the FEBS Journal

MINIREVIEW

Protein aggregation and amyloid fibril formation prediction software from primary sequence: towards controlling the formation of bacterial inclusion bodies

Stavros J. Hamodrakas

Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Greece

Keywords
aggregation-prone amino acid stretches; amyloid-fibril forming regions; amyloidoses; functional amyloids; prediction software

Correspondence
S. J. Hamodrakas, Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 157 01, Greece
Fax: +30 210 727 4254
Tel: +30 210 727 4931
E-mail: shamodr@biol.uoa.gr
Website: http://biophysics.biol.uoa.gr

(Received 28 January 2011, revised 18 April 2011, accepted 3 May 2011)

doi:10.1111/j.1742-4658.2011.08164.x

Proteins might aggregate into ordered or amorphous structures, utilizing relatively short sequence stretches, usually organized in β-sheet-like assemblies. Here, we attempt to list all available software, developed during the last decade or so, for the prediction of such aggregation-prone stretches from protein primary structure, without distinguishing whether these algorithms predict amino acid sequences destined to be involved in ordered fibrillar amyloids or amorphous aggregates. The results of application of four of these programs on 23 proteins related to amyloidoses are compared. Because protein aggregation during protein production in bacterial cell factories has been shown to resemble amyloid formation, the algorithms might become useful tools to improve the solubility of recombinant proteins and for screening therapeutic approaches against amyloidoses under conditions that mimic physiologically relevant environments. One such example is given.

Background and aims

Normally soluble proteins or peptides convert under certain conditions into ordered fibrillar aggregates known as amyloid deposits. The fibrils which constitute these amyloid deposits are known as amyloid fibrils and the amyloid fibrils or their precursors appear to be related to several neurodegenerative diseases including Alzheimer's, Parkinson's, Huntington's, and also type II diabetes, prion diseases and many others, collectively called amyloidoses. Amyloidogenic proteins are quite diverse, with little similarity in sequence and native three-dimensional structure [1,2]. Additionally, several proteins and peptides not related to amyloidoses have the potential to form amyloid fibrils *in vitro*, suggesting that this ability for structural rearrangement and aggregation may be inherent to proteins [3].

All amyloid fibrils share the same cross-beta architecture and several functional proteins found in bacteria, fungi, insects and humans have also been found to adopt the same architecture under physiological conditions, as part of their functional role ([4–8] and references therein), despite the diversity of origin of their constituent proteins. Attention was given to these functional amyloids after our finding that silkmoth chorion is a natural protective amyloid [9,10].

Theoretical and experimental evidence indicates that short sequence stretches may be responsible for amyloid formation [11–13] and several methods have been published recently that attempt to predict aggregation-prone or amyloidogenic regions, based on various

Abbreviations
HST, hot-spot threshold; IB, inclusion body.

properties of proteins (TANGO [14], PASTA [15–21], AGGRESCAN [22], SALSA [23,24], ZYGGREGATOR [25]). We should perhaps mention here that some of the prediction methods try to distinguish amyloid fibril (ordered aggregates) prediction from amorphous aggregate prediction, providing also the relevant physical reasoning and influencing factors. However, we shall not attempt to distinguish between the two, obviously functionally different, cases hereinafter.

This minireview aims to provide (a) a short description of prediction algorithms and available software, (b) results of their use on a set of 23 well-known amyloidogenic proteins and (c) guidance towards applying this software as a useful tool for improving the solubility of recombinant proteins and for controlling the formation of bacterial inclusion bodies (IBs).

Short description of prediction algorithms and available software

Each method makes its own assumptions and implements its own predictors, which range from quite simplistic to quite complex. The ability to form β-strands is a predominant feature in most works, either in the form of statistical propensities or in the form of structural stability. Yoon and Welsh [15] searched for hidden beta-propensity in sequences, in other words regions that appear to be natively α-helical but have nonetheless the ability to form β-strands. Hamodrakas *et al*. [26] have similarly looked for 'conformational switches' in sequences – regions with a high predicted tendency to form both α-helices and β-strands – using the consensus secondary structure prediction program SECSTR [27] and Zibaee *et al*. [24] looked for β-contiguity, essentially a derivative of β-strand propensity based on the Chou and Fasman [28,29] set of secondary structure preference values. In a more structural approach, Thompson *et al*. [20] and Zhang *et al*. [23] identified regions computationally that can be stable β-strands in a stacked β-sheet crystal, similar to the one obtained from the peptides GNNQQNY and NNQQNY [30], known amyloidogenic regions from the yeast prion Sup35, while Trovato *et al*. [21] looked for regions with the ability to pair with each other and form β-sheets, with their program termed PASTA.

The formation of β-strands is not the only predictor though. Conchillo-Solé *et al*. [22] defined a set of aggregation propensities upon which they calculate the presence of aggregation 'hot-spots' in sequences. Their AGGRESCAN software is based on an aggregation-propensity scale for the 20 natural amino acids derived from *in vivo* experiments and on the assumption that short and specific sequence stretches are responsible for protein aggregation. In some more detail: relative experimental aggregation propensities, for each of the 20 natural amino acids, were initially derived from the intracellular aggregation of mutants, performing single-point mutations at the central position (19) of the central hydrophobic cluster comprising residues 17–21 of amyloid $A\beta_{1-42}$ Alzheimer's peptide ([22] and references therein). Then, a value is assigned to each residue of a given polypeptide sequence, which is taken from the table giving the relative experimental (*in vivo*) aggregation propensities of the 20 natural amino acids (a3v). Next, calculations are based on the sliding-window averaging technique: a sliding window of a given length is chosen and the program calculates the average of a3v values over the sliding window and assigns it to the central residue of the window (sliding-window lengths of 5, 7, 9 and 11 residues were trained against a database of 57 amyloidogenic proteins in which the location of aggregation hot-spots was known from experiment). This average is called a4v [22]. A plot of a4v over the entire sequence defines the aggregation profile of the polypeptide. The hot-spot threshold (HST) was defined as the average of the a3v of the 20 natural amino acids weighted by their frequencies in the SwissProt database [22]. A segment of the polypeptide sequence is considered as a putative aggregation hot-spot if there are five or more consecutive residues with an a4v larger than the HST and none of them is a proline (aggregation breaker). Several other parameters are calculated and reported, such as the average a4v in each hot-spot, the area of the aggregation profile above the HST, the total area (the HST being the zero axis) and the area above the HST of each profile peak identified as a hot-spot. These areas are calculated numerically using the trapezoidal rule [22]. The best predictions were obtained utilizing a sliding-window size of 5 for protein sequences with a length ≤ 75 residues, 7 for ≤ 175 residues, 9 for ≤ 300 residues and 11 for > 300 residues.

Galzitskaya *et al*. [18,19] also defined a novel intrinsic property for amino acid residues, the average expected packing density, which they found to be correlated to amyloidogenesis, while López de la Paz and Serrano [11] identified a sequence pattern that is involved in the formation of amyloid-like fibrils.

A variety of multi-parametric methods exist as well. Pawar *et al*. [17] and Tartaglia *et al*. [25] combine intrinsic properties of amino acid sequences to calculate aggregation propensities, while Tartaglia *et al*. [25] and Fernandez-Escamilla *et al*. [14] additionally include the effect of environmental variables in their equations for calculating aggregation rates.

We demonstrated that a consensus approach might be better suited for the task of predicting amyloidogenic stretches [26] and we developed a consensus algorithm, AMYLPRED [31], which combines some of these methods, representing most of the above-mentioned categories. These amyloidogenic determinants may often act as 'conformational switches' and thus they may play the role of templates initiating amyloid formation, through perhaps local structural rearrangements. We have shown that this tool successfully predicts nearly all experimentally verified amyloidogenic determinants in the sequences of proteins causing amyloidoses. Furthermore, AMYLPRED predicts on the sequences of amyloidogenic proteins several short potential amyloidogenic stretches that have not yet been experimentally verified [31]. A rather important finding by the application of this tool is that nearly all experimentally verified amyloidogenic determinants or aggregation-prone sequences and most predicted but not yet experimentally verified amyloidogenic regions reside on the surface of the crystallographically solved crystal structures of the relevant amyloidogenic proteins. This is shown in Figs 1 and 2 and, in more detail, in [31].

Several other methods have also been proposed recently that attempt to predict aggregation-prone or amyloidogenic regions in protein sequences. Clarke and Parker [32] combined a coarse-grained physicochemical protein model with a highly efficient Monte Carlo sampling technique to identify amyloidogenic sequences in four proteins for which respective experimental peptide fragmentation data exist. Peptide sequences were defined as amyloidogenic if the ensemble structure predicted for three interacting peptides described a stable and regular three-stranded β-sheet. Tian *et al.* [33] proposed a method, named PAFIG (prediction of amyloid fibril forming segments) based on support vector machines, to identify hexapeptides associated with amyloid fibrillar aggregates. PAFIG was used to predict the potential fibril-forming hexapeptides in all of the 64 000 000 possible hexapeptides. As a result, approximately 5.08% of hexapeptides showed a high aggregation propensity.

NETCSSP, an algorithm developed by Kim *et al.* [34], implements the latest version of the CSSP algorithm and provides a Flash-chart-based graphic interface that enables an interactive calculation of CSSP values for any user-selected regions in a given protein sequence. The CSSP algorithm (calculation of contact-dependent secondary structure propensity) is a sensitive method that detects non-native secondary structure propensities in protein primary structures. The method predicts local conformational changes, usually associated with protein aggregation and amyloid fibril formation, and can quantitatively estimate the mutational effect on changes in native or non-native secondary structural propensities in local sequences. This web tool provides pre-calculated non-native secondary structure propensities for over 1 400 000 fragments that are seven residues long, collected from Protein Data Bank (PDB) structures. They are searchable for chameleon subsequences (sequences that have the ability to form both α-helix and β-sheet) that can serve as the nucleating core of amyloid fibril formation.

The algorithm BETASCAN [35] calculates likelihood scores for potential β-strands and strand-pairs based

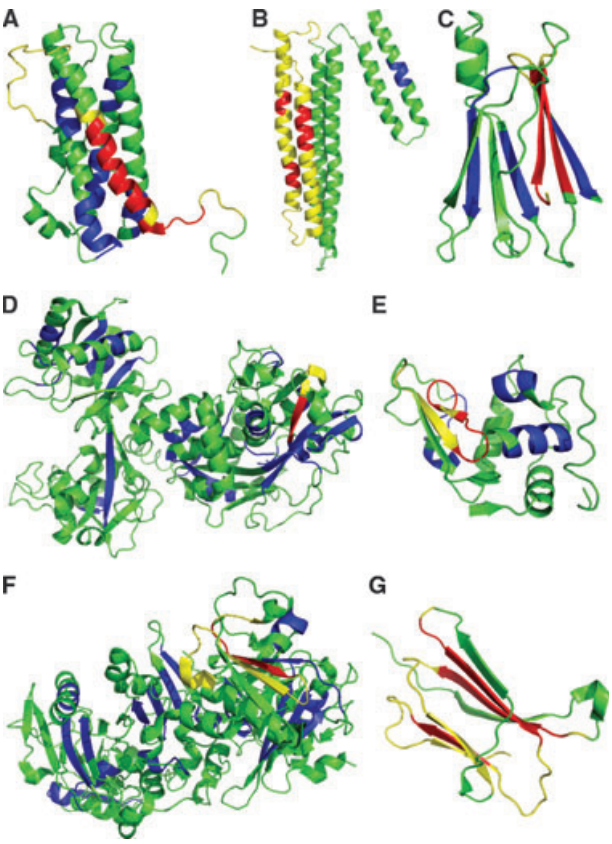

Fig. 1. Cartoon representations of seven proteins related to amyloidoses, with experimentally determined structures, which contain experimentally determined amyloidogenic regions. These seven protein models (see also Table S1), which were produced utilizing PYMOL [42] are (A) prolactin (PDB 1RWS); (B) apolipoprotein A-I (2A01); (C) transthyretin (1BMZ); (D) lactoferrin (1CB6); (E) lysozyme C (1LZ1); (F) gelsolin (2FGH); (G) β_2-microglobulin (1LDS). Experimentally determined amyloidogenic regions are shown in yellow. Theoretically predicted amyloidogenic regions, utilizing AMYLPRED [31], which coincide with experimentally determined regions are coloured red, whereas predicted amyloidogenic regions by AMYLPRED are shown in blue. The remainder of each protein is shown in green. Adapted from [31] with permission of BiomedCentral Ltd.

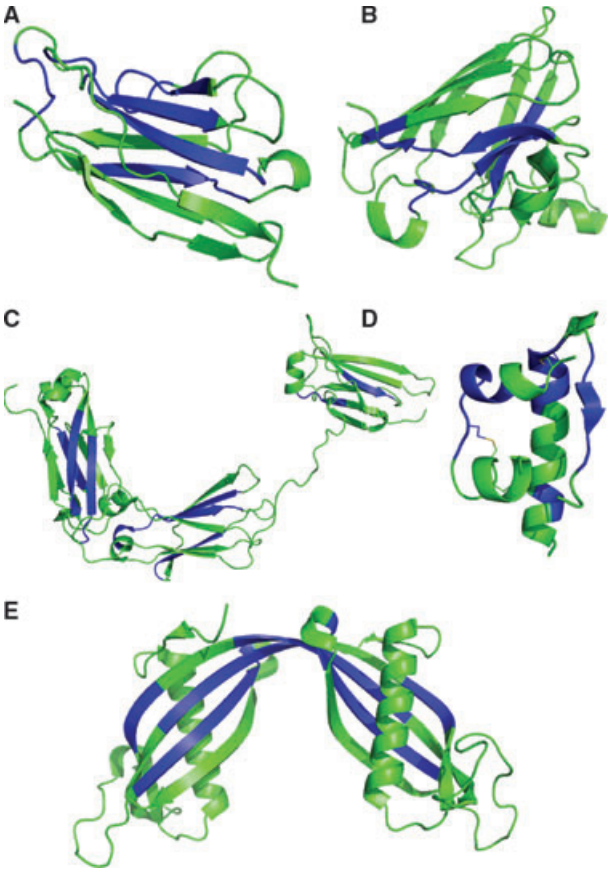

Fig. 2. Cartoon representations of five proteins related to amyloidoses, with experimentally determined structures which do not contain experimentally determined amyloidogenic regions. These five protein models (see also Table S1), which were produced utilizing PYMOL [42] are (A) immunoglobulin κ-4 light chain (PDB 1LVE); (B) superoxide dismutase (2C9V); (C) immunoglobulin G1 heavy chain (1HZH); (D) insulin (1ZNJ); (E) cystatin C (1R4C). Predicted amyloidogenic regions by AMYLPRED [31] are shown in blue (see also Table S1). The remainder of each protein is shown in green. Adapted from [31] with permission of BiomedCentral Ltd.

on correlations observed in parallel β-sheets. The program then determines the strands and pairs with the greatest local likelihood for all of the sequence's potential β-structures. BETASCAN suggests multiple alternative folding patterns and assigns relative *a priori* probabilities based solely on amino acid sequence, probability tables and pre-chosen parameters.

In the FOLDAMYLOID method [36], which is an extension of a method published by the same authors [18,19] based on the expected packing density of residues, two characteristics (expected probability of hydrogen bond formation and expected packing density of residues) are simultaneously used to detect amyloidogenic regions in a protein sequence. The authors claim that regions with high expected probability of formation of backbone–backbone hydrogen bonds as well as regions with high expected packing density are mostly responsible for the formation of amyloid fibrils. In more detail, the observed packing density for each amino acid residue was calculated from a database of 3769 three-dimensional protein structures (which have $< 25\%$ sequence identity between each other) obtained from the SCOP database [37], containing proteins which belong to the four main SCOP classes (classes a, b, c and d, which are all-α, all-β, α/β and α + β proteins, respectively) [36]. The observed packing density for each amino acid residue is defined as the number of amino acid residues in contact with the given residue (two residues are considered to be in contact if any pair of their non-hydrogen atoms is at a distance < 8 Å). Neighbouring residues in the amino acid sequence were excluded from this consideration. The calculated values (average observed packing density values for each amino acid residue, for the entire database) are used as a prototype scale for constructing a packing density profile for a certain protein sequence. Calculations are based on the sliding-window averaging technique. First, an expected value is assigned to each residue of the protein, equal to the average packing density value observed for this type of residue; then, the obtained values are averaged inside the window and the average is assigned to the central residue of the window. The 'smoothed' expected values for every position of the polypeptide chain provide the final profile, which is directly used for the prediction of amyloidogenic regions. On the 'smoothed' profile, a region is predicted as an amyloidogenic one if all its residues lie above a given cut-off (have numbers of expected contacts higher than the cut-off) and the size of the region is greater than or equal to the size of the sliding window used. Optimum values for the cut-off (threshold) and the sliding-window length are 21.4 contacts per residue and five residues, respectively [36].

The authors of FOLDAMYLOID also constructed two separate, different probability scales for the 20 amino acid residue types, acting separately either as donors or acceptors of backbone–backbone hydrogen bonds, calculated from the same database of 3769 proteins, utilizing the DSSP program [38]. The probability of backbone–backbone hydrogen bond formation, for each residue type, was calculated separately as the total number of hydrogen bonds this residue forms, acting either as donor or as acceptor, respectively, divided by the total number of residues of the same type in the database. The two, apparently, separate scales of probability of hydrogen bond formation are also used for constructing profiles over a protein sequence. Similarly as above, for the construction of

the profiles calculations are based on the sliding-window (five residues in length) averaging technique. First, an expected value is assigned to each residue of the protein, equal to the probability of backbone–backbone hydrogen bond formation observed for this type of residue; then, the obtained values are averaged inside the window and the average is assigned to the central residue of the window. The smoothed expected values for every position of the polypeptide chain provide the final profile, which is directly used for the prediction of amyloidogenic regions. On the smoothed profile, a region is predicted as an amyloidogenic one if all its residues lie above a given cut-off and the size of the region is greater than or equal to the size of the sliding window used. Optimum values for the cut-offs (thresholds), determined from receiver–operator characteristic curves, are 0.697 for the method based on the donor scale and 0.671 for the method based on the acceptor scale [36].

Thus, there are three scales which allow the prediction of amyloidogenic regions in a protein sequence (or rather, the ability of a peptide to be amyloidogenic): the scale of the packing density, and two scales of the probability of formation of backbone–backbone hydrogen bonds (assigned to donor and to acceptor residues, termed donor and acceptor scales, respectively). The authors, in order to take into consideration the above-mentioned scales simultaneously, have constructed several 'hybrid' scales by merging the individual scales with equal weights. The 'hybrid' scale, which includes all three scales (contacts + donors + acceptors) with equal weights, correctly predicts 80% of amyloidogenic peptides (115 of 144 peptides) and 72% of non-amyloidogenic ones (189 of 263 peptides), with a cut-off value of 0.062, from a database of 407 amyloidogenic and non-amyloidogenic peptides provided at the FOLDAMYLOID site (Table 1) [36].

WALTZ is a web-based tool that uses, mainly, a position-specific scoring matrix (PSSM) to determine amyloid-forming sequences [39]. The PSSM was built based on the experimental exploration of the sequence space of amyloid hexapeptides. According to its authors, WALTZ allows for identification and better distinction between amyloid sequences and amorphous β-sheet aggregates, and also allows for identification of amyloid-forming regions in functional amyloids. In more detail, the WALTZ algorithm was developed by combining specific sequence information with physicochemical as well as structural information.

The PSSM for amyloid propensity of WALTZ, was constructed from an experimentally defined training set comprising 116 'positive' (amyloid-forming) hexapeptides and 162 'negative' (non-forming) hexapeptides (http://waltz.switchlab.org/). This data set is an extension of the AmylHex database, which contains community-generated, experimentally verified amyloidogenic hexapeptides, consisting of 67 'positive' and 91 'negative' examples that have been used to benchmark novel prediction methods [20]. The additional examples/hexapeptides were identified experimentally by the authors of WALTZ [39]. The position-specific score for an amino acid was calculated as a standard log-odd score in a position-specific scoring matrix (the value for each amino acid at each position is the logarithm of the ratio of its frequency in the training set and the background database). As there is a positive and a negative set that both sample well the amino acid space over the motif (hexapeptide) positions, one profile was created for each set (positive and negative, respectively) and the score against the negative profile is subtracted (compliance with the negative set) from the score against the positive profile. Apparently, the sequence profile (S_{profile}) is the sum of position-specific scores for all amino acids in the hexapeptide.

Nineteen selected physical properties which best describe amyloid propensity enter the scoring function as a physical property term S_{physprop} consisting of the sum of the products of the amino acid frequency with the normalized property value of the respective amino acid for each position. Essentially, these properties can be assigned to three major groups representing beta, helical and solvation-related hydrophobicity propensities.

As the analysis of the hexapeptide experimental data sets (positive and negative) may impose sequence bias specific to the available data, the authors of WALTZ

Table 1. Protein aggregation and amyloid fibril formation prediction servers (URLs) and software.

Method	URL or software
TANGO [14]	http://tango.crg.es/
PASTA [21]	http://protein.cribi.unipd.it/pasta/
AGGRESCAN [22]	http://bioinf.uab.es/aggrescan/
PRE-AMYL [23]	Available at ftp://mdl.ipc.pku.edu.cn/pub/software/pre-amyl/
SALSA [24]	To obtain the software, contact Louise Serpell (l.c.serpell@sussex.ac.uk)
ZYGGREGATOR [25]	http://www-vendruscolo.ch.cam.ac.uk/zyggregator_test.php
AMYLPRED [31]	http://biophysics.biol.uoa.gr/AMYLPRED/
PAFIG [33]	Available at http://www.mobioinfor.cn/pafig/
NETCSSP [34]	http://cssp2.sookmyung.ac.kr/
BETASCAN [35]	http://groups.csail.mit.edu/cb/betascan/
FOLDAMYLOID [36]	http://antares.protres.ru/fold-amyloid/oga.cgi
WALTZ [39]	http://waltz.switchlab.org/

estimated the preference or non-preference of amino acids at the hexapeptide motif positions on a structural basis using the atomic force field FOLDX. The fibril crystal structure of the GNNQQNY peptide from Sup35 (PDB 1YJP) was first simplified to polyalanine. Then, all possible pair combinations of all 20 natural amino acids at all positions were generated and energy-optimized using FOLDX [40]. Energy estimates were calculated with FOLDX as the ΔG difference ($\Delta\Delta G$) to the reference polyalanine. To retrieve a position-specific pseudoenergy matrix for the prediction scoring function (and calculate S_{struct}), they averaged for each amino acid the energies for all its occurrences at a certain position in combination with all amino acids at other positions [39]. WALTZ combines sequence, physicochemical as well as structural information into a composite scoring function: $S_{total} = S_{profile} + S_{physprop} - 0.2S_{struct}$.

The authors of WALTZ claim that, when omitting the physicochemical property and structural descriptors in the prediction function, the sequence profile alone performs better than other prediction algorithms, although less than the complete scoring function. For more details, an interested reader should consult the original publication.

Table 1 provides a list of available servers and also sites for downloading available software developed for protein aggregation/amyloid fibril formation prediction.

Conclusions

Table S1 contains the results of the application of four (AMYLPRED [31], AGGRESCAN [22], WALTZ [39] and FOLDAMYLOID [23]) of these servers on 23 well-known amyloidogenic proteins [31]. Three of these methods, AGGRESCAN, FOLDAMYLOID and WALTZ, were analysed in more detail above. A comparison of 'aggregation-prone' stretches/amyloid fibril forming regions predicted by all programs with experimentally derived available information, given in Table S1, emphasizes what is believed to be true for 'aggregation-prone'/amyloid fibril forming regions prediction software: it appears that all methods tend to overpredict ([31] and references therein).

However, this might not actually be the case. We have undertaken a systematic study of synthesizing possible amyloidogenic peptide stretches, predicted by AMYLPRED [31], and testing them experimentally by transmission electron microscopy, X-ray diffraction, attenuated total reflection FTIR spectroscopy and Congo Red binding for their ability to form amyloid-like fibrils in water solutions. Out of 16 peptides synthesized so far, only one peptide was not found to be amyloidogenic (V. A. Iconomidou and S. J. Hamodrakas, unpublished data).

A number of amyloidogenic proteins related to human diseases that accumulate as insoluble IBs when synthesized recombinantly in bacteria have already been tested (Table 2 of [41] and references therein). Most of these proteins are included in Table S1. This suggests the exciting possibility of performing *in silico* (producing suitably designed variants, especially in the aggregation-prone/amyloidogenic regions) combined with *in vivo* (suitably engineered variants in bacterial cell factories) experiments for the detailed study of amyloid aggregation in various amyloidoses. Furthermore, the introduction of aggregation-disrupting amino acid substitutions in the aggregation-prone/amyloidogenic short sequence regions suggests the possibility of fine-tuning and controlling the solubility of proteins, synthesized by recombinant technology in bacterial cell factories. An example of how this can be accomplished, utilizing prediction algorithms as a first, guiding step, is given in Fig. 3. Furthermore, in Fig. 3, it is indicated how this procedure can be used for the synthesis of peptides, possible potent 'anti-amyloid' drugs, in association with recent findings.

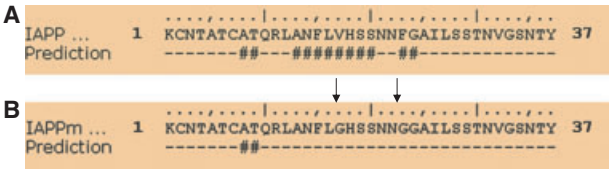

Fig. 3. A schematic example of how protein aggregation and amyloid fibril formation prediction software might be used for fine-tuning and control of protein solubility in bacterial IBs is shown. (A) The amino acid sequence of the 37 amino acid human islet amyloid polypeptide hormone (IAPP, amylin), a peptide forming amyloid-like fibrils, probably associated with a well-known amyloidosis, diabetes type II [1,2,4], is shown. Predicted amyloidogenic determinants by AMYLPRED [31] are marked by # below the sequence (see also Table S1 and references therein). This protein is known to accumulate as insoluble IBs when attempts are made for its synthesis, recombinantly, in bacteria ([41] and references therein). (B) Performing two single amino acid substitutions in the IAPP sequence (V17G and F23G, arrows), the AMYLPRED output suggests that the protein has 'lost' two, crucial, amyloidogenic determinants/'aggregation-prone' short peptides (compare with (A) above) and may therefore be soluble, not forming IBs. Thinking along similar lines may lead to the synthesis of peptides, potent 'anti-amyloid' drugs. Recently, a synthetic analogue of human amylin with proline (P) substitutions at positions 25, 28 and 29 (brand name Symlin or pramlintide), was approved for adult use in patients with diabetes mellitus types I and II, knowing that rat and mice amylin, which are not amyloidogenic, have similar substitutions at these positions [43]. Pramlintide (positively charged) is delivered as an acetate salt.

The testing of 'anti-amyloid' drugs that would prevent the formation of bacterial IBs in bacterial cell cultures should also not be excluded. These views are further discussed in detail by García-Fruitós *et al*. in this series, and also in [41].

Acknowledgements

We thank the University of Athens for financial support and the anonymous reviewers for useful criticism.

References

1 Chiti F & Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* **75**, 333–366.
2 Uversky VN & Fink AL (2004) Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim Biophys Acta* **1698**, 131–153.
3 Dobson CM (1999) Protein misfolding, evolution and disease. *Trends Biochem Sci* **24**, 329–332.
4 Harrison RS, Sharpe DC, Singh Y & Fairlie DP (2007) Amyloid peptides and proteins in review. *Rev Physiol Biochem Pharmacol* **159**, 1–77.
5 Fowler DM, Koulov AV, Balch WE & Kelly JW (2007) Functional amyloid – from bacteria to humans. *Trends Biochem Sci* **32**, 217–224.
6 Otzen D & Nielsen PH (2008) We find them here, we find them there: functional bacterial amyloid. *Cell Mol Life Sci* **65**, 910–927.
7 Fändrich M (2007) On the structural definitions of amyloid fibrils and other polypeptide aggregates. *Cell Mol Life Sci* **64**, 2066–2078.
8 Maji SK, Schubert D, Rivier C, Lee S, Rivier JE & Riek R (2008) Amyloid as a depot for the formulation of long-acting drugs. *PLoS Biol* **6**, 240–252.
9 Iconomidou VA, Vriend G & Hamodrakas SJ (2000) Amyloids protect the silkmoth oocyte and embryo. *FEBS Lett* **479**, 141–145.
10 Iconomidou VA & Hamodrakas SJ (2008) Natural protective amyloids. *Curr Prot Pept Sci* **9**, 291–309.
11 López de la Paz M & Serrano L (2004) Sequence determinants of amyloid fibril formation. *Proc Natl Acad Sci* **101**, 87–92.
12 Esteras-Chopo A, Serrano L & López de la Paz M (2005) The amyloid stretch hypothesis: recruiting proteins toward the dark side. *Proc Natl Acad Sci* **102**, 1639–1648.
13 Teng PK & Eisenberg D (2009) Short protein segments can drive a non-fibrilizing protein into the amyloid state. *Protein Eng Des Sel* **22**, 531–536.
14 Fernandez-Escamilla AM, Rousseaux F, Schymkowitz J & Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* **22**, 1302–1306.
15 Yoon S & Welsh WJ (2004) Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci* **13**, 2149–2160.
16 Tartaglia GG, Cavalli A, Pellarin A & Cafliesch A (2005) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci* **14**, 2723–2734.
17 Pawar AP, DuBay KF, Zurdo J, Chiti F, Vendruscolo M & Dobson CM (2005) Prediction of 'aggregation-prone' and 'aggregation-susceptible' regions in protein associated with neurodegenerative diseases. *J Mol Biol* **350**, 379–392.
18 Galzitskaya OV, Garbuzynskiy SG & Lobanov MV (2006) Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput Biol* **2**, 1639–1648.
19 Galzitskaya OV, Garbuzynskiy SO & Lobanov MY (2006) A search for amyloidogenic regions in protein chains. *Mol Biol* **40**, 821–828.
20 Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D & Eisenberg D (2006) The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci* **103**, 4074–4078.
21 Trovato A, Chiti F, Maritan A & Seno F (2006) Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PloS Comp Biol* **2**, 1608–1618.
22 Conchillo-Solé O, de Groot NS, Aviles FX, Vendrell J, Daura X & Ventura S (2007) AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinformatics* **8**, 65–81.
23 Zhang Z, Chen H & Lai L (2007) Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. *Bioinformatics* **23**, 2218–2225.
24 Zibaee S, Makin OS, Goedert M & Serpell LC (2007) A simple algorithm locates β-strands in the amyloid fibril core of α-synuclein, Aβ, and tau using the amino acid sequence alone. *Protein Sci* **16**, 906–918.
25 Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F & Vendruscolo M (2008) Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* **380**, 425–436.
26 Hamodrakas SJ, Liappa C & Iconomidou VA (2007) Consensus prediction of amyloidogenic determinants in amyloid-forming proteins. *Int J Biol Macromol* **41**, 295–300.
27 Hamodrakas SJ (1988) A protein secondary structure prediction scheme for the IBM PC and compatibles. *Comput Appl Biosci* **4**, 473–477.
28 Chou PY & Fasman GD (1974) Conformational parameters for amino acids in α-helical, β-sheet, and

random coil regions calculated from proteins. *Biochemistry* **13**, 211–222.
29 Chou PY & Fasman GD (1974) Prediction of protein conformation. *Biochemistry* **13**, 222–245.
30 Nelson R, Sawaya MR, Balbirnie M, Madsen AØ, Riekel C, Grothe R & Eisenberg D (2005) Structure of the cross-β spine of amyloid-like fibrils. *Nature* **435**, 773–778.
31 Frousios KK, Iconomidou VA, Karletidi CM & Hamodrakas SJ (2009) Amyloidogenic determinants are usually not buried. *BMC Struct Biol* **9**, 44.
32 Clarke OJ & Parker MJ (2009) Identification of amyloidogenic peptide sequences using a coarse-grained physicochemical model. *J Comp Chem* **30**, 621–630.
33 Tian J, Wu N, Guo J & Fan Y (2009) Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC Bioinformatics* **10**(Suppl I), S45.
34 Kim S, Choi J, Lee SJ, Welsh WJ & Yoon S (2009) NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation. *Nucl Acids Res* **37**, W469–W473.
35 Bryan AW Jr, Menke M, Cowen LJ, Lindquist SL & Berger B (2009) BETASCAN: probable β-amyloids identified by pairwise probabilistic analysis. *PLoS Comput Biol* **5**, 1–11.
36 Garbuzynskiy SO, Lobanov MY & Galzitskaya OV (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* **26**, 326–332.
37 Murzin AG, Brenner SE, Hubbard T & Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536–540.
38 Kabsch W & Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637.
39 Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L *et al*. (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods* **7**, 237–245.
40 Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F & Serrano L (2005) The FoldX web server: an online force field. *Nucleic Acids Res* **33**, W382–388.
41 de Groot NS, Sabate R & Ventura S (2009) Amyloids in bacterial inclusion bodies. *Trends Biochem Sci* **34**, 408–416.
42 Delano WL (2005) *The PyMOL Molecular Graphics System*. DeLano Scientific LLC, San Francisco, CA.
43 Jones MC (2007) Therapies for diabetes: pramlintide and exenatide. *Am Fam Physician* **75**, 1831–1835.

Supporting information

The following supplementary material is available:
Table S1. Prediction of amyloidogenic regions or 'aggregation-prone' stretches, for 23 amyloidogenic proteins [31] by four methods, for comparison.

This supplementary material can be found in the online version of this article.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.