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 [11]. 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
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 amylooidogenic pathway have been proposed. Here we discuss these approaches and the principles behind them. New data about amyloidogenesis which may be critical for the understanding of the protein aggregation pathways are more complicated involving multiple intermediate stages for both natively structured and unstructured proteins [7–8].

2. Methods for the prediction of amyloid fibril formation

2.1. Methods that rely on individual amino acid aggregation propensity scores 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 amylooidogenic 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 (Table 1).

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 Aβ peptide interferes with the correct folding of the GFP and thus with the emission of fluorescence.

![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 likelihood 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].](image-url)

### Table 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Basic approach</th>
<th>Server/Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGGRESCAN</td>
<td>Composition of amino acids</td>
<td><a href="http://bioinf.uab.es/aggrescan/">http://bioinf.uab.es/aggrescan/</a></td>
</tr>
<tr>
<td>FoldAmyloid</td>
<td>Composition of amino acids</td>
<td>[FoldAmyloid, Garbuzynsky et al. (2010)]</td>
</tr>
<tr>
<td>Zygggregator</td>
<td>Amino acid aggregation propensities and properties of β-structural conformation</td>
<td><a href="http://www-vendruscolo.ch.cam.ac.uk/zygggregator.php">http://www-vendruscolo.ch.cam.ac.uk/zygggregator.php</a></td>
</tr>
<tr>
<td>TANGO</td>
<td>Properties of β-structural conformation</td>
<td><a href="http://tango.crg.es/">http://tango.crg.es/</a></td>
</tr>
<tr>
<td>PASTA</td>
<td>Pairwise interactions within the β-sheets</td>
<td>[β-sheets]</td>
</tr>
<tr>
<td>BetaScan</td>
<td>Pairwise interactions within the β-sheets</td>
<td>[β-sheets]</td>
</tr>
<tr>
<td>3D Profile method</td>
<td>Amyloid-like structures of short peptides</td>
<td>[β-sheets]</td>
</tr>
<tr>
<td>(ZipperDB)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waltz</td>
<td>Amyloid-like structures of short peptides</td>
<td><a href="http://walzt.switchlab.org/">http://walzt.switchlab.org/</a></td>
</tr>
<tr>
<td>NetCSSP</td>
<td>Conformational switches</td>
<td>[β-sheets]</td>
</tr>
<tr>
<td>AmylPred</td>
<td>Conformational switches</td>
<td>[β-sheets]</td>
</tr>
</tbody>
</table>

* 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 trifluoroethanol 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 (Cγ 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 hydrophobic 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 NNQNY 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 NNQNY 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 GNQQNY 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 amyloidogenicity 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 SVIIIE 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
They accurately predict short amyloid-forming peptides, and are two slightly different beta-strands alternating along the fibril axis. The Het-s cross-amyloid-amyloid peptide in its training set, or by some differences of Het-s prion domain. This can be explained by the absence of program does not find any amyloid-forming region within the prion region, the predictions are less credible. For example, Waltz predictions of amyloidogenic "hot spots" in fibril-forming regions of such tests are amyloid-forming regions (Fig. 3). The most frequently used examples for such 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.

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 beta-structural pattern representing alternation of polar and apolar residues, then to the analysis of side-chain interactions within a beta-sheet, before finally supplementing it with the analysis of side-chain packing between the beta-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 “beta-arcade” (Fig. 4) [48]. This fold represents a columnar structure produced by stacking of beta-strand-loop-beta-strand motifs called “beta-arches” [49]. Each beta-arcade has a double-layer structure in which 2 parallel in-register beta-sheets face each other. The side chains protruding from opposing beta-sheets form tight inter-digitated packing. Variations of beta-arch
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


