FEBS Letters 587 (2013) 1089–1095

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

Breaking the amyloidogenicity code: Methods to predict amyloids from amino acid sequence

Abdullah B. Ahmed, Andrey V. Kajava *

Centre de Recherches de Biochimie Macromoléculaire, UMR5237 CNRS, Montpellier 1 et 2, 1919, Route de Mende, 34293 Montpellier Cédex 5, France
Institut de Biologie Computationnelle, 95 rue de la Galéra, 34095 Montpellier Cédex, France

ARTICLE INFO

Article history:
Received 4 December 2012
Revised 11 December 2012
Accepted 11 December 2012
Available online 20 December 2012

Edited by Wilhelm Just

Keywords:
Amyloid fibril
Bioinformatics
Neurodegenerative disease
Protein misfolding

ABSTRACT

Numerous studies have shown that the ability to form amyloid fibrils is an inherent property of the polypeptide chain. This has lead to the development of several computational approaches to predict amyloidogenicity by amino acid sequences. Here, we discuss the principles governing these methods, and evaluate them using several datasets. They deliver excellent performance in the tests made using short peptides (~6 residues). However, there is a general tendency towards a high number of false positives when tested against longer sequences. This shortcoming needs to be addressed as these longer sequences are linked to diseases. Recent structural studies have shown that the core element of the majority of disease-related amyloid fibrils is a β-strand-loop-β-strand motif called β-arch. This insight provides an opportunity to substantially improve the prediction of amyloids produced by natural proteins, ushering in an era of personalized medicine based on genome analysis.

1. Introduction

Scientists have been interested in the ability of the amino acid sequence of a protein to determine its structural state for over 50 years. The foremost efforts were devoted to studying globular proteins [1]. Later on, researchers set their sights on the intrinsically unstructured regions of proteins making significant progress in the understanding of their sequence code [2,3]. However, it has been shown that these lines of thought are insufficient to understand the complexities of protein folding. Over the last two decades, numerous studies have demonstrated that, depending on conditions and (or) the amino acid sequence, otherwise globular or unstructured proteins can assemble into insoluble, stable structures of unlimited dimensions consisting of either amyloid fibrils or amorphous aggregates [4–8]. It is becoming evident that an accurate estimation of the structural state(s) encoded by a given amino acid sequence requires evaluation of the individual probabilities of the protein to have either soluble 3D structure, an unstructured state, or insoluble structures, as well as the likelihoods of transition between the states of this triad (Fig. 1). It is important to note that the amyloidogenic form of the insoluble state is attracting special interest as it is linked to a number of human diseases. In this review we focus on existing approaches to predict the propensity of proteins and peptides to form amyloids based on the analysis of their amino acid sequences.

Although amyloidogenic precursor proteins vary with respect to their amino acid sequence and native fold, the resulting amyloid fibrils share similar generic properties. They are typically straight, rigid, between 4 and 13 nm in diameter, thermostable, protease-resistant, and rich in β-structure [9–12]. Amyloid fibrils are the subject of special interest mainly due to their link to a broad range of human diseases, which include, but are not limited to, type II diabetes, rheumatoid arthritis, and perhaps most importantly, debilitating neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, and Huntington's disease. Although, admittedly, it has been shown that in some organisms amyloid structures can also play important, "beneficial" roles [5]. The scope of amyloid studies has broadened with the discovery of many proteins that are not normally amyloidogenic but may be induced to form amyloid fibrils in vitro [13]. Currently, in addition to this, the problem of amyloid formation is receiving increasing attention from biotechnologists searching for ways to avoid the accumulation of recombinant proteins into aggregates [6].

However, despite considerable interest, and much effort put toward understanding of the sequence-structure relationship of amyloid fibrils, this structural state remains the least studied compared with soluble structured and unstructured proteins. This situation may be attributed in part to the limited number of studied amyloids and that the methods of determining high-resolution

Abbreviations: GFP, green fluorescent protein; ROC, receiver operator characteristic; H-bond, hydrogen bond; PSSM, position specific scoring matrix

* Corresponding author.
E-mail address: andrey.kajava@crbm.cnrs.fr (A.V. Kajava).

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2012.12.006

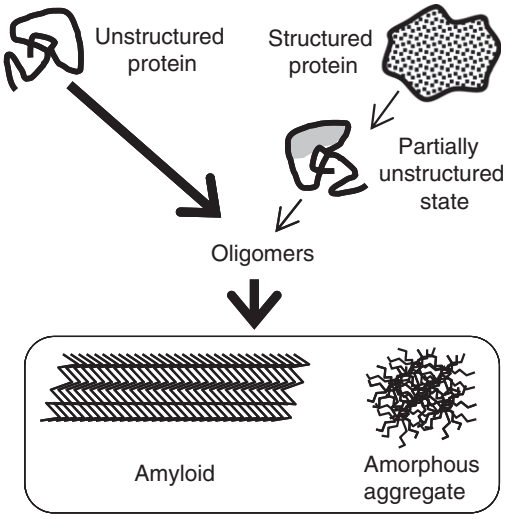

Fig. 1. Simplified scheme of relationship between the principal structural states of proteins (soluble structured, unstructured states and insoluble amorphous aggregates and amyloid fibrils). Most of the known disease-related proteins form amyloids. The likelihoods of transition between these states are denoted by the thickness of the arrows. In the majority of cases a polypeptide chain is unfolded prior to aggregation. A structured protein with the amyloidogenic potential must become partially or completely unfolded to form the amyloid fibril or amorphous aggregates. In reality, the protein aggregation pathways are more complicated involving multiple intermediate stages for both natively structured and unstructured proteins [7–8].

structure (protein crystallography and NMR spectroscopy) cannot be used because of the insolubility of fibrils. Nevertheless, over the last decade, numerous studies have demonstrated that just like globular and unstructured states, the propensity to form amyloids is coded by the amino acid sequence. Based on this data, several methods for prediction of amyloidogenicity have been proposed. Here we discuss these approaches and the principles behind them. New data about amyloidogenesis which may be critical for improvement of the current methods are also presented. The list of described methods is not exhaustive. Our intention was to cover most of them, selecting those that are the most popular, most original and diverse in terms of the basic principles, and those that can be downloaded or used via web-servers (Table 1).

2. Methods for the prediction of amyloid fibril formation

2.1. Methods that rely on individual amino acid aggregation propensities, and the composition of amyloidogenic regions

Unlike soluble structured proteins where similar sequence motifs correspond to 3D structural resemblance, the proteins and peptides that form amyloids have very different sequences. This suggests that it is the amino acid composition rather than sequence motifs that may be of critical influence to amyloidogenicity. As a result, several approaches rely on experimental or theoretical data of individual amino acid aggregation propensities and the evaluation of amino acid composition of amyloidogenic regions [14–17]. Here, we discuss the two most recent programs, AGGRESCAN [16] and FoldAmyloid [17], as approaches having different backgrounds and easily accessible by their corresponding web-servers.

The AGGRESCAN program is based on the assumption that short (5–11 residues) sequences or "hot spots" can nucleate aggregation in peptides and proteins, and that the propensity of these "hot spots" is determined by their amino acid composition. The aggregation–propensity scale for individual amino acids (Table 2) was derived from the following experimental data. The C-terminus of the 42-residue long human Aβ-peptide was linked by 12 residue fragment to a green fluorescent protein (GFP) [18]. It was shown that *Escherichia coli* cells express a high amount of this fusion protein but exhibit little fluorescence, indicating that the presence of the aggregation–prone Aβ42 peptide interferes with the correct folding of the GFP and thus with the emission of fluorescence.

Table 2
Aggregation–propensity scales for individual amino acids derived by different approaches.*

AGGRESCAN, de Groot et al. (2006)	FoldAmyloid, Garbuzynskiy et al. (2010)	Pawar et al. (2005) (pH 7)
I 1.822	**I 1.217**	**W 2.92**
F 1.754	**W 1.027**	**F 2.80**
V 1.594	**L 1.015**	**C 1.61**
L 1.380	**F 0.958**	**Y 1.03**
Y 1.159	**V 0.92**	**I 0.93**
W 1.037	**Y 0.851**	**V 0.49**
M 0.910	**M 0.725**	**L −0.25**
C 0.604	C 0.568	**M −1.06**
———————	A 0.086	———————
A −0.036	———————	T −2.12
T −0.159	R 0.032	A −3.31
S −0.294	H 0.025	G −3.96
P −0.334	S −0.73	H −4.31
G −0.535	Q −0.271	S −5.08
K −0.931	T −0.349	Q −6.00
H −1.033	K −0.565	N −6.02
Q −1.231	E −0.632	D −9.42
R −1.240	N −0.713	K −9.55
N −1.302	D −0.776	E −10.38
E −1.412	G −1.088	R −11.93
D −1.836	P −2.303	P −11.96

* Aggregation propensity decreases from top to bottom. Bulky apolar residues are in bold. Dashed lines indicate the boundaries between amyloidogenic and non-amyloidogenic amino acids (for AGGRESCAN −0.02, for FoldAmyloid triple hybrid scale 0.062).

Table 1
Methods to predict amyloids from amino acid sequences.*

Name	Basic approach	Server/Website
AGGRESCAN	Composition of amino acids	http://bioinf.uab.es/aggrescan/
FoldAmyloid	Composition of amino acids	http://bioinfo.protres.ru/fold-amyloid/oga.cgi
Zyggregator	Amino acid aggregation propensities and properties of β-structural conformation	http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php
TANGO	Properties of β-structural conformation	http://tango.crg.es/
PASTA	Pairwise interactions within the β-sheets	http://protein.bio.unipd.it/pasta/
BetaScan	Pairwise interactions within the β-sheets	http://groups.csail.mit.edu/cb/betascan/betascan.html
3D Profile method (ZipperDB)	Amyloid-like structures of short peptides	http://services.mbi.ucla.edu/zipperdb/submit
Waltz	Amyloid-like structures of short peptides	http://waltz.switchlab.org/
NetCSSP	Conformational switches	http://cssp2.sookmyung.ac.kr/index.html
AmylPred	Conformational switches	http://biophysics.biol.uoa.gr/AMYLPRED/

* Described in this review and accessible via internet.