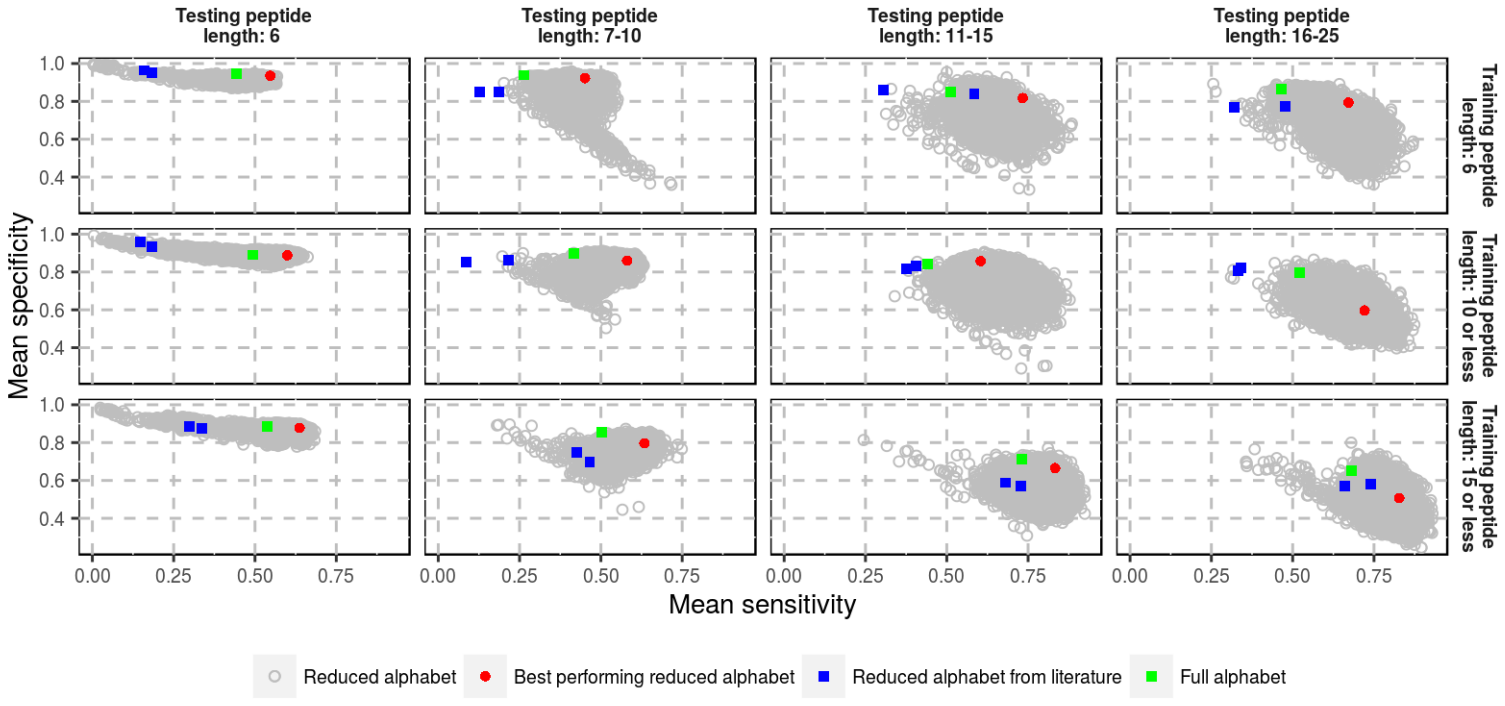

Fig. 3. Sensitivity and specificity of classifiers in cross-validation. Red circle: classifier employing best encoding of amino acid. Green square: classifier using full amino acid alphabet. Blue square: classifiers employing encodings from literature. The classifier based on the best encoding always have good specificity and sensitivity.