Bioinformatics, 31(20), 2015, 3395–3397
doi: 10.1093/bioinformatics/btv375
Advance Access Publication Date: 17 June 2015
Applications Note

Databases and ontologies

AmyLoad: website dedicated to amyloidogenic protein fragments

Pawel P. Wozniak and Malgorzata Kotulska*

Department of Biomedical Engineering, Wroclaw University of Technology, Wroclaw, Poland

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 29, 2015; revised on May 25, 2015; accepted on June 12, 2015

Abstract

Analyses of amyloidogenic sequence fragments are essential in studies of neurodegenerative diseases. However, there is no one internet dataset that collects all the sequences that have been investigated for their amyloidogenicity. Therefore, we have created the AmyLoad website which collects the amyloidogenic sequences from all major sources. The website allows for filtration of the fragments and provides detailed information about each of them. Registered users can both personalize their work with the website and submit their own sequences into the database. To maintain database reliability, submitted sequences are reviewed before making them available to the public. Finally, we re-implemented several amyloidogenic sequence predictors, thus the AmyLoad website can be used as a sequence analysis tool. We encourage researchers working on amyloid proteins to contribute to our service.

Availability and implementation: The AmyLoad website is freely available at http://comprec-lin.iiar.pwr.edu.pl/amyload/.

Contact: malgorzata.kotulska@pwr.edu.pl

1 Introduction

The mechanisms for development of amyloidogenic diseases such as the Alzheimer's or Parkinson's diseases are still not completely known. It has been proved that for these diseases to occur, specific protein fragments must become exposed in the structures. These fragments are characterised by a high aggregation propensity, which leads to the formation of fibres with a sterric zipper form, soluble oligomers or amorphous aggregates (Uversky and Fink, 2004). The amyloidogenic sequence fragments are essential to understand the development of amyloid diseases.

Several difficulties are encountered in studies of amyloidogenic fragments. First, there are a few internet data sources which contain such fragments. The most popular are AMYPdb (Pawlicki *et al.*, 2008), WALTZ-DB (Beerten *et al.*, 2015), AmylHex with AmylFrag (Thompson *et al.*, 2006) and ZipperDB (Goldschmidt *et al.*, 2010), or datasets used to validate analytical methods, such as TANGO (Fernandez-Escamilla *et al.*, 2004) or AGGRESCAN (Conchillo-Solé *et al.*, 2007). None of these sources contains all available data. The WALTZ-DB, for instance, is designed for hexapeptides, while ZipperDB collects amyloidogenic sequences obtained with computer modelling. AMYPdb, on the other hand, includes amino acid templates characteristic of specific amyloid protein families, rather than amyloidogenic fragments as such. Secondly, most of the datasets do not provide a user-friendly way of data browsing and filtering. To the best of our knowledge, currently only WALTZ-DB enables its users to filter data. Finally, various computational methods have been developed to predict the occurrence of amyloidogenic fragments in protein sequence. Most of these are based on empirical amino acid features, e.g. contact sites density or hydrogen bonding (Conchillo-Solé *et al.*, 2007; Garbuzynskiy *et al.*, 2010), machine learning methods (Gasior and Kotulska, 2014) or consensus methods (Tsolis *et al.*, 2013), to mention only a few. As with the fragment data, these tools are not collected together in one place on the Internet. Therefore, if a sequence is analyzed with several different predictors, each method must be used separately. To address these difficulties, we present the AmyLoad internet database which consolidates a great majority of the currently available amyloidogenic and non-amyloidogenic sequences, and several prediction methods. The AmyLoad website allows a user to fully personalize their work with over 1300 unique sequence database entries and implemented

prediction methods, and also to submit new data to the database. Links to other yet not implemented amyloid predictors are also available in our service.

2 Methods

The AmyLoad website currently contains almost 1400 unique, experimentally derived sequence fragments mostly selected from five datasets: WALTZ-DB, AmylHex, AmylFrag and validation datasets of AGGRESCAN and TANGO, with 908, 158, 43, 25 and 248 sequences, respectively. The more detailed information provided for each sequence, as well as a set of other sequences, was obtained by manual studies of over 90 publications for which references are accessible on the AmyLoad website. The prediction methods FoldAmyloid (Garbuzynskiy *et al.*, 2010), AGGRESCAN (Conchillo-Solé *et al.*, 2007) and FISH (Gasior and Kotulska, 2014) were re-implemented in Python 2.7. Their performance was tested comparing with the results of their original online implementations. All the website data is stored in a MySQL database. The web server is built with Django 1.7 framework and functionality implemented in Python 2.7.

3 Results

3.1 Browsing the database

Sequences presented on the AmyLoad website are listed in a table with columns indicating protein name, fragment name, amino acid sequence and the information about amyloidogenicity. Additional details are separately presented for each fragment, such as an examination method, source dataset references (see Section 2), references to relevant publications and other information. Furthermore, if the sequence fragment belongs to the WALTZ-DB, there is an additional link provided leading to the WALTZ-DB sequence information page (Beerten *et al.*, 2015). Data presented in the table can be filtered by selected features, such as the protein name, amyloidogenicity, minimum and maximum sequence length or the amino acid subsequence occurrence. Each fragment can be downloaded in the CSV, SSV (semicolon-separated-value), XML or FASTA file format.

3.2 Sequence analysis

One of the most distinguishing functionalities of AmyLoad is additional implementation of sequence analysis methods. The analysis is performed on the submitted FASTA sequences. There are three amyloidogenic sequence predictors available on the website, FoldAmyloid, AGGRESCAN and FISH. Each method can be applied using the same options as in their original online implementations. For example, FoldAmyloid can be run for different empirical amino acid features, sizes of the averaging frame and the threshold values (Garbuzynskiy *et al.*, 2010). For implementation we selected our own method and two other classical and well performing methods which were exhaustively described and relatively easy to re-implement. Other methods will be implemented gradually. Until then, we list all available tools on the *Analysis* webpage with their references and online directories. Another option is a search of the AmyLoad database to determine if any fragment of the analysed sequence is already known to AmyLoad. To enhance the service performance, the results are sent by the web server to the provided e-mail address, so the user does not need to wait online for completion of selected actions.

3.3 Personalized studies

A registered user obtains the access to additional functionalities of the AmyLoad website. First, it is possible to create *temporary sessions*, which makes the selection of specific fragment subsets easier. A *temporary session* can be saved and loaded later by the user, from the history of saved sessions, any time it is needed. Also, the AmyLoad website allows the users to contribute information into it. Registered users may submit their own sequence fragments to the database in two different ways. The first method is by submission form where one sequence can be added at a time. The second method is to upload a properly built XML file which can contain information about more than one fragment. The structure of such an XML file is described in the AmyLoad help page along with an example of a properly-constructed XML file. Users are apprised of the status of their submission by e-mail message, and errors are flagged for correction. Properly submitted sequence fragment data is immediately available to the user who added it in the *Added fragments* page. The data is available to the general public, in the database browsing page, only after review. The review is a compatibility assessment of the provided references and other sequence details. If the references are not provided, the database curator may ask for additional information necessary for the sequence evaluation. The database will accept only experimental data that have been published in a reliable peer-reviewed scientific journal.

4 Summary

The AmyLoad website contains almost 1400 unique amyloidogenic and non-amyloidogenic sequence database entries with detailed information and references. Since it gathers data from most of the well-known amyloidogenic datasets, we believe that it is currently the largest internet data source dedicated to amyloidogenic fragments. Compared to ZipperDB, AmyLoad contains only experimental data. Also, AmyLoad collects amyloidogenic fragments of various lengths, which is different from WALTZ-DB designed for hexapeptides. Further, AmyLoad provides explicit sequences, differing from AMYPdb with templates of amyloid families. Finally, our website allows submission of new sequences into the database and personalization of a user's work. We believe that, upon time, it will grow and become a popular place to deposit and obtain data about newly discovered sequences. By taking advantage of the implemented sequence analysis methods, the AmyLoad website has the potential of becoming a powerful amyloidogenicity predicting tool. Results from different methods can be easily compared, indicating differences and supporting their evaluation. Also, a growing number of deposited data will influence a development of new, improved analytical methods. Therefore, we strongly encourage all authors of such methods and all scientists analysing amyloidogenic sequence fragments to contribute to this service.

Acknowledgements

We would like to thank Prof. Jacek Cichoń for discussion and software-related suggestions, and Dr. Wayne Fisher for critical review of the manuscript.

Funding

This work was in part supported by the grant N N519 643540 from National Science Centre of Poland.

References

Beerten,J. *et al.* (2015) WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. *Bioinformatics*, **31**, 1698–1700.

Conchillo-Solé,O. *et al.* (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65.

Fernandez-Escamilla,A.-M. *et al*. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.

Garbuzynskiy,S.O. *et al*. (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, **26**, 326–332.

Gasior,P. and Kotulska,M. (2014) FISH Amyloid—a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids. *BMC Bioinformatics*, **15**, 54.

Goldschmidt,L. *et al*. (2010) Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proc. Natl. Acad. Sci. USA*, **107**, 3487–3492.

Pawlicki,S. *et al*. (2008) AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics*, **9**, 273.

Thompson,M.J. *et al*. (2006) The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl. Acad. Sci. USA*, **103**, 4074–4078.

Tsolis,A.C. *et al*. (2013) A Consensus method for the prediction of 'Aggregation-Prone' peptides in globular proteins. *PLoS One*, **8**, e54175.

Uversky,V.N. and Fink,A.L. (2004) Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim. Biophys. Acta*, **1698**, 131–153.