Prediction of peptide and protein propensity for amyloid formation
Carlos Familia1,3, Sarah R. Dennison1, Alexandre Quintas3 and David A. Phoenix2

1School of Pharmacy and Biological Sciences, University of Central Lancashire, Preston PR1 2HE, UK.
2Office of the Vice Chancellor, London South Bank University, 103 Borough Road, London SE1 0AA, UK.
3Instituto Superior de Ciências da Saúde Egas Moniz, Campus Universitário, Qta. Da Granja, Monte de Caparica 2829-511 Caparica, Portugal.

Amyloid fibers are unbranched filamentous protein aggregates with an indefinite length and a diameter that can range from 5-8 nm. They are commonly formed by polypeptide chains arranged in a characteristic cross-β conformation with strands perpendicular oriented to the fiber long axis. This results in a series of stacked β-chains (Figure 1) that propagate along the fiber, where the polypeptides are found arranged in a highly ordered fashion.

Despite fibril structural similarity, proteins that can undergo structural changes that lead to amyloid fibril formation have been characterized (Figure 2). Furthermore, a number of researchers have suggested that the ability to form amyloid fibrils is an intrinsic property of the polypeptide backbone.

The development of a new amyloidogenic propensity predictor was based on a machine learning approach through recurrent feature selection and feed-forward neural networks, after sequence encoding with amino acid physicochemical and biochemical properties (Figure 3).

Sequence datasets
Two distinct sequence datasets (training sequences and external validation sequences) were constructed from the literature, containing sequences of peptides and proteins with experimental in vitro evidence of amyloid formation. The training sequences dataset is exclusively formed by six amino acids peptides in length, with a total of 298 sequences, from which 161 have been reported negatively and 125 have been reported positively for amyloid formation. The external validation sequences dataset is a more general dataset comprising a total of 483 peptide and protein sequences with lengths greater than six amino acids, from which 142 have been reported negatively and 342 have been reported positively for amyloid formation.

Sequence encoding
Sequence information was encoded into numerical vectors through the use of two datasets of amino acid physicochemical and biochemical properties, the Amмо Acid Index Database version 9.1 (Akindele) and the Ammino Acid Physicochemical Properties Database (APDPbase) based on the single characteristics, their cumulative summation and some basic measures of these characteristics (summation, mean, harmonic mean, median, mode, standard deviation, interquartile range, mean absolute deviation, range, kurtosis and skewness).

Feature selection
Feature selection was performed with two recurrent feature selection wrapper methods, from the caret package v5.14-8 and boruta package v 2.1.21 with the performance of improving protein encoding.

Artificial neural networks
Feed-forward fully connected artificial neural networks were created with MATLAB’s Neural Networks Toolbox, and trained after random division of the training sequences dataset into three distinct subsets, the training (70%), test (15%) and validation (15%) subsets. The best neural network was selected from a total of 1000 trained networks based on the values of accuracy and standard deviation obtained for the training, test, validation subsets and overall dataset. The selected neural network was post-processed by the classification of the sequences present in the external validation dataset, which was performed by the submission of the pre-processed individual input vectors, generated by a sliding window of six amino acids that was run through the polypeptide sequence, to the corresponding neural network (Figure 4). A sequence was considered amyloidogenic if at least one of the six amino acids windows that went through the sequence was classified as an amyloidogenic protein.

Input sequence (APDBase) encoding

In this study we have developed a highly accurate and effective method for the prediction of amyloid propensity based on the polypeptide amino acidic sequence alone (Figure 5). This has been achieved using a very small subset of highly relevant physicochemical and biochemical amino acid properties. Overall, this study not only provides a new amyloidogenic propensity prediction method, but also new insights into the understanding of the key driving forces underpinning the self-assembly of peptides and proteins into amyloid-like fibers.

The first author would like to thank Víctor Familia, Branca Proença and Ana Santos for all their encouragement and support. The authors thank John Edwards and Robert Leigh for their support given to make this new predictor available online. The authors also thank Professor Silvio Tissot and Professor Sebastian Maurer-Stroh for the help regarding the preparation of the results provided by Paula and Waltz prediction methods, respectively.

Amyloidogenic propensity profile

Figure 1 – Schematic image of the amyloid structure showing the stacked β-chains that propagate along the fiber rendering the fibril unbranched.

Figure 2 – Schematic image of the native structure of some of the proteins that are known to form amyloid fibers, highlighting the sequence motifs of amyloidogenic and non-amyloidogenic proteins.

Figure 3 – Overview of the methods used for the development of the new amyloidogenic propensity predictor based on a machine learning approach. Through recurrent feature selection and feed-forward neural networks.

Figure 4 – Sensitivity, specificity, positive predictive value, negative predictive value and accuracy obtained in the classification of the training sequences (top) and external validation sequences (bottom) datasets for all predictors, with corresponding 95% confidence intervals and accuracy comparison p-values.

Figure 5 – Image of the developed neural network showing the classification of sequences with lengths higher than six amino acids. The output is obtained through a sliding window of six amino acids that was run through the polypeptide sequence, to the corresponding neural network. A sequence was considered amyloidogenic if at least one of the six amino acids windows that went through the sequence was classified as an amyloidogenic protein.

Figure 6 – Overview of the developed neural network showing the classification of sequences with lengths higher than six amino acids. The output is obtained through a sliding window of six amino acids that was run through the polypeptide sequence, to the corresponding neural network. A sequence was considered amyloidogenic if at least one of the six amino acids windows that went through the sequence was classified as an amyloidogenic protein.

Figure 7 – Schematic diagram of the amyloidogenic propensity prediction software and the functional protein sequence database.

Figure 8 – Schematic diagram of the amyloidogenic propensity prediction software and the functional protein sequence database.

Figure 9 – Schematic diagram of the amyloidogenic propensity prediction software and the functional protein sequence database.

Figure 10 – Schematic diagram of the amyloidogenic propensity prediction software and the functional protein sequence database.