Prediction of peptide and protein propensity for amyloid formation

Carlos Família[1,3], Sarah R. Dennison[1], Alexandre Quintas[3] and David A. Phoenix[2]

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

INTRODUCTION

Amyloid fibers are unbranched filamentous protein aggregates with an indefinite length and a diameter that can range from 6-12 nm[1]. They are commonly formed by polypeptide chains arranged in a characteristic cross-β conformation with strands perpendicularly oriented to the fiber long axis. This, results in a series of stacked β-chains (Figure 1) that propagate along the fiber[2], where the polypeptides are found arranged in a highly ordered fashion[3].

Despite fibril structural similarity, proteins that can undergo structural changes that ultimately lead to amyloid formation are quite diverse, sharing no obvious sequence or structural homology[4] (Figure 2). Furthermore, a number of researchers have suggested that the ability to form amyloid fibrils is an intrinsic property of the polypeptide backbone[5,6].

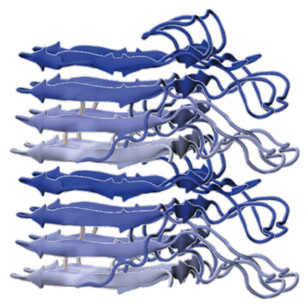

Figure 1 – Schematic image of the amyloid structure showing the stacked β-chains that propagate along the fiber axis constituting the fiber protofilament.

Amyloid fiber formation has long been associated with several debilitating diseases and currently there are around forty reported human diseases linked to amyloid. These include localized amyloidosis such as pancreatic amyloidosis, atrial amyloidosis of the heart, Alzheimer's disease, Parkinson's disease, Huntington's disease and Creutzfeldt-Jakob's disease, as well as systemic diseases such as familial amyloid polyneuropathy or immunoglobulin light-chain amyloidosis[7].

Over the last two decades several publications have shown that amyloid could be produced through "controlled" fibrillization and with specific biological functions instead of an off-pathway product of protein folding that leads to disease[8]. Examples include the bacterial pili, the curly fibrils expressed in *Escherichia coli* and *Salmonella,* which are involved in surface colonization and biofilm formation, human pigment binding templates, regulation of the expression reading-through stop-codon in yeast (*Saccharomyces cerevisiae*, Sup35p), among others.

Protein aggregation and subsequent assembly into amyloid like structures is commonly seen as a major problem in large scale expression of peptides and proteins of potential interest within the field of biotechnology, which frequently result in their accumulation as insoluble aggregates within inclusion bodies, reducing the yield of extraction and purification, and ultimately, the economic viability of the purification process[9]. Recent works, however, exploited protein aggregation into amyloid fibrils and subsequent accumulation in inclusion bodies with the purpose of improving protein expression[10].

In the area of material science, amyloid fibrils can be seen as an important source of innovation, since they may provide insights into a wide range of properties that could be explored in the design of new nanomaterials. The ability of amyloid to self-assemble or self-replicate into well-defined structures, their nanoscale dimensions, the diversity of associated protein sequences, the ease of production and low cost make them key systems for investigation[11].

Due to the relevance of amyloid in such different areas of study as biochemistry, medicine, microbiology, biotechnology and materials science, the knowledge of which and how peptides and proteins undergo amyloid formation is of paramount importance. Experimental identification of amyloidogenic proteins *in vitro* is extremely laborious and time-consuming. Hence, computational approaches that can accurately and reliably predict the amyloidogenic propensity of peptides and assess their amyloidogenic potential based on the sequence information alone are extremely valuable. Additionally, such work can help to elucidate the driving forces responsible for amyloid-like fiber formation and stabilization and provide new insights into the self-assembly problem.

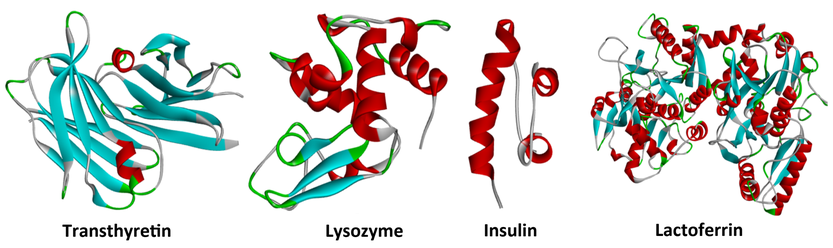

Figure 2 – Schematic image of the native structure of some of proteins that are known to form amyloid fibers, evidencing their secondary structure.

METHODS

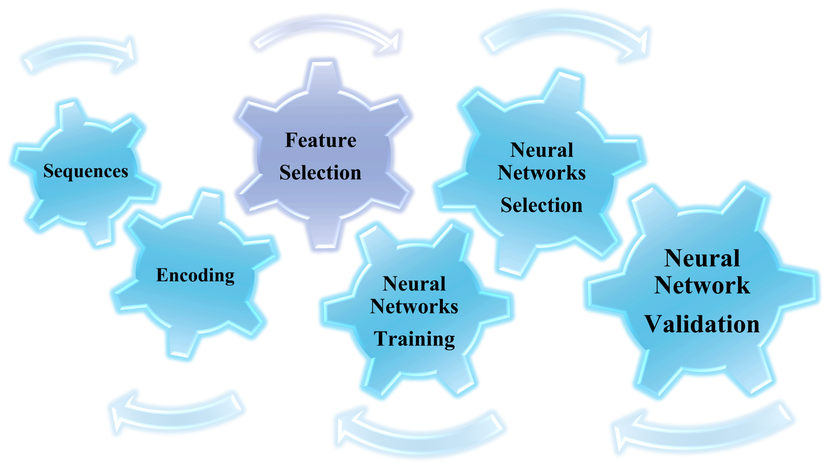

Figure 3 – Overview of the methodology used for the development of the new phenomenological predictor based on a machine learning approach, through recursive feature selection and feed-forward neural networks.

The development of a new amyloidogenicity propensity predictor was based on a machine learning approach through recursive 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 296 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 341 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 Amino Acid Index Database version 9.1 (AAindex)[12] and the Amino Acid Physicochemical Properties Database (APDBase)[13] 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 recursive feature selection wrapper methods, from the caret package v.5.15-48 and boruta package for R v.2.15.1 with five different internal classifiers (sparse partial least squares (spls), shrinkage discriminant analysis (sda), both linear, penalized support vector machines (psvm) and naïve bayes (nb) for caret, and random forests (rf) for boruta).

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 networks were posteriorly validated 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 acid windows that went through the sequence was classified amyloidogenic.

Input sequence (Aβ42): DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA

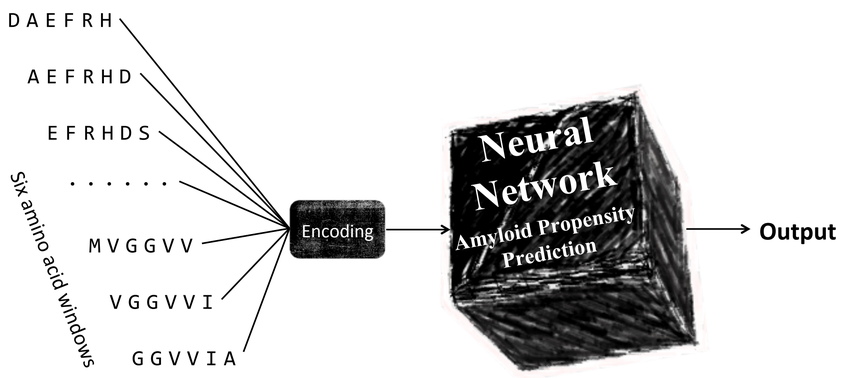

Figure 4 – Image of the developed neural network showing the classification of sequences with lengths higher than six amino acids through a sliding window of six amino acids for the example input sequence of Aβ42 protein.

RESULTS

Recursive feature selection allowed the identification of 5 different features subsets per encoding dataset, from each 1000 neural networks were trained and 10 were selected based on the accuracy value obtained for the classification of the external validation sequences dataset. The overall best neural network was trained with the subset of features discovered by recursive feature selection that used Naïve Bayes as internal classifier.

The features consisted in a subset of 14 features was selected, from which, three corresponded to summation of the amino acid properties (Normalized frequency of beta-sheet, Normalized frequency of beta-sheet from LG and Weights for beta-sheet at the window position of 1), one corresponded to the values of standard deviation and range (Isoelectric Point), and three corresponded to the standard deviation, range and mean absolute deviation (Atom-based hydrophobic moment, Helix termination parameter at position j+1 and ΔG° values for the peptides extrapolated to 0M urea.

Additionally the selected neural networks was compared with other published methods, where 95% confidence intervals were computed through bootstrapping with 2000 replicates, and p-values were computed for the comparison of the accuracy values between the selected predictor and each given predictor using the post-hoc methods for the Friedman's test after a 10-fold stratified resampling of the data (Figure 5). Showing to be significantly more accurate than Aggrescan, FoldAmyloid and Tango for the training sequences dataset and then Pasta, Tango, Waltz and Zyggregator for the external validation sequences dataset, both at a significance level of 0.05.

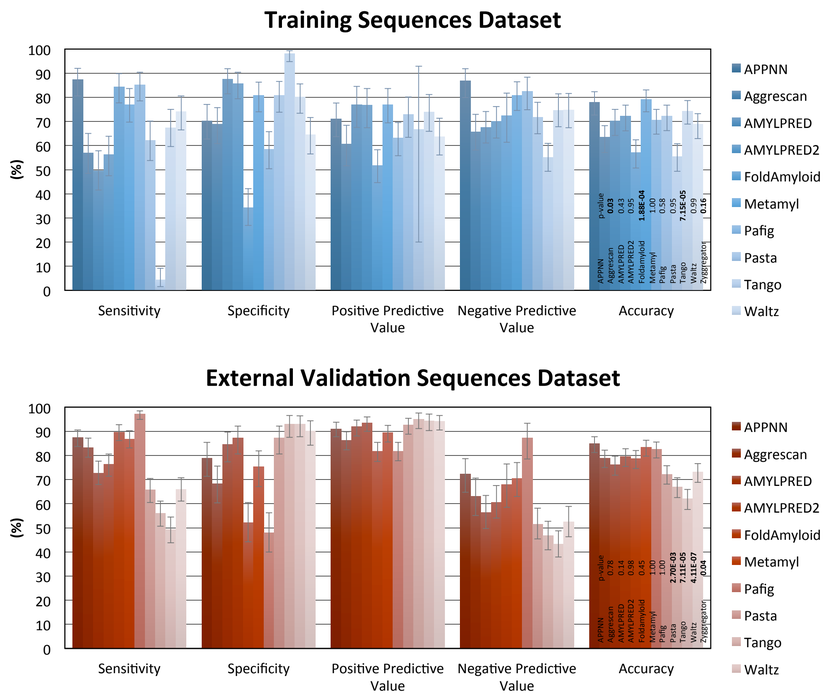

Figure 5 – Sensitivity, specificity positive predictive value, negative predictive value and accuracy obtained in the classification of the training sequences (top) and external validation sequences (bottom) datasets for all predictors, with corresponding 95% confidence intervals and accuracy comparison p-values.

CONCLUSION

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 6). 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 amyloidogenicity 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.

ACKNOWLEDGEMENTS

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

SAMPLE OUTPUT

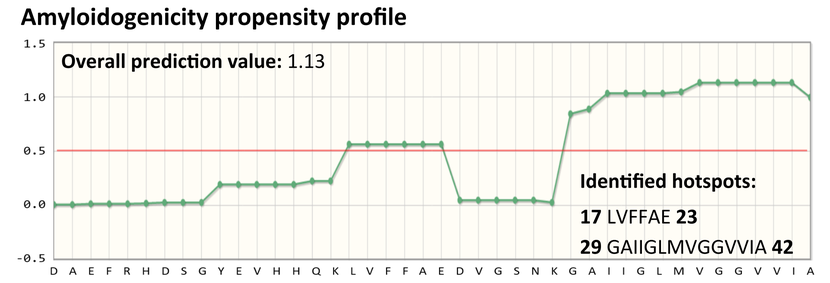

Figure 6 – Sample output from the web interface of the developed neural network using the Aβ42 protein sequence as example.

REFERENCES

1. Sunde, M. & Blake, C. The structure of amyloid fibrils by electron microscopy and X-ray diffraction. Adv. Protein Chem. 50, 123–159 (1997).
2. Kelly, J. W. Alternative conformations of amyloidogenic proteins govern their behavior. Curr. Opin. Struct. Biol. 6, 11–17 (1996).
3. Astbury, W. T., Dickinson, S. & Bailey, K. The X-ray interpretation of denaturation and the structure of the seed globulins. Biochem. J. 29, 2351–2360 (1935).
4. Hamodrakas, S. J. Protein aggregation and amyloid fibril formation prediction software from primary sequence: towards controlling the formation of bacterial inclusion bodies. FEBS J. 278, 2428–2435 (2011).
5. Dobson, C. M. The structural basis of protein folding and its links with human disease. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 356, 133–145 (2001).
6. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 424, 805–808 (2003).
7. Selkoe, D. J. Folding proteins in fatal ways. Nature 426, 900–904 (2003).
8. Fowler, D. M., Koulov, A. V, Balch, W. E. & Kelly, J. W. Functional amyloid - from bacteria to humans. Trends Biochem. Sci. 32, 217–224 (2007).
9. Ventura, S. & Villaverde, A. Protein quality in bacterial inclusion bodies. Trends Biotechnol. 24, 179–185 (2006).
10. Idicula-thomas, S. & Balaji, P. V. Protein aggregation : A perspective from amyloid and inclusion-body formation. Curr. Sci. 92, 758–767 (2007).
11. Pastor, M. T., Esteras-Chopo, A. & López de la Paz, M. Design of model systems for amyloid formation: lessons for prediction and inhibition. Curr. Opin. Struct. Biol. 15, 57–63 (2005).
12. Kawashima, S., Ogata, H. & Kanehisa, M. AAindex: Amino Acid Index Database. Nucleic Acids Res. 27, 368–369 (1999).
13. Mathura, V. S. & Kolippakkam, D. APDbase: Amino acid Physico-chemical properties Database. Bioinformation 1, 2–4 (2005).

WEBPAGE

http://www.uclan.ac.uk/research/environment/projects/analysis_prediction_of_peptides_proteins_propensity_for_amyloidogenicity.php

CONTACT

carlosfamilia@gmail.com