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

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

1 Introduction

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 1 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. (**?**) wants to know about …… text follows.

$$\sum x + y = Z \tag{1}$$

2 Approach

Equation (1) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text .....

3 Methods

3.1 Data set

The data used in the study was extracted from AmyLoad data base. 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 18 537 reduced amino acid alphabets with different lengths (from three to six letters).

We created the reduced alphabet of amino acids using Ward's clusterization on the selected physicochemical properties. We picked 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.

We further reduced the number of properties to 17, by selecting measures uncorrelated with others (with the Pearson's correlation coefficient for normalized values larger than 0.95 or smaller than 0.05) since they would create very similar encodings.

3.3 Traning of learners

In addition to the reduced amino acid alphabets created through the clusterization of physicochemical properties, we also examined two reduced alphabets from the literature and raw amino acid sequences.

Assuming that in longer amyloids only the short part of the sequence is responsible for amyloidogenicity, we restricted the maximum length of peptides in training data set to fifteen amino acids to easy the extraction of

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

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

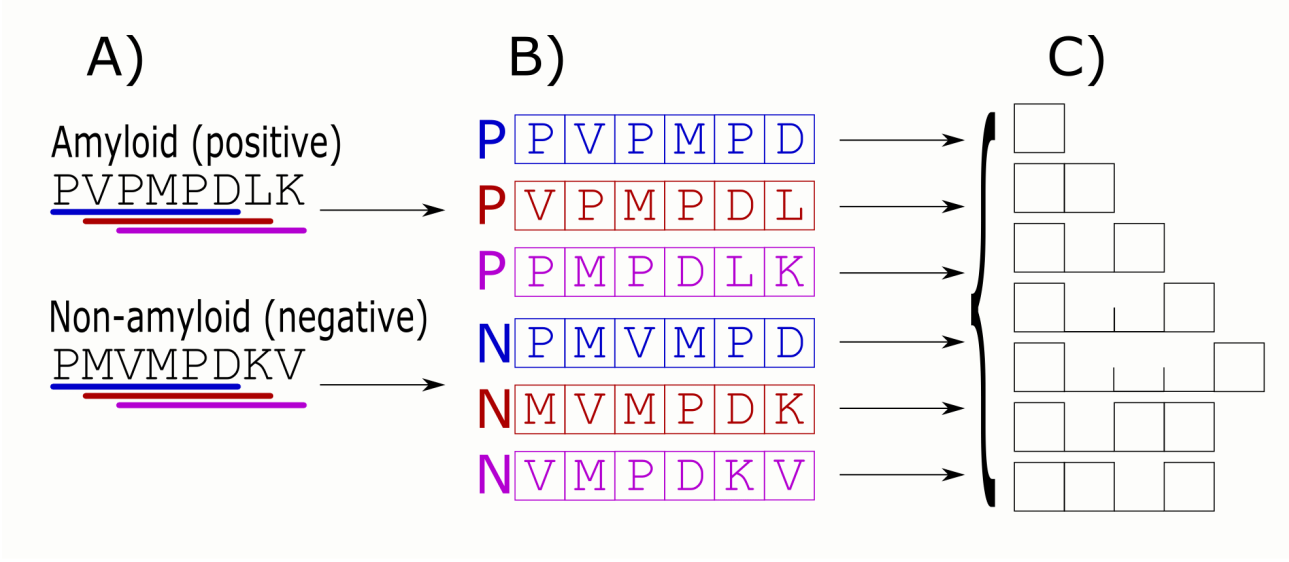

Fig. 1. Caption, caption.

probable hot-spots. 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 yields 3 hexamers.

The inquire the exact length of amyloidogenicity signal, we trained nine classifiers for each encoding on the sequences of different length. We considered sequences of length 6, shorter of equal to 10 residues and shorter or equal to 15 residues. We specified separately length of peptides for negative and positive training set, obtaining in total nine classifiers per each encoding.

3.4 Cross-validation

The cross-validation was repeated five times for each combination of the encoding as well as the length of sequences in positive data set and negative data set.

Since we are interested if our classiffiers 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.

4 Discussion

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text

Text Text Text Text Text Text Text Text Text Text. **?** might want to know about 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. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text.

Table **??** shows that Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text.

5 Conclusion

(Table **??**) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about 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. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about 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. Figure 2 shows that the above method 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. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about 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. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about 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. Figure 2 shows that the above method Text Text Text Text

Acknowledgements

Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Funding

The conference fee was funded by the KNOW Consortium.