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 sh [illegible] understand t [illegible] 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 [illegible] 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 © 2012 Federation of European Biochemical Societies. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.febslet.2012.12.006

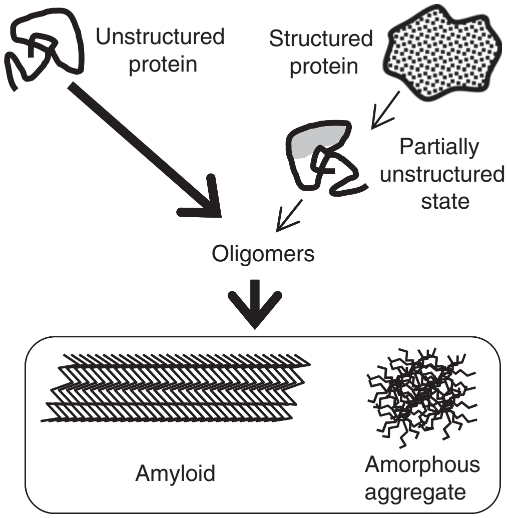

Fig. 1. Simplified scheme of relationship between the principal structural states of proteins (soluble structured, unstructured states and insoluble amorphous aggregates and amyloid fibrils). Most of the known disease-related proteins form amyloids. The likelihoods of transition between these states 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].

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).

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 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.

Using this in vivo system, the fluorescence of mutants with Phe19 substituted by all other amino acids was tested and these results were used to create the aggregation–propensity scale for amino acids (Table 2). A "hot spot" is defined as a region that contains five or more consecutive residues with an average aggregation propensity value higher than the average of the 20 naturally occurring amino acids weighted by their frequencies in the Swiss-Prot database.

The FoldAmyloid program, like AGGRESCAN, uses the assumption that short sequences of 5 residues are sufficient for amyloidogenesis and applies a sliding window averaging technique to find them [17]. The main difference is the derivation of individual amino acid aggregation propensities from the statistical analysis of known 3D structures of globular proteins. It was shown that two characteristics (the mean number of atom–atom contacts per residue, and the mean number of backbone H-bonds per residue) correlate well with amyloidogenicity. The FoldAmyloid program predicts amyloidogenic regions using either one of these amino acids scales (contacts, backbone H-bonds of acceptors or donors), or a hybrid scale which includes all three (Table 2). The cut-off values optimal for amyloidogenic prediction were selected based on receiver operator characteristic (ROC) curves obtained by tests on sets of the known amyloidogenic and non-amyloidogenic peptides.

2.2. Methods that rely on individual amino acid aggregation propensities and the properties of β-structural conformation

The major building blocks of amyloids are β-strands, which have an extended conformation with conserved apolar and variable (generally polar) residues alternating along the chain. A number of methods use this information to improve the prediction of amyloidogenic regions.

One of them is the Zyggregator method that takes into consideration patterns of seven or more residues with alternating apolar and polar residues [19]. To calculate the aggregation propensity, this method also uses a set of physico-chemical properties of amino acid residues such as hydrophobicity, charge, and the propensity to adopt α-helical or β-structural conformations. These properties were derived by fitting the expression used to calculate the aggregation propensity on a database of mutational variants for which aggregation was measured in vitro [14,20]. Zyggregator also considers the flanking residues ("gatekeeper" residues) of a given sliding window for the presence of charged residues of the same sign, as this may reduce aggregation by electrostatic repulsion. In a majority of cases a polypeptide chain should be unfolded to aggregate. Therefore, when applied to structured proteins, prediction methods need to estimate probability of the protein or parts of it to be unstructured. Zyggregator has this option, evaluating the local stability of protein structure by CamP program [21].

The TANGO predictor of β-structural aggregation [22] uses a statistical mechanics approach to make secondary structure predictions. For a given sequence this method considers different competing conformations (random coil, β-turn, α-helix, and β-sheets) and predicts which is most likely to occur. The algorithm is based on the following assumptions: (i) a particular amino acid sequence is aggregation–prone if it has high propensity to form β-structure, (ii) all residues of the β-region are buried in the hydrophobic interior of the aggregate, (iii) complementary charges in the selected window establish favorable electrostatic interactions, and (iv) deviating from neutral net charge disfavors aggregation of the peptide. TANGO considers that peptides have a tendency for aggregation when they possess segments of at least five consecutive residues in the predicted β-aggregate conformation. Zyggregator and TANGO both take into account the effect of physico-chemical conditions such as pH, temperature, ionic strength, and the trifluorethanol concentration on aggregation.

2.3. Methods that rely on pairwise side-chain to side-chain interactions within the β-sheets

A β-strand cannot exist alone. It is stabilized only when involved in β-sheets, where several β-strands interact with each other via peptide group H-bonds. Although the main interaction occurs by the H-bonds, side-chain to side-chain interactions of the adjacent β-strands may provide additional sequence specific stability. Consideration of these side-chain interactions may improve the correct prediction of the β-strand regions and their arrangement within the β-sheets. Therefore, some methods use the data on the propensity of pairwise side-chain side-chain interactions within the β-sheets to predict amyloidogenicity.

For example, the PASTA program [23,24] uses a non-redundant set of known globular structures to count for pairs of amino acids that form contacts (C_α atoms less than 6.5 Å) between the adjacent β-strands of the β-sheets. The occurrence of the amino acid pairs was analyzed separately for parallel and anti-parallel β-strands. Finally, the pairwise scores were calculated by using a Boltzmann energy function derived from the amino acid contact occurrence. The score was used to predict localization and a preferable 3D arrangement of β-strand pairs (parallel or antiparallel, shifted or in register) in a given protein. It was assumed that protein can form amyloids via interaction of a short (four residues or more) β-structural region.

The BETASCAN program, also relies on β-pairing propensities, specifically focusing on the parallel orientation of β-strands, as it occurs most frequently in amyloid fibrils [25]. It calculates the likelihood scores for potential β-strands and strand-pairs based on correlations observed in known amphipathic parallel β-sheets with one face contacting hydrophibic interior and the other face exposed to the solution. The likelihood of a sequence to form parallel β-strands is calculated as the propensity for this sequence to occur as a β-strand multiplied by its propensity to form β-strand pairs. BETASCAN then uses a hill-climbing algorithm to determine if rotation of the β-strands by 180° around its axis, the addition or subtraction of residues to the fibril forming region, or the shifting the first or second β-strand pairs can give rise to structures more likely to form parallel β-strands and therefore predicted to be more amyloidogenic. BETASCAN uses a β-strand window of 3–13 residues.

2.4. Methods inspired by the establishment of the amyloid-like structures of short peptides

The approaches mentioned above were based on the analysis of known 3D globular structures. However, since 2005, several crystal structures in which short peptides engage in amyloid-like interactions have been determined [26,27]. These structures provided details of side-chain interactions inside the double β-sheet, which is also called a "cross-β spine". The basic template represents two parallel β-sheets oriented antiparallel to one another with the interface formed by the like-sides of each sheet. The improved understanding of the interactions of β-strands in microcrystals of short amyloidogenic peptides inspired new approaches.

The 3D profile method [28] uses the "cross-β spine" structure formed by NNQQNY fragment from the sup35 prion protein of *Saccharomyces cerevisiae* [26]. Initially, a set of 3D templates were built from the crystal structure of the peptide NNQQNY by small displacements of one of the two interacting β-sheets relative to the other. Each six-residue peptide of an analyzed protein is mapped onto these templates and the energy of each sequence to the profile mapping was evaluated using the ROSETTADESIGN program [29,30]. A region is considered to be amyloidogenic if the energy evaluated in this manner is below a threshold value. To test the performance of the program the AmylHex database of 67 known fibril-forming and 91 non-fibril-forming hexapeptides was compiled from the literature.

The similar, Statistical Potential Method [31] also uses the 3D templates generated by small displacements of the crystal structure of peptide NNQQNY [26]. However, residue-based statistical potential calculations, instead of ROSETTADESIGN analysis were applied to evaluate the energy of each sequence mapped onto these templates.

Another method, called Waltz, [32] used the AmylHex dataset [28] not for benchmarking, but as a learning set to generate the position specific scoring matrix (PSSM) of hexapeptides for identification of amyloid-forming sequences. For this purpose, the AmylHex dataset was supplemented by a number of new experimentally determined amyloidogenic and non-amyloidogenic peptides. In addition to the PSSM, 19 physical properties of amino acids (such as β-structure propensity, and hydrophobicity) that strongly correlated with their frequencies in the positive and negative hexapeptide learning sets were selected to predict amyloidogenicity. Finally, Waltz program uses a position specific pseudo energy matrix derived as follows. The crystal structure of Sup35 GNNQQNY peptide [26] was reduced to poly-alanine and all possible combinations of naturally occurring amino acids were made. The structures were optimized and their energy was estimated by using the FoldX program [33]. The three terms (PSSM, physicochemical and structure-derived) are combined in the composite scoring function used to predict the amyloidogenicity.

2.5. Methods that estimate probability of structured proteins to be partially unfolded

To form cross-β amyloids, a polypeptide chain with high amyloidogenic potential needs to be unstable within its native 3D structure or be completely unfolded. Indeed, experimental studies show that most of the known amyloid-forming sequences (for example, amyloid-β, α-synuclein, Ure2p, and Sup35p) are unstructured in their non-amyloid state. Proteins that fold into soluble 3D structures may also contain a number of amyloidogenic regions hidden in their structures. Significant efforts have been dedicated to the identification of such hidden regions (also known as 'conformational switches' or "chameleon" sequences) within globular proteins that are innocuous in their normal state [34].

Some methods developed for prediction of amyloidogenicity address this problem. For example, the Zyggregator method includes an option to evaluate the local stability of protein structure [21]. The Net-CSSP method (contact-dependent secondary structural propensity) [35,36] quantifies the influence of tertiary interactions on secondary structure preference by using an artificial neural network-based algorithm and seeks to find short regions with a hidden potential to form β-sheets.

Another web-based tool, AmylPred, combines the results of amyloidogenicy predictions with the SecStr secondary structure prediction tool [37]. The SecStr tool uses six different algorithms to give a secondary structure prediction. If it predicts that amino acid stretches have ambivalent propensities for α-helix and β-strand, they are considered to be regions with potential 'conformational switches'. After that several approaches such as, FoldAmyloid [17] and scanning of proteins with amyloidogenic motif extracted from the known fibril-forming peptides [38] are applied to the sequence. Regions of the structured protein that are simultaneously identified as the 'conformational switches' and highly amyloidogenic considered to be the amyloidogenic determinants.

3. Performance of methods

To evaluate prediction methods, benchmark datasets of amyloid-forming and non-forming sequences are required. When doing so, the primary problem is the limited number of known amyloid-forming proteins. Today, only about 20 amyloid-forming proteins are known to be linked to diseases [39]. Although datasets can be enriched by adding known mutants of these proteins, this is not a solution, as the datasets become biased towards certain over-represented sequences. Moreover, prediction methods are designed to exclusively detect cross-β amyloids, whereas disease-related fibrils are heterogeneous in terms of their 3D structure. Some are formed by stacks of native or refolded globular structures, [40–42] and do not necessarily exhibit cross-β structure. Care must also be taken when developing the negative set. It is tempting to use globular proteins as they are soluble and non-amyloidogenic. Most prediction programs, however, operate using only sequence information, and will incorrectly predict amyloidogenic candidates that are in fact hidden inside the protein structure. Furthermore, when one considers that different amyloid-forming proteins form fibrils at different conditions (concentration, ionic strength, pH, etc.) it becomes evident that the task to construct testing datasets of high quality is extremely challenging.

Most of the methods use datasets of short peptides. The reasons are that short peptides can be synthesized easily and tested in the same or similar experimental conditions for the formation of amyloid fibrils. Moreover, soluble short peptides can be used directly as a non-amyloidogenic set. As these peptides are unfolded, they do not have the problem of structurally hidden regions found in folded proteins. Finally, the usage of short peptides is in agreement with the predominant paradigm underlying existing prediction algorithms: short (about 6 residues) regions are sufficient for forming amyloid fibrils of full-length proteins.

There are several popular benchmark datasets of short peptides. The first large dataset was compiled for the testing of the TANGO algorithm [22] and consisted of 78 amyloidogenic and 172 non-amyloidogenic peptides mostly from human disease related proteins. Peptides were considered to be aggregating when their circular dichroism or NMR spectra had concentration dependence in the range between 1 μM and 5 mM, or when binding to an amyloid-reporting dye (thioflavine T) was observed. Another set of experimentally determined amyloid-forming peptides was selected from the literature and used to test AGGRESCAN program [16]. The most frequently used data set is AmylHex. It contains 158 six-residue peptides of which 67 have been shown to form fibrils and 91 are soluble [28]. A majority of the dataset consists of mutants of STVIIE peptide, as well as hexapeptides and their mutants from amylin, tau, insulin, β2-microglobulin. Recently, the AmylHex dataset was supplemented by 49 new amyloid-forming and 71 non-amyloid-forming hexapeptide sequences [32] to bring the total number of amyloid forming hexapeptides to 116 positive and 103 negative sequences. Several other predictors of amyloidogenicity used one of the datasets mentioned above or their combinations.

Fig. 2 shows our benchmarking results for three programs (TANGO, AGGRESCAN and Waltz) on a combined set of the sequences from all the datasets mentioned above. The tested programs display good results, correctly identifying 65%, 71% and 80% of the amyloid-forming peptides, correspondingly, and having only 17%, 25% and 15% of false positives in the set of non-amyloidogenic peptides. Waltz performs better than the other programs, however, it is necessary to remember that a large number of peptides from the combined dataset were used by this program as a training set [32].

The other approach typically used to demonstrate the power of the methods was the prediction of known pathogenic or protective mutants of amyloid-forming proteins to demonstrate the ability to predict the observed change in the amyloidogenicity [16,22]. In addition, the programs are tested for the prediction of locations of amyloid-forming regions in longer peptides (30–40 residues) and full-length proteins. Especially those, with a natively unfolded monomeric state, and experimentally verified locations of amyloid

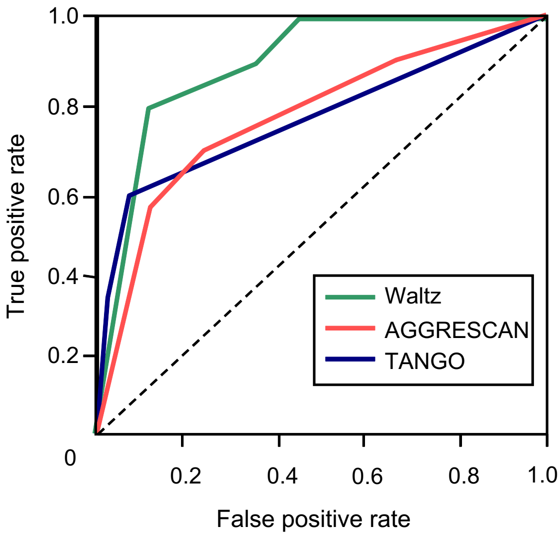

Fig. 2. Comparison of ROC curve performance on the combined dataset composed of sequences compiled by the authors of TANGO, AGGRESCAN and AmylHex (including the sequences added by the authors of the Waltz program). The combined dataset contains 243 amyloid forming and 333 non-forming sequences. The programs were chosen for their ability to analyse multiple sequences simultaneously at their web-servers.

Table 3
Performance of different methods on datasets of proteins.*

Program**	True positive rate	False positive rate
Waltz	0.666 (12/18)	0.346 (18/52)
Tango	0.277 (5/18)	0.500 (26/52)
Aggrescan	0.722 (13/18)	0.769 (40/52)
FoldAmyloid	0.388 (7/18)	0.750 (39/52)
AmylPred	0.833 (15/18)	0.673 (35/52)

True positive rate: (Number of true positives)/(Total number of amyloid-forming sequences).
False positive rate: (Number of false positives)/(Total number of non-amyloidogenic sequences).

* The set of amyloid-forming sequences is composed of proteins or peptides known to form amyloids in vivo that were taken from literature with the following criteria: they are non-globular in their native state, form cross-beta fibrils under physiological conditions, and their sequences are longer than 20 amino acids. This set contains 18 sequences (see Supplementary data).The negative set was extracted from the DisProt database of disordered proteins (Vucetic et al., 2005) with the following criteria: sequences are disordered in their entirety and have less than 150 residues. The negative set contains 52 sequences (see Supplementary data).

** The default settings of the web-servers were used.

forming regions (Fig. 3). The most frequently used examples for such tests are amyloid-β, α-synuclein and amylin. In Fig. 3, the predictions of amyloidogenic "hot spots" in fibril-forming regions of amyloid-β and Het-s prion are shown. The programs generate satisfactory predictions for amyloid-β peptide, while in the Het-s prion region, the predictions are less credible. For example, Waltz program does not find any amyloid-forming region within the Het-s prion domain. This can be explained by the absence of the Het-s peptides in its training set, or by some differences of the Het-s fibril structure from the typical cross-β amyloids. The amyloid-β structure represents a stack of identical peptides, but the Het-s cross-β fibril is formed by the repetitive element with two slightly different beta-strands alternating along the fibril axis.

The performance of the programs can be summarized thusly. They accurately predict short amyloid-forming peptides, and are adept at determining experimentally established fibril-forming regions in full-length proteins. However, most of the methods generate a large number of false positives when applied to the sequences of longer than 30–40 residues. Another problem of these methods is the over prediction of amyloids in hydrophobic regions and their poor predictive capability of amyloidogenic sequences rich in polar Gln and (or) Asn. This shortcoming can be explained by the fact that some methods use aggregation propensities values obtained from the analysis of globular proteins which have the hydrophobic residues as the predominant structure-stabilizing factor.

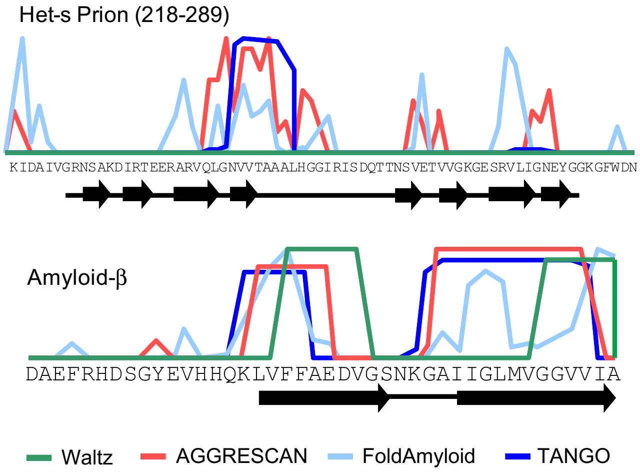

Fig. 3. Prediction of localization of amyloidogenic regions in the Het-s prion domain (218–289) and in amyliod-β peptide. Lines with arrows below the sequence denote the known regions involved in the fibril structure. The β-strands are shown by arrows.

It must be emphasised that the eventual goal for all methods be the correct prediction of amyloid fibril formation in naturally occurring proteins and peptides. However, the existing programs yield unconvincing results, generally predicting a sizable number of false positives (Table 3).

4. New structural data as a basis for improvement of the algorithms

Our retrospective analysis of the methods for prediction of amyloidogenicity reveals that an appropriate consideration of the structural properties of amyloids is a key factor for improving of the performance of these programs. Indeed, the progression of known methods shows a pattern of increasing usage of structural information. They started with the consideration of the amino acid composition of proteins, then moved onto the β-structural pattern representing alternation of polar and apolar residues, then to the analysis of side-chain interactions within a β-sheet, before finally supplementing it with the analysis of side-chain packing between the β-sheets.

Recently, new experimental approaches have shed more light on the details of the 3D structural arrangement of amyloid fibrils. Progress was made by the application of new experimental techniques such as solid state nuclear magnetic resonance, cryo-electron microscopy, scanning transmission electron microscopy mass measurements, and electron paramagnetic resonance spectroscopy, in conjunction with more established approaches such as X-ray fiber diffraction, conventional electron microscopy, and optical spectroscopy [43–47] As a result, it was shown that a majority of structural models of disease-related amyloid fibrils can be reduced to a so called "β-arcade" (Fig. 4) [48]. This fold represents a columnar structure produced by stacking of β-strand-loop-β-strand motifs called "β-arches" [49]. Each β-arcade has a double-layer structure in which 2 parallel in-register β-sheets face each other. The side chains protruding from opposing β-sheets form tight inter-digitated packing. Variations of β-arch

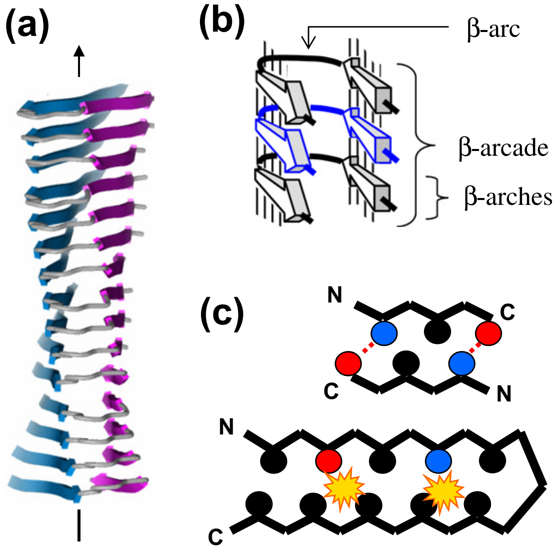

Fig. 4. The β-arcade as a common structural motif of disease-related amyloid fibrils. (a) Overall view of the β-arcade protofibril with the first β-sheet in blue and the second one in magenta. (b) Detailed view of the β-arcade arrangement. The structure is formed by the axial stacking of identical β-arches consisting of 2 long β-strands (shown as arrows) connected by a β-arc. The arches are H-bonded (thin lines) along the fibril axis and form a double layer of parallel β-sheets. (c) An example demonstrating that amyloidogenic sequence motifs in short peptides can be non-amyloidogenic in the β-arcade structure. In the first case, an axial view of a short double layer formed by identical parallel β-sheets is shown. The short β-strands have a sequence motif with positively (blue circle) and negatively (red circle) charged residues separated by one apolar residue (black circle) on their interior side. The charged side chains can form favourable salt bridges. In contrast, such a sequence motif within the β-arcade structure should be energetically unfavourable due to the location of uncompensated charges incapable of forming salt bridges inside the apolar environment. For the sake of clearness the outside side chains are not shown.

arrangements may lead to either single β-arcade structures [50,51], superpleated β-structures with several adjacent β-arcades [52] or β-arches within the β-solenoids [53].

The prevalence of β-arches in disease-related amyloids will certainly have implications for identifying amyloidogenic sequences. The amyloidogenic region capable of forming the β-arcade structure needs to be over 15–20 residues. This may lead to the revision of the paradigm that short 6–10 residue segments of protein can initiate its amyloidosis. In addition, one can imagine cases where methods based on the prediction of short amyloidogenic regions will fail to detect the β-arch regions of high amyloidogenic potential (see for example, Fig. 4).

5. Conclusions

Several computational methods have been developed to predict the propensity of polypeptides to form amyloids based on sequence analysis. Many of the methods have rendered excellent performance capabilities in the numerous tests. These algorithms use the assumption that a short sequence (about 6 residues) is sufficient to trigger the amyloid formation of a given protein. Consequently, they achieve their best results among short peptides. However, the analysis of short peptides is largely un-equivalent to the in vivo formation of disease related amyloids. Indeed, peptides of less than about 15 residues rarely reach fibril-forming concentrations in human cells, as once produced, they are rapidly degraded by endogenous proteases [54]. Although it is true that a short fibril-forming region may occur within a longer polypeptide chain, fusion of short amyloidogenic peptides with soluble proteins has not yielded convincing results, only triggering fibrillation at high concentrations [55,56]. Additionally, known naturally occurring amyloid-forming proteins have amyloidogenic regions that are longer than 15 residues. Finally, recent experimental techniques reveal that the minimal structural element of the majority of disease-related amyloid fibrils is a columnar structure produced by stacking of β-strand-loop-β-strand motifs spanning over 15–20 residues.

Given these considerations, we may expect the development of new bioinformatics tools with improved prediction when applied to the long peptides or full-length proteins. These kinds of tools are particularly relevant to the disease-related amyloids and especially needed because currently no reliable ways to diagnose the early stages of such diseases are available. Furthermore, thanks to a radical drop in the cost of sequencing an individual's genome, such bioinformatics tools are becoming extremely timely. With further research, an accurate risk profile might enable individuals to take steps to prevent diseases for which they are at increased risk based on genetics.

Acknowledgements

The authors thank M. Pervaiz and F. Richard for critical reading of the manuscript and suggestions, and M.-T. Chateau for assistance with the datasets. Funding: Overseas Scholarship Scheme Phase II grant by the Higher Education Commission of Pakistan to A. B. A.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.febslet.2012.12.006.

References

[1] Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. Science 181, 223–230.
[2] Uversky, V.N., Gillespie, J.R. and Fink, A.L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41, 415–427.
[3] Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G., et al. (2005) DisProt: a database of protein disorder. Bioinformatics 21, 137–140.
[4] Dobson, C.M. (2001) Protein folding and its links with human disease. Biochem. Soc. Symp., 1–26.
[5] Otzen, D. and Nielsen, P.H. (2008) We find them here, we find them there: functional bacterial amyloid. Cell. Mol. Life Sci. 65, 910–927.
[6] Ventura, S. and Villaverde, A. (2006) Protein quality in bacterial inclusion bodies. Trends Biotechnol. 24, 179–185.
[7] Uversky, V.N. and Fink, A.L. (1698) Conformational constraints for amyloid fibrillation: the importance of being unfolded. Biochim. Biophys. Acta 2004, 131–153.
[8] Fandrich, M. (2012) Oligomeric intermediates in amyloid formation: structure determination and mechanisms of toxicity. J. Mol. Biol. 421, 427–440.
[9] Shirahama, T. and Cohen, A.S. (1967) High-resolution electron microscopic analysis of the amyloid fibril. J. Cell Biol. 33, 679–708.
[10] Eanes, E.D. and Glenner, G.G. (1968) X-ray diffraction studies on amyloid filaments. J. Histochem. Cytochem. 16, 673–677.
[11] Kirschner, D.A., Inouye, H., Duffy, L.K., Sinclair, A., Lind, M. and Selkoe, D.J. (1987) Synthetic peptide homologous to beta protein from Alzheimer disease forms amyloid-like fibrils in vitro. Proc. Natl. Acad. Sci. USA 84, 6953–6957.
[12] Serpell, L.C., Fraser, P.E. and Sunde, M. (1999) X-ray fiber diffraction of amyloid fibrils. Methods Enzymol. 309, 526–536.
[13] Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G. and Dobson, C.M. (1999) Designing conditions for in vitro formation of amyloid protofilaments and fibrils. Proc. Natl. Acad. Sci. USA 96, 3590–3594.
[14] DuBay, K.F., Pawar, A.P., Chiti, F., Zurdo, J., Dobson, C.M. and Vendruscolo, M. (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. J. Mol. Biol. 341, 1317–1326.
[15] Rojas Quijano, F.A., Morrow, D., Wise, B.M., Brancia, F.L. and Goux, W.J. (2006) Prediction of nucleating sequences from amyloidogenic propensities of tau-related peptides. Biochemistry 45, 4638–4652.
[16] Conchillo-Sole, O., de Groot, N.S., Aviles, F.X., Vendrell, J., Daura, X. and Ventura, S. (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinf. 8, 65.
[17] Garbuzynskiy, S.O., Lobanov, M.Y. and Galzitskaya, O.V. (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics 26, 326–332.

[18] de Groot, N.S., Aviles, F.X., Vendrell, J. and Ventura, S. (2006) Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. FEBS J. 273, 658–668.

[19] Tartaglia, G.G. and Vendruscolo, M. (2008) The Zyggregator method for predicting protein aggregation propensities. Chem. Soc. Rev. 37, 1395–1401.

[20] Chiti, F., Stefani, M., Taddei, N., Ramponi, G. and Dobson, C.M. (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 424, 805–808.

[21] Tartaglia, G.G., Cavalli, A. and Vendruscolo, M. (2007) Prediction of local structural stabilities of proteins from their amino acid sequences. Structure 15, 139–143.

[22] Fernandez-Escamilla, A.M., Rousseau, F., Schymkowitz, J. and Serrano, L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat. Biotechnol. 22, 1302–1306.

[23] Trovato, A., Seno, F. and Tosatto, S.C. (2007) The PASTA server for protein aggregation prediction. Protein Eng., Des. Sel. 20, 521–523.

[24] Trovato, A., Chiti, F., Maritan, A. and Seno, F. (2006) Insight into the structure of amyloid fibrils from the analysis of globular proteins. PLoS Comput. Biol. 2, e170.

[25] Bryan Jr., A.W., Menke, M., Cowen, L.J., Lindquist, S.L. and Berger, B. (2009) BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. PLoS Comput. Biol. 5, e1000333.

[26] Nelson, R., Sawaya, M.R., Balbirnie, M., Madsen, A.O., Riekel, C., Grothe, R. and Eisenberg, D. (2005) Structure of the cross-beta spine of amyloid-like fibrils. Nature 435, 773–778.

[27] Sawaya, M.R., Sambashivan, S., Nelson, R., Ivanova, M.I., Sievers, S.A., Apostol, M.I., Thompson, M.J., Balbirnie, M., Wiltzius, J.J., McFarlane, H.T., et al. (2007) Atomic structures of amyloid cross-beta spines reveal varied steric zippers. Nature 447, 453–457.

[28] Thompson, M.J., Sievers, S.A., Karanicolas, J., Ivanova, M.I., Baker, D. and Eisenberg, D. (2006) The 3D profile method for identifying fibril-forming segments of proteins. Proc. Natl. Acad. Sci. USA 103, 4074–4078.

[29] Liu, Y. and Kuhlman, B. (2006) RosettaDesign server for protein design. Nucleic Acids Res. 34, W235–238.

[30] Simons, K.T., Bonneau, R., Ruczinski, I. and Baker, D. (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins (Suppl 3), 171–176.

[31] Zhang, Z., Chen, H. and Lai, L. (2007) Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. Bioinformatics 23, 2218–2225.

[32] Maurer-Stroh, S., Debulpaep, M., Kuemmerer, N., Lopez de la Paz, M., Martins, I.C., Reumers, J., Morris, K.L., Copland, A., Serpell, L., Serrano, L., et al. (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat. Methods 7, 237–242.

[33] Guerois, R., Nielsen, J.E. and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. 320, 369–387.

[34] Chiti, F., Taddei, N., Bucciantini, M., White, P., Ramponi, G. and Dobson, C.M. (2000) Mutational analysis of the propensity for amyloid formation by a globular protein. EMBO J. 19, 1441–1449.

[35] Yoon, S. and Welsh, W.J. (2004) Detecting hidden sequence propensity for amyloid fibril formation. Protein Sci. 13, 2149–2160.

[36] Kim, C., Choi, J., Lee, S.J., Welsh, W.J. and Yoon, S. (2009) NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation. Nucleic Acids Res. 37, W469–473.

[37] Hamodrakas, S.J., Liappa, C. and Iconomidou, V.A. (2007) Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins. Int. J. Biol. Macromol. 41, 295–300.

[38] Lopez de la Paz, M. and Serrano, L. (2004) Sequence determinants of amyloid fibril formation. Proc. Natl. Acad. Sci. USA 101, 87–92.

[39] Pepys, M.B. (2006) Amyloidosis. Annu. Rev. Med. 57, 223–241.

[40] Westermark, P., Sletten, K., Johansson, B. and Cornwell 3rd, G.G. (1990) Fibril in senile systemic amyloidosis is derived from normal transthyretin. Proc. Natl. Acad. Sci. USA 87, 2843–2845.

[41] Sanders, A., Jeremy Craven, C., Higgins, L.D., Giannini, S., Conroy, M.J., Hounslow, A.M., Waltho, J.P. and Staniforth, R.A. (2004) Cystatin forms a tetramer through structural rearrangement of domain-swapped dimers prior to amyloidogenesis. J. Mol. Biol. 336, 165–178.

[42] Elam, J.S., Taylor, A.B., Strange, R., Antonyuk, S., Doucette, P.A., Rodriguez, J.A., Hasnain, S.S., Hayward, L.J., Valentine, J.S., Yeates, T.O., et al. (2003) Amyloid-like filaments and water-filled nanotubes formed by SOD1 mutant proteins linked to familial ALS. Nat. Struct. Biol. 10, 461–467.

[43] Margittai, M. and Langen, R. (2008) Fibrils with parallel in-register structure constitute a major class of amyloid fibrils: molecular insights from electron paramagnetic resonance spectroscopy. Q. Rev. Biophys. 41, 265–297.

[44] Benzinger, T.L., Gregory, D.M., Burkoth, T.S., Miller-Auer, H., Lynn, D.G., Botto, R.E. and Meredith, S.C. (1998) Propagating structure of Alzheimer's beta-amyloid(10–35) is parallel beta-sheet with residues in exact register. Proc. Natl. Acad. Sci. USA 95, 13407–13412.

[45] Goldsbury, C., Baxa, U., Simon, M.N., Steven, A.C., Engel, A., Wall, J.S., Aebi, U. and Muller, S.A. (2011) Amyloid structure and assembly: insights from scanning transmission electron microscopy. J. Struct. Biol. 173, 1–13.

[46] Sharma, D., Shinchuk, L.M., Inouye, H., Wetzel, R. and Kirschner, D.A. (2005) Polyglutamine homopolymers having 8–45 residues form slablike beta-crystallite assemblies. Proteins 61, 398–411.

[47] Sachse, C., Fandrich, M. and Grigorieff, N. (2008) Paired beta-sheet structure of an Abeta(1–40) amyloid fibril revealed by electron microscopy. Proc. Natl. Acad. Sci. USA 105, 7462–7466.

[48] Kajava, A.V., Baxa, U. and Steven, A.C. (2010) Beta arcades: recurring motifs in naturally occurring and disease-related amyloid fibrils. FASEB J. 24, 1311–1319.

[49] Hennetin, J., Jullian, B., Steven, A.C. and Kajava, A.V. (2006) Standard conformations of beta-arches in beta-solenoid proteins. J. Mol. Biol. 358, 1094–1105.

[50] Luhrs, T., Ritter, C., Adrian, M., Riek-Loher, D., Bohrmann, B., Dobeli, H., Schubert, D. and Riek, R. (2005) 3D structure of Alzheimer's amyloid-beta(1–42) fibrils. Proc. Natl. Acad. Sci. USA 102, 17342–17347.

[51] Petkova, A.T., Yau, W.M. and Tycko, R. (2006) Experimental constraints on quaternary structure in Alzheimer's beta-amyloid fibrils. Biochemistry 45, 498–512.

[52] Kajava, A.V., Baxa, U., Wickner, R.B. and Steven, A.C. (2004) A model for Ure2p prion filaments and other amyloids: the parallel superpleated beta-structure. Proc. Natl. Acad. Sci. USA 101, 7885–7890.

[53] Wasmer, C., Lange, A., Van Melckebeke, H., Siemer, A.B., Riek, R. and Meier, B.H. (2008) Amyloid fibrils of the HET-s(218–289) prion form a beta solenoid with a triangular hydrophobic core. Science 319, 1523–1526.

[54] Saveanu, L., Fruci, D. and van Endert, P. (2002) Beyond the proteasome: trimming, degradation and generation of MHC class I ligands by auxiliary proteases. Mol. Immunol. 39, 203–215.

[55] Guo, Z. and Eisenberg, D. (2008) The structure of a fibril-forming sequence, NNQQNY, in the context of a globular fold. Protein Sci. 17, 1617–1623.

[56] Esteras-Chopo, A., Serrano, L. and Lopez de la Paz, M. (2005) The amyloid stretch hypothesis: recruiting proteins toward the dark side. Proc. Natl. Acad. Sci. USA 102, 16672–16677.