Bioinformatics, 31(10), 2015, 1698–1700
doi: 10.1093/bioinformatics/btv027
Advance Access Publication Date: 18 January 2015
Applications Note

OXFORD

Databases and ontologies

WALTZ-DB: a benchmark database of amyloidogenic hexapeptides

Jacinte Beerten[1,2], Joost Van Durme[1,2], Rodrigo Gallardo[1,2], Emidio Capriotti[1,2,3], Louise Serpell[4], Frederic Rousseau[1,2,*] and Joost Schymkowitz[1,2,*]

[1]VIB Switch Laboratory, Leuven, Belgium, [2]Department of Cellular and Molecular Medicine, KU Leuven, Herestraat 49 Box 802, Leuven, Belgium, [3]Division of Informatics, Department of Pathology, University of Alabama at Birmingham, Birmingham, AL 35249, USA and [4]School of Life Sciences, University of Sussex, Falmer, East Sussex BN1 9QG, UK

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on September 22, 2014; revised on January 8, 2015; accepted on January 12, 2015

Abstract

Summary: Accurate prediction of amyloid-forming amino acid sequences remains an important challenge. We here present an online database that provides open access to the largest set of experimentally characterized amyloid forming hexapeptides. To this end, we expanded our previous set of 280 hexapeptides used to develop the Waltz algorithm with 89 peptides from literature review and by systematic experimental characterisation of the aggregation of 720 hexapeptides by transmission electron microscopy, dye binding and Fourier transform infrared spectroscopy. This brings the total number of experimentally characterized hexapeptides in the WALTZ-DB database to 1089, of which 244 are annotated as positive for amyloid formation.

Availability and implementation: The WALTZ-DB database is freely available without any registration requirement at http://waltzdb.switchlab.org.

Contact: frederic.rousseau@switch.vib-kuleuven.be or joost.schymkowitz@switch.vib-kuleuven.be

1 Introduction

Amyloid formation by proteins is widely recognized as a pathogenic mechanism in diverse diseases such as Alzheimer Disease and type II diabetes, but also as a functional mechanism of biological nanostructure formation, such as the chorion protein that stabilizes insect eggshell (Fowler *et al*., 2006). It has been generally established that the formation of amyloid fibrils by proteins is nucleated by short aggregation prone segments (APR) of the polypeptide chain, which are a necessary and sufficient requirement to allow amyloid conversion of a folded protein (De Baets *et al*., 2014). In amyloid, these short stretches form an intermolecular beta-sheet that runs parallel to the fiber axis, indicated as the cross-beta structure (Sunde *et al*., 1997; Eisenberg and Jucker, 2012). The sequence that constitute APRs are usually characterized by a high beta-sheet propensity and hydrophobicity and a low net charge (Chiti *et al*., 2003). Computational tools have been developed to predict amylogenic or aggregation propensities of proteins by detecting APRs in polypeptide sequences, reviewed elsewhere (Ho *et al*., 2006; De Baets *et al*., 2014). We earlier developed the Waltz amyloid prediction algorithm (Maurer-Stroh *et al*., 2010), which is a data-based statistical method that uses a position-specific scoring matrix for its data representation. As the quality of statistical methods depends critically on the quantity (number and sequence diversity of known positive and negative examples) and quality (confidence of the amyloid status) of the available data, we decided to increase the available high confidence learning data by an order of magnitude. This was achieved by expanding the set of 280 hexapeptides of known amyloid forming proteins with an additional experimentally verified 720 hexapeptides derived from 63 different proteins, combined with an additional 89 peptides derived from literature review, bringing the total number of hexapeptides to 1089.

2 Methods

In order to study the amylogenic properties of hexapeptides, 720 uncharacterized hexapeptides were synthesized by JPT Technologies GmbH. Peptides were dissolved in 50 mM phosphate buffer pH 7.4, 0.05% sodium-azide to a final concentration of 1 mM. The peptide solutions were incubated for 4 weeks at 25°C with shaking at 1000 r.p.m. prior to analysis. Transmission electron microscopy was applied to gather high contrast images of the peptides in triplicate. Morphological differences between the peptides were examined with a JEM-2100 microscope (JEOL, Japan) at 80 keV using negative staining with uranyl acetate of samples adsorbed on formvar film coated 400-mesh copper grids (Agar Scientific Ltd., England). If a peptide did not reveal fibrils by TEM after 4 weeks, the incubation period was prolonged for at least two weeks. ProteoStat dye staining was carried out in duplicate on all peptides using a BMG PolarStar platereader in 96-well format, mixing 30 µl of peptide solution with 70 µl of Proteostat Reagent (ENZO Life Sciences). Upon interaction with the cross-beta sheet structure of protein aggregation, the dye shows an increase in fluorescence intensity. Fourier transform infrared spectroscopy (FTIR) was measured using a Bruker Tensor 27 infrared spectrophotometer equipped with a Bio-ATR II accessory to determine the secondary structure of the peptide formations. A hexapeptide was identified as an amyloid forming fibril when the FTIR spectrum showed peaks around 1635 cm^{-1} and/or 1680 cm^{-1}. Additional hexapeptide properties such as WALTZ, TANGO (Fernandez-Escamilla *et al.*, 2004) and PASTA scores (Trovato *et al.*, 2007), hydrophobicity, Chou-Fasman values for helix and strand propensity (Chou and Fasman, 1974) were calculated and amyloid structural class (Eisenberg *et al.*, 2009) prediction and predicted atomic structure models were obtained through comparative modeling using the FoldX force field (Schymkowitz *et al.*, 2005) on the publically available structures of amyloid cores (Eisenberg *et al.*, 2005; Morris *et al.*, 2013). All peptide data is stored in a MySQL database and is available through a webserver built with the Drupal content management system to provide fast and secure data access. An extended version of these experimental methods are available in the help page of the website.

3 Results

3.1 Datasets

The database consists of these distinct sets of hexapeptides: (1) FUS: 49 peptides from Fused in sarcoma protein, (2) TDP-43: 94 peptides from TAR-DNA Protein 4, (3) SOD-1: 30 peptides from Superoxide Dismutase, (4) Sup35: 205 peptides from the yeast Sup35 prion protein, (5) Lindquist: 105 hexapeptides from a bioinformatics amyloid prediction study from the Lindquist group, (6) Functionals: 140 peptides from bacterial and yeast adhesins, (7) Diversity set: collection of 50 wild type and mutant hexapeptides from different unrelated disease proteins, (8) Newcores: 47 peptides where the hexapeptides cores positions 3 and 5 were systematically explored to eliminate residue bias on these positions, (9) Literature: 200 peptides from the Amylhex database (Maurer-Stroh, *et al.*, 2010) and 169 hexapeptides mined from the literature annotating their amylogenic properties.

3.2 Online database content

3.2.1 Peptide listing and filtering

The homepage of WALTZ-DB immediately lists a paged table of the entire database content. The visible columns are the peptide sequence, source (in-house or literature), morphology decision (amyloid or non-amyloid), WALTZ and TANGO score. The extensive filter block on the right serves to fine-tune a peptide search. Filtering can be done on sequence, morphology decision, source, hydrophobicity, UniProt identifier, PDB identifier, availability of TEM image or FTIR spectrum and amyloid- or aggregation-prediction by WALTZ, TANGO and PASTA. At the bottom of the table, two buttons activate a download of the resulting peptide list in CSV or Excel format.

3.2.2 Detailed peptide data

Clicking on a peptide sequence opens a peptide-centered page with detailed annotation and experimental data. When a TEM image was positive for fiber formation or even a vague impression of fiber formation was present, we show three TEM images, the FTIR spectrum and the ProteoStat dye staining values. For non-synthetic peptides, the location of the peptide sequence in the parent protein is colored red inside the full sequence. On top of the page is a summary including FTIR peak values, ProteoStat binding values, WALTZ, TANGO and PASTA scores, hydrophobicity and Chou-Fasman values for helix and strand propensity. WALTZ-DB also provides an atomic structure model for the peptide including information about the preferred amyloid structure class. The model is a PDB file and can be downloaded for analysis. At the bottom of each peptide-centered page there are links to external sites that offer amyloid prediction tools (De Baets, *et al.*, 2014), so the user can run the peptide through other predictors and annotation tools.

4 Summary

WALTZ-DB contains 1089 hexapeptides and is currently the largest database for amyloid morphology annotation. To this end, we experimentally verified the fiber forming potential for 720 hexapeptides through TEM, FTIR and ProteoStat dye staining. We make available the TEM images and FTIR spectra to allow researchers to reach independent decisions on the amyloid status of each peptide. In addition, we provide our own classification based on the presence of amyloid aggregates by TEM, supported by at least one evidence for beta-sheet structure by FTIR or Proteostat dye binding. As the database contains amyloid-positive as well as amyloid-negative samples, it serves perfectly as a reference set to develop novel prediction tools.

Funding

The Switch Laboratory was supported by grants from VIB, University of Leuven, the Funds for Scientific Research Flanders (FWO G.0509.13), the Flanders Institute for Science and Technology (IWT) and the Federal Office for Scientific Affairs of Belgium (Belspo IUAP P7/16). J.V.D. held a personal grant from the Fonds Wetenschappelijk Onderzoek Vlaanderen (FWOTM443). E.C. has been partially supported by an EMBO Short-Term Fellowship (EMBO ASTF 224-2012). GDB was supported by a postdoc fellowship of KU Leuven.

Conflict of Interest: none declared.

References

Chiti,F. *et al.* (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **424**, 805–808.

Chou,P.Y. and Fasman,G.D. (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211–222.

De Baets,G. *et al.* (2014) Predicting aggregation-prone sequences in proteins. *Essays Biochem.*, **56**, 41–52.

Eisenberg,D. and Jucker,M. (2012) The amyloid state of proteins in human diseases. *Cell*, **148**, 1188–1203.

Eisenberg,D. *et al.* (2005) Structural studies of amyloid. *FEBS J.*, **272**, 78–79.

Eisenberg,D. *et al.* (2009) Amyloid and prion structures. *FASEB J.*, **23**, 423.1.

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.

Fowler,D.M. *et al.* (2006) Functional amyloid formation within mammalian tissue. *PLoS Biol.*, **4**, e6.

Ho,M.-R. *et al.* (2006) Human pancreatitis-associated protein forms fibrillar aggregates with a native-like conformation. *J. Biol. Chem.*, **281**, 33566–33576.

Maurer-Stroh,S. *et al.* (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods.*, **7**, 237–242.

Morris,K.L. *et al.* (2013) Exploring the sequence-structure relationship for amyloid peptides. *Biochem. J.*, **450**, 275–283.

Schymkowitz,J. *et al.* (2005) The FoldX web server: an online force field, *Nucleic Acids Res.*, **33**, W382–W388.

Sunde,M. *et al.* (1997) Common core structure of amyloid fibrils by synchrotron X-ray diffraction, *J. Mol. Biol.*, **273**, 729–739.

Trovato,A. *et al.* (2007) The PASTA server for protein aggregation prediction, *Protein Eng. Desig. Select.*, **20**, 521–523.