Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Structural Biology

Computational methods to predict protein aggregation

Susanna Navarro and Salvador Ventura

Abstract

In most cases, protein aggregation stems from the establishment of non-native intermolecular contacts. The formation of insoluble protein aggregates is associated with many human diseases and is a major bottleneck for the industrial production of protein-based therapeutics. Strikingly, fibrillar aggregates are naturally exploited for structural scaffolding or to generate molecular switches and can be artificially engineered to build up multi-functional nanomaterials. Thus, there is a high interest in rationalizing and forecasting protein aggregation. Here, we review the available computational toolbox to predict protein aggregation propensities, identify sequential or structural aggregation-prone regions, evaluate the impact of mutations on aggregation or recognize prion-like domains. We discuss the strengths and limitations of these algorithms and how they can evolve in the next future.

Addresses

Institut de Biotecnologia I de Biomedicina, Departament de Bioquímica I Biologia Molecular, Universitat Autònoma de Barcelona, 08193, Bellaterra, Barcelona, Spain

Corresponding author: Ventura, Salvador (salvador.ventura@uab.es)

Current Opinion in Structural Biology 2022, **73**:102343

This review comes from a themed issue on **Folding and Binding**

Edited by **Fabrizio Chiti** and **Anna Sablina**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.sbi.2022.102343

Introduction

Proteins are the most abundant biomolecules in living systems playing a pivotal role in most biological processes. The establishment of functional native intra- and interchain interactions is a fundamental aspect of protein biology, governing protein folding, binding, and activity. In the crowded environment of living cells, proteins transit through different conformational states in the search for a free energy minimum, which can correspond to a monomeric state or a wide variety of assemblies [1], depending on the polypeptide and cellular conditions.

Intracellular assemblies present different natures ranging from biomolecular condensates, which are multi-component, dynamic, and reversible assemblages, often formed via liquid—liquid phase separation, to irreversible protein aggregates, like amyloids [2]. During protein assembly, native contacts might preserve the surface conformation that interacts to drive the supramolecular structure, in a process known as agglomeration [3,4]; however, more often, protomers undergo partial or global unfolding, and native contacts are replaced by non-native intermolecular interactions leading to the formation of non-structured amorphous aggregates or highly ordered amyloid fibrils, characterized by a cross-β conformation formed by the repetitive stacking of β-strand monomers perpendicularly to the fibril axis [5].

The formation of insoluble aggregates is often associated with a loss of protein activity and/or a gain in toxicity, and is intimately related to a broad array of diseases and aging, including neurodegenerative disorders, such as Alzheimer's and Parkinson's [6], and non-neuronal localized or systemic diseases [7].

Protein aggregation constitutes a major bottleneck in producing protein-based therapeutics and biotechnological products. During manufacturing, proteins are exposed to unnatural stresses and formulated at concentrations far from their physiological abundance, which exacerbate their aggregation potential. Aggregation directly reduces production yields, final product activities and might triggerunpredictable immune responses in patients [8].

Although amyloids are best known for their deleterious effects, they have also evolved to play a pivotal role in biological functions that cannot be attained by globular proteins, from bacteria to humans [9—12]. These activities, associated with their common cross-β architecture, have become a source of inspiration for designing amyloid-based materials with a broad range of applications in nanotechnology [13—15].

Extensive research has been devoted to understanding the molecular determinants behind uncontrolled and aberrant protein oligomerization. However, empirical methods are costly, time-consuming, and limited by the availability of proper protein models. In this context, computational methods have emerged as powerful complementary tools for studying aggregation at the individual protein and proteome scales [16]. Understanding and anticipating protein aggregation has a translational impact in five main areas: pathology, evolution, protein manufacturing, development of novel nanomaterials, and unveiling the physiological role played by functional amyloids.

Here, we provide a compendium of computational methods available to predict aggregation propensity, identify aggregation-prone regions, evaluate the impact of mutations on aggregation and solubility, and for prion-like proteins identification. We also discuss how the incorporation of yet unattended parameters might improve such predictions.

Prediction of aggregation propensities in protein sequences

Intrinsic determinants behind protein aggregation

Many of the molecular determinants governing protein aggregation are naturally imprinted in the primary sequence and depend on intrinsic properties of the polypeptide chain, including the amino acid composition or residues patterning. Hydrophobicity is a major driving force in the establishment of intrachain and interchain contacts leading to aggregation, whereas the protein local and net charges often play an opposing role, promoting electrostatic repulsion between individual residues or molecules. Most amyloid aggregates adopt a cross-β fold, implying that aggregation is favoured by amino acids with high β-sheet propensity and counteracted by β-sheet breaking residues, such as Pro and Gly. However, the aggregation propensity of a protein is not equally distributed along the sequence and tends to concentrate in short linear stretches, known as hot spots or aggregation-prone regions (APRs). These regions are sufficient and necessary to nucleate protein aggregation. They are often enriched in hydrophobic residues and, in globular proteins, they generally map at the hydrophobic core. However, exposed APRs involved in catalysis and binding have also been described. The presence of APRs is a requirement for protein stability and it seems that positive selection for an amyloid structure might be behind the emergence of globular folds [17], their potency being modulated by the presence of adjacent residues of low aggregation propensity, known as gatekeepers. In addition, aggregation propensities are balanced by extrinsic factors, such as pH, ionic strength, temperature, or even gravity [18]. This dependence on the environment is often behind the frequently observed amyloid polymorphism [19,20].

Our increasing understanding of the intrinsic and extrinsic factors governing protein aggregation and their interplay has crystallized in the development of *in silico* methods to predict this phenomenon departing both from sequences and structures.

Exploiting sequential features to predict protein aggregation

Protein sequences encode the information for folding into native globular states but also to remain partially or fully unstructured in the case of intrinsically disordered proteins (IDPs). This linear code overlaps with the one accounting for the formation of the cross-β fold common to amyloids and thus first-generation aggregation predictors were aimed to read aggregation propensities from the primary sequence.

To date, more than 20 prediction algorithms exist for the recognition of linear APRs in polypeptides using sundry metrics to derive aggregation propensity, solubility, and thermodynamic stability. All of them use protein sequences as an input, but they differ on the strategies employed for APRs identification and quantification, relying on either phenomenologically or theoretically derived parameters. A first set of algorithms exploits scales of aggregation propensity for amino acids determined experimentally, exemplified by AGGRESCAN [21] or Zyggregator [22]. A second set of methods combines different key features of APRs, such as specific physicochemical properties of amino acids, their distribution along the sequence or the feasibility to adopt a β-sheet conformation, employing particular functions to weigh each of these parameters. They include Waltz [23], SALSA [24], PAGE [25], Tango [26], FoldAmyloid [27] and PASTA [28], among others. For a description of first-generation methods see Table 1.

Given the intricacy of protein aggregation reactions, averaging the predictions provided by algorithms relying on distinct strategies seems reasonable. This is the concept behind AmylPred2, which combines the outputs of eleven different algorithms, identifying an APR when the stretch is predicted by at least n/2 of the methods [29]. In METAMYL, scores are weighted using a logistic regression model arising from the combination of four predictive methods [30] (Table 1).

Artificial intelligence to evaluate sequential aggregation propensity

The need to train and perfect linear predictors boosted the generation of datasets and databases compiling experimentally validated aggregating regions in peptides and proteins [31,32]. Most of these repositories provide a binary classification regarding the capacity of the proteins or segments to assemble or not into amyloid fibrils. Thus, they constitute the perfect substrate for machine learning classifiers, which can automatically uncover unrelated features of polypeptide chains. The effectiveness of these algorithms heavily depends on the accurate annotation of the reference training data, although a certain degree of misannotated instances still allows robust predictions, which suggests that, indeed,

Table 1

Computational methods to predict protein aggregation.

Algorithm	Characteristics	URL and Ref
Sequence-based methods		
AGGRESCAN	Intracellular aggregation propensity scale for each of the 20 amino acids derived after introducing single point mutations on Aβ42 peptide	http://bioinf.uab.es/aggrescan [21]
Zyggregator	Relative propensities for aggregation based on hydrophobicity, secondary structure propensity, hydrophobic/hydrophilic patterning, net charge, the presence of gatekeepers and the influence of structural protection	[22]
Waltz	Statistical method that exploits Position-Specific Substitution Matrices (PSSM) obtained from amyloidogenic hexapeptides, combined with physicochemical properties of β-amyloids and conformational features of amyloid backbone structures	https://waltz.switchlab.org/ [23]
SALSA	β-strand propensity is calculated from a β-strand contiguity score based on Chou and Fasman's secondary structure propensity scale and applied using a sliding window	http://amypdb.genouest.org/e107_plugins/amypdb_aggregation/db_prediction_salsa.php [24]
PAGE	Prediction of parallel or anti-parallel β-sheet organization in fibrils and aggregation rates based on physicochemical properties and computational design of β-aggregating peptide sequences	[25]
TANGO	Statistical mechanics-based method that calculates the partition function of the phase-space considering conformational states and energy terms, together with physico-chemical and protein stability parameters	http://tango.crg.es [26]
FoldAmyloid	Predicts amyloidogenic regions according to the expected probability of formation of backbone−backbone hydrogen bonds and the expected packing Density	(http://bioinfo.protres.ru/fold-amyloid/ [27]
PASTA 2.0	Estimation of β-strand inter-molecular pairing probability between polypeptide segments, based on experimentally-resolved β-sheets structures, and a statistical energy function to determine fibril formation	http://protein.bio.unipd.it/pasta2/ [28]
SecStr	Combines the prediction of amyloidogenic regions with the identification of 'conformational switches'	http://biophysics.biol.uoa.gr [82]
ArchCandy	Detection of amyloidogenic regions based on the propensity to form parallel in register stacking of β-arches (β-strand-loop-β-strand motif)	https://bioinfo.crbm.cnrs.fr/index.php?route=tools&tool=7 [83]
BetaSerpentine	Reconstruction and ranking of β-serpentine arrangements of adjacent β-arches predicted by ArchCandy	github.com/stanislavspbgu/BetaSerpentine [84]
BETASCAN	Prediction of β-strands and strand pairs in a sequence based on pairwise probabilities for each pair of residues to form hydrogen bonds in amphiphilic β-sheets	http://betascan.csail.mit.edu [85]
AmyloidMutants	Quantifies the energetic effects of sequence mutation on fibril conformation and stability using a potential energy scoring function derived from the frequency of specific residue/residue interactions in PDB protein structures	http://amyloid.csail.mit.edu/ [86]
STITCHER	Predicts the formation of strand-pairs into complete β-sheets based on a free-energy model accounting for amino acid sidechain stacking contributions, entropic estimation, and steric restrictions for amyloidal parallel β-sheet formation. A dynamic program returns the top scored structures	http://stitcher.csail.mit.edu [87]
GAP (Aggregation Proneness)	Evaluation of residue pair propensities to occur at adjacent or alternate positions in globular proteins, and calculation of thermodynamic energy potentials. This tool can predict whether APRs would form amorphous β-aggregates or amyloid fibrils.	http://www.iitm.ac.in/bioinfo/GAP/ [88]
3D Profile	Profile generated by mutating the side chains in the cross-β spine of NNQQNY crystal structure and employing ROSETTADESIGN to evaluate the sequence energetic fit	www.rosettacommons.org [89]
Machine-learning methods		
ANuPP (Aggregation Nucleation Prediction in Peptides and Proteins)	Ensemble-classifier that identifies potential aggregation-nucleating regions trained on an experimental dataset of amyloidogenic and nonamyloidogenic hexapeptides	https://web.iitm.ac.in/bioinfo2/ANuPP/ [34]

(*continued on next page*)

Table 1 (*continued*)

Algorithm	Characteristics	URL and Ref
PATH (Prediction of Amyloidogenicity by Threading)	The input sequence is threated into templates of different structural amyloid classes and the model with the lowest energy value score according to PyRosetta is used as an input for machine learning classifiers	Scripts available: https://github.com/KubaWojciechowski/PATH [35]
NetCSSP	Detects non-native secondary structure propensities based on the calculation of contact-dependent secondary structure propensity (CSSP), and the search for chameleon subsequences using a PDB structures collection	http://cssp2.sookmyung.ac.kr/ [36]
FiSH amyloid	Machine learning method trained to recognise and classify amyloidogenic segments based on position-specific amino acid co-occurrence patterns in protein sequences	http://www.comprec.pwr.wroc.pl/COMPREC_home_page.html [37]
RF Amyloid	A random forest protein classifier based on composition and physicochemical features from protein sequences	http://server.malab.cn/RFAmyloid/ [38]
Budapest	Linear Support Vector Machine (SVM)-based predictor for hexapeptides trained and tested with the experimental hexapeptide Waltz dataset.	https://pitgroup.org/bap/ [39]
CORDAX	Logistic regression approach that detects APRs and predicts the structural topology and architecture of the fibril core. It exploits a curated amyloid template structural database generated from the WALTZ-DB 2.0 repository	https://cordax.switchlab.org [40]
AgMata	Machine learning-based classifier not trained on aggregation data. It uses manually selected parameters accounting for predicted secondary structure propensities, side-chain and backbone dynamics, and a β-pairing energy function	https://bitbucket.org/bio2byte/agmata [90]
Pre-Amyl-MLP	Machine learning-based prediction using a multilayer perceptron-based classification that exploit a selected combination of amyloid associated features	http://106.12.83.135:8080/amyWeb_Release/index.jsp [91]
AbAmyloid	Random Forest classifier adopted to evaluate amino acid composition, dipeptide composition and physicochemical properties	http://iclab.life.nctu.edu.tw/abamyloid [92]
Pafig (Prediction of amyloid fibril-forming segments)	Identification of amyloid fibril-prone hexapeptides with a scale derived from machine supervised learning of >500 physicochemical properties. Suitable for large-scale analysis	[93]
APPNN (Amyloidogeniciy Propensity Prediction Neural Network)	Machine learning approach that analyses features correlated with self-assembly of peptides and proteins into amyloids: frequency of β-sheet, isoelectric point, atom-based hydrophobic moment, helix termination parameters and ΔG° values for peptides extrapolated in 0 M urea	http://cran.r-project.org/web/packages/appnn/index.html) [94]
Amylogram	N-grams and random forest machine learning classifiers that recognise sequential patterns in the amyloids considering hydrophobicity, tendency to form β-sheets, and low flexibility of amino acid residues	http://smorfland.uni.wroc.pl/shiny/AmyloGram/ [95]
Consensus methods		
Amylpred 2	It combines 11 individual methods (Aggrescan, AmyloidMutants, Amyloidogenic, Pattern, Average Packing Density, Beta-strand contiguity, Hexapeptide Conformational Energy, NetCSSP, Pafig, SecStr, Tango and Waltz) to identify amyloid-forming regions. Consensus between the output from at least n/2 of n selected algorithms	http://aias.biol.uoa.gr/AMYLPRED2/ [29]
MetAmyl	Uses a logistic regression model after combining and weighting the output of 4 popular predictors (SALSA, PAFIG, Waltz and FoldAmyloid). It can handle large sequence datasets	http://metamyl.genouest.org/ [30]
3D structure-based methods		
Solubis	Identifies mutations that reduce protein aggregation employing the statistical thermodynamics algorithm TANGO to evaluate the sequence intrinsic aggregation propensity. This is projected onto high-resolution 3D structures, and the thermodynamic stability of the variant assessed using FoldX	http://solubis.switchlab.org [55]
Aggscore	Trained on a dataset of 31 adnectin proteins with varying aggregation propensities. It uses the distribution of hydrophobic and electrostatic patches on the surface together with their intensity and orientation to implement and aggregation propensity scoring function	Schrödinger's BioLuminate Suite as of software release 2018–1 [56]

Table 1 (*continued*)

Algorithm	Characteristics	URL and Ref
SAP (Spatial aggregation propensity)	Based on full antibody atomistic simulations, it measures the effective dynamically exposed hydrophobicity of a certain patch on the protein surface to identify aggregation prone regions. It implements a mutation option on predicted regions. SAP has been applied either in high-throughput developability screening of therapeutic protein candidates or to enhance stability at later stages of manufacturing	[57]
AGGRESCAN3D 2.0	The method integrates the 3D-structural information of PDBs and evaluates the contribution of solvent-exposed APRs using the amino acid aggregation scale from Aggrescan method. Predictions can be run in static and dynamic modes. An extended option allows the evaluation of stable mutations using FoldX method	http://biocomp.chem.uw.edu.pl/A3D2/ [58,59]
Camsol	Aimed to design protein variants with enhanced solubility by the calculation of the intrinsic solubility profile with a structural correction accounting for the residues' structural environment and for their solvent exposure. The structurally corrected solubility profile is used to identify the most suitable solubilizing mutations	http://www-vendruscolo.ch.cam.ac.uk/camsolmethod.html [60]

these programs can be used to evaluate the quality of the experimental data [33].

Several protein aggregation predictors employ machine learning classifiers to recognise sequence-specific features and position-specific patterns, or they use energy functions of cross-β pairings; they include ANuPP [34], Amylogram [35], netCSSP [36], FISH amyloid [37], RF amyloid [38], and Budapest [39] (Table 1). The most recently developed PATH [35] and CORDAX [40] algorithms exploit crystallographic structures of amyloid steric zippers [41] to combine threading and machine learning approaches, providing insights into the forces that govern the formation of amyloid assemblies by short peptides (Table 1). This class of algorithms has succeeded in protein redesign [42], solubility forecasting [43], or ranking aggregation rates [44]. Further improvements would require connecting their outputs with understandable physical, chemical, or structural parameters to feedback the predictive model.

Prion-like proteins can be identified from their sequences

Prions and prion-like proteins are a particular subset of amyloids that can interconvert between a soluble conformer and a cross-β structure. Both states can be functional, and the conversion to the amyloid state can either switch off the initial function or switch on a new activity. The aggregated state is often transmissible, templating homologous polypeptides conversion. They differ from classical amyloids since, in most cases, this activity is located in disordered regions of low sequence complexity enriched in polar residues, like glutamine and asparagine, and depleted in hydrophobic amino acids, which are known as prion or prion-like domains (PrDs, PrLDs). Their enrichment in hydrophilic residues contrasts with the eminent hydrophobic nature of classical APRs and precludes identifying these aggregating regions with pre-existent algorithms. Thus, the development of new tools was required (Table 2). A first class of methods compares target proteins' amino acid composition with the one of PrDs of similar length in bona fide yeast prions using different metrics and scanning procedures. They include DIANA [45], LPS [46], PAPA [47], PLAAC [48], PrionScan [49], and the machine learning method pRANK [50]. All these methods consider that the transition towards the aggregated state is mediated by a large number of weak interactions distributed along the PrD, according to the view that the PrD composition and not its specific sequence is relevant for prion conversion. A second class of algorithms, like pWALTZ [51] and PrionW [52], relies on identifying and ranking soft-amyloid cores within PrD, defined as short linear stretches of low complexity with a moderate but significant amyloid propensity that nucleates the aggregation reaction. The predictions of the two kinds of methods are complementary, and their joint application has allowed the identification of thousands of new prion-like candidates in proteomes belonging to all kingdoms of life [53].

Prediction of aggregation propensities in globular protein structures

Overall, linear algorithms have demonstrated to be fast and cost-effective tools to predict sequences aggregation propensities, with a remarkable overlay between predicted and experimentally validated APRs [54]. They are of particular interest when dealing with short

Table 2

Computational methods to identify proteins bearing prion domains.

Method	Description	URL and Ref
Composition-based predictors		
DIANA (Defined Interval Amino acid Numerating Algorithm)	Inspired by the length and composition of Sup35p and Ure2p yeast prions and the length of pathogenic polyQ expansions. The algorithm searches for consecutive 80 residue-long sequence and retrieves the most Q/N-rich stretch containing at least 30 Q and/or N residues	[45]
LPS (Lowest-Probability Subsequences)	Identifies compositional-biased regions by defining the lowest-probability sub-sequences (LPSs) for a given amino acid composition in a defined proteomic context	http://libaio.biol.mcgill.ca/lps-annotate.html [46]
PAPA (Prion Aggregation Prediction Algorithm)	Combination of analysis of disordered regions using FoldIndex and calculation of the prion propensity of each amino acid using a scale obtained experimentally from a randomly mutated Sup35p segment	http://combi.cs.colostate.edu/supplements/papa/ [47]
PLAAC (Prion-like amino acid composition)	Implements a Hidden Markov model derived from 28 characterised yeast prion proteins. The program employs the HMM-derived residue log-likelihood for every position, and calculates the probability of each amino acid to be part of a PrLD, considering the background frequencies in the selected proteome	http://plaac.wi.mit.edu/ [48]
PrionScan	Bimodal method incorporating an open-source database of prion predictions for all the proteins in UniProt KB and a sequence analysis tool to test prionogenicity of user's protein sequences, relying on the amino acid frequencies of validated prion and non-prion of similar composition	http://webapps.bifi.es/prionscan [49]
pRank	Employs supervised multiple-instance learning to solve the problem of inaccurately annotated data. Trained on top of 22 known Q/N-rich yeast prions against the rest of the yeast proteome, it ranks and classifies prion sequences	http://faculty.pieas.edu.pk/fayyaz/prank.html [50]
Soft Amyloid Core-based predictors		
pWaltz	Employs the amyloid propensity scoring matrix of the Waltz linear predictor and scans for soft-amyloid cores (SACs) by implementing a 21-residue window and calculating its average amyloid load. It identifies SACs as the highest scoring stretches above a calibrated threshold	http://bioinf.uab.es/pWALTZ/ [23,51]
PrionW	Intrinsically disordered segments of at least 80 residues are identified with FoldIndex and their Q/N content computed. Q/N enriched sequences are then scored using pWaltz. The Q/N and pWaltz identification thresholds can be adjusted for each specific protein dataset.	http://bioinf.uab.cat/prionw/ [52]

peptides, IDPs, or fluctuating protein regions, where APRs are permanently or transiently exposed to solvent. However, they mistakenly overestimate the aggregation propensity of globular proteins, and native oligomers, for which the structural context strongly modulates the contribution of sequential determinants. Overprediction occurs because, in native states, APRs are usually embedded within the hydrophobic core, being sheltered from the solvent and thus not contributing significantly to the aggregation propensity. On the contrary, the clustering of sequentially distant hydrophobic residues at protein surfaces may generate structural aggregation-prone regions (STAP) in the native state, to which linear predictors remain blind.

The above-described limitations, together with the evidence that many of the biotechnologically relevant proteins are globular (e.g., antibodies or enzymes for industrial and therapeutic applications), have prompted the development of a generation of structure-based prediction methods to address the location of solvent-accessible aggregation hot-spots, like Solubis [55] and Aggscore [56], or the identification of spatially close aggregation-prone regions, such as SAP [57], AGGRESCAN3D 2.0 [58,59], and CamSol [60] (Table 1). Globular proteins are flexible in solution, and this impacts the degree of exposure of their STAPs. Accordingly, molecular dynamics have been incorporated in SAP and AGGRESCAN3D 2.0 predictions to simulate structural fluctuations under native conditions [58,59], using full atomistic calculations and the coarse-grained CABS-flex approach [61], respectively. Noteworthy, several of these tools allow for predicting structural aggregation propensity and thermodynamic stability upon automatically introducing multiple mutations, which is extremely useful for selecting and producing highly soluble and stable protein variants [62].

The predictions of structure-based algorithms are accurate when a high-resolution experimental structure is available, being less precise when working with molecular models, and this has restricted their use to the subset of structures deposited in the protein data bank. However, the apparition of programs like AlphaFold2 [63] and RoseTTaFold [64] providing high fidelity structural models for natural or mutated sequences should bypass this limitation. We foresee a future in which structure and aggregation prediction algorithms would be sequentially implemented in protein production pipelines, making it fast and economically affordable to redesign therapeutic products with increased solubility [65].

Finally, later advances in solid-state NMR (ssNMR) [66] and cryo-EM [67] are providing an increasing repertoire of fibrillar structures of pathological and functional amyloids, which can be used to train structure-based methods in order to assist the redesign of amyloid-like structures for therapeutic and nanotechnology applications [65].

Conclusions and perspectives

In silico aggregation predictors have contributed to guide and assist experimental endeavours in elucidating the molecular mechanisms underlying protein aggregation-related diseases [68]. They have also boosted the design of engineered protein variants with improved solubility and stability [69,70], saving time and costs in therapeutic proteins production pipelines. Each of the discussed approaches has its own pros and cons (Table 3), and users should bear in mind the specific problem they want to address. This is because they capture different aspects of protein aggregation, that might or might not be relevant for the application of interest.

Protein aggregation reactions are complex processes in which, besides the primary sequence, multiple parameters might impact the stability, structure, cooperativity, solubility, kinetics, and dynamics of polypeptides. Thus, the binary classification into aggregating and non-aggregating proteins or protein regions provided by predictors might be accurate only in conditions close to those in which the initial experiments or calculations that feed the training data sets were performed.

The influence of factors intrinsic to the sequence is evident because we need different programs to predict archetypical amyloids and prion-like proteins. Indeed, two recent works [40,71] indicate that the amyloid sequence space is much larger than previously thought, including highly soluble sequences with low aliphatic content and/or high net charge. At the core of most of

Table 3

Pros and cons of the different classes of computational methods to predict protein aggregation.

Method	Pros	Cons
Sequence Based	- Low computational demand - User-friendly - Many webservers available - Fast analysis of complete proteomes - Perform well on small peptides and IDPs - Consensus methods available	- Dismiss the 3D-structural context - Disregard the impact of mutations on protein stability - Do not account for protein micro-environment
Machine learning	- Produce outputs from apparently unrelated protein sequence features - Attain high accuracy - Implemented in webservers - Suitable for fast and wide-range analyses	- Accurate annotation of the reference experimental training data is required - Do not calculate full-length proteins average propensities - Disconnection between the output and understandable physicochemical interpretations
Structure Based	- Predict on the native, functional state of proteins - Consider the quaternary structure - Analyse surface-exposed hydrophobic and hydrophilic patches - Incorporate protein dynamics calculations - Accurate in predicting globular proteins solubility - Consider the impact of mutations on protein stability - Allow automated design of more soluble protein variants	- Dependence on high-resolution structures availability - Not suitable for IDPs - High computational cost - Difficulty to perform proteome-wide analysis - Protein chemistry/engineering knowledge required - Do not account for protein micro-environment

the discussed algorithms is the concept that low solubility and aggregation propensity are interchangeably properties, and this should be revisited to fish sequences in this uncharted amyloid territory. In addition, it should not be overlooked that aggregation is not uniquely dependent on APRs since flexible segments often located remote from these stretches may act as conformational switches, masking APRs and thus modulating the fibrillation propensity [72]. Moreover, even if we often assume that APRs are exposed to solvent during the process of protein folding and accordingly that sequence-based predictions are accurate at this stage, these sequences are usually protected by chaperones and self-chaperoning *in vivo*, not contributing to fibrillation [73]. Finally, most existing programs do not consider post-translational modifications, even though they can dramatically alter aggregation propensities [74]. Thus, it is apparent that despite current progress, we are addressing a complex process with rather simplistic approaches. Because we have in our hands accurate tools to predict conformational switches, posttranslational modifications, and interactions with chaperones, co-factors, or other non-proteinaceous molecules, we should make an effort to integrate these parameters in present programs to provide accurate predictions in biological relevant scenarios.

Regarding extrinsic factors, viscosity, temperature, pH, ionic concentration, protein concentration, solvent identity, and the interaction with other molecules are all known to influence intrinsic aggregation propensities [75]. Subtle variations in these factors are behind the formation of structurally different amyloid polymorphs [76], leading to distinct pathological phenotypes in neurodegenerative disorders [77]. Despite this evidence, scarce efforts have been made to include them in the predictions, and only recently the effect of the pH has been accurately parametrized [78] and incorporated into a webserver [79], although only for disordered proteins. The main limitation to developing methods that can mimic the protein microenvironment in their predictions is the lack of systematic experimental data covering all possible variables combinations for a set of structurally and sequentially unrelated proteins. Gathering these data is often seen as a “low regard” objective, but we have all witnessed how artificial intelligence (AI) has revolutionized structural biology, which would not have been possible without the previous existence of an extensive collection of protein structures.

Our opinion is that building up algorithms that can effectively predict aggregation in the specific conditions occurring at the neuronal synapsis or allow stable formulations of antibodies for immunotherapies, just to mention a couple of applications, is indeed a “high reward” objective. 2021 has witnessed the development of impressive advances that should help in attaining these challenging objectives, like the obtention of high-resolution cryo-EM structures of amyloid fibrils extracted from the organs of patients [80] and AI-assisted elucidation of thousands of human protein structures [81]. Third-generation algorithms and databases are just around the corner.

Funding

Our lab is funded by the Spanish Ministry of Science and Innovation (PID2019-105017RB-I00) by ICREA, (ICREA-Academia 2020) and by EU (PhasAge/H2020-WIDESPREAD-2020-5).

Conflict of interest statement

Nothing declared.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Kuan SL, Bergamini FRG, Weil T: **Functional protein nanostructures: a chemical toolbox**. *Chem Soc Rev* 2018, **47**: 9069–9105.

2. Alberti S, Hyman AA: **Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing**. *Nat Rev Mol Cell Biol* 2021, **22**:196–213.

3. Garcia-Seisdedos H, Villegas JA, Levy ED: **Infinite assembly of folded proteins in evolution, disease, and engineering**. *Angew Chem Int Ed Engl* 2019, **58**:5514–5531.

4. Dey S, Levy ED: **PDB-wide identification of physiological hetero-oligomeric assemblies based on conserved quaternary structure geometry**. *Structure* 2021, **29**: 1303–1311.e3.

5. Riek R, Eisenberg DS: **The activities of amyloids from a structural perspective**. *Nature* 2016, **539**:227–235.

6. • Dobson CM, Knowles TPJ, Vendruscolo M: **The amyloid phenomenon and its significance in biology and medicine**. *Cold Spring Harbor Perspect Biol* 2020, **12**.

An overview of the field that addresses the nature of the amyloid state, its conection with protein homoestasis and strategies to target its link with disease.

7. Lall D, Lorenzini I, Mota TA, Bell S, Mahan TE, Ulrich JD, Davtyan H, Rexach JE, Muhammad AKMG, Shelest O, *et al.*: **C9orf72 deficiency promotes microglial-mediated synaptic loss in aging and amyloid accumulation**. *Neuron* 2021, **109**: 2275–2279.

8. Hamrang Z, Rattray NJW, Pluen A: **Proteins behaving badly: emerging technologies in profiling biopharmaceutical aggregation**. *Trends Biotechnol* 2013, **31**:448–458.

9. • Shanmugam N, Baker MODG, Ball SR, Steain M, Pham CLL, Sunde M: **Microbial functional amyloids serve diverse purposes for structure, adhesion and defence**. *Biophys Rev* 2019, **11**:287–302. 2019 113.

An extensive and detailed review on microbial functional amyloids and their roles.

10. Levkovich SA, Gazit E, Laor Bar-Yosef D: **Two decades of studying functional amyloids in microorganisms**. *Trends Microbiol* 2021, **29**:251–265.

11. Rubel MS, Fedotov SA, Grizel AV, Sopova JV, Malikova OA, Chernoff YO, Rubel AA: **Functional mammalian amyloids and amyloid-like proteins**. *Life* 2020, **10**:1–32.

12. Santos J, Ventura S: **Functional amyloids germinate in plants**. *Trends Plant Sci* 2021, **26**:7–10.

13. Zozulia O, Dolan MA, Korendovych IV: **Catalytic peptide assemblies**. *Chem Soc Rev* 2018, **47**:3621–3639.

14. Shen Y, Levin A, Kamada A, Toprakcioglu Z, Rodriguez-Garcia M, Xu Y, Knowles TPJ: **From protein building blocks to functional materials**. *ACS Nano* 2021, **15**:5819–5837.

15. Zeng R, Lv C, Wang C, Zhao G: **Bionanomaterials based on protein self-assembly: design and applications in biotechnology**. *Biotechnol Adv* 2021, **52**:107835.

16. Santos J, Pujols J, Pallarès I, Iglesias V, Ventura S: **Computational prediction of protein aggregation: advances in proteomics, conformation-specific algorithms and biotechnological applications**. *Comput Struct Biotechnol J* 2020, **18**:1403–1413.

17. • Langenberg T, Gallardo R, van der Kant R, Louros N, Michiels E, Duran-Romaña R, Houben B, Cassio R, Wilkinson H, Garcia T, *et al.*: **Thermodynamic and evolutionary coupling between the native and amyloid state of globular proteins**. *Cell Rep* 2020, **31**.

Describes how positive selection of amyloid propensity favors globular proteins stability.

18. Yagi-Utsumi M, Yanaka S, Song C, Satoh T, Yamazaki C, Kasahara H, Shimazu T, Murata K, Kato K: **Characterization of amyloid β fibril formation under microgravity conditions**. *NPJ Microgravity* 2020, **6**.

19. • Kollmer M, Close W, Funk L, Rasmussen J, Bsoul A, Schierhorn A, Schmidt M, Sigurdson CJ, Jucker M, Fändrich M: **Cryo-EM structure and polymorphism of Aβ amyloid fibrils purified from Alzheimer's brain tissue**. *Nat Commun* 2019, **10**.

Evidence for the *in vivo* structural polymorphism of Aβ amyloid fibrils.

20. Hoppe SO, Uzunoğlu G, Nussbaum-Krammer C: **α-Synuclein strains: does amyloid conformation explain the heterogeneity of synucleinopathies?** *Biomolecules* 2021, **11**:931.

21. Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S: **AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides**. *BMC Bioinf* 2007, **8**:65.

22. Tartaglia GG, Vendruscolo M: **The Zyggregator method for predicting protein aggregation propensities**. *Chem Soc Rev* 2008, **37**:1395.

23. Maurer-Stroh S, Debulpaep M, Kuemmerer N, de la Paz ML, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, *et al.*: **Exploring the sequence determinants of amyloid structure using position-specific scoring matrices**. *Nat Methods* 2010, **7**:237–242.

24. Zibaee S, Makin OS, Goedert M, Serpell LC: **A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone**. *Protein Sci* 2007, **16**:906–918.

25. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A: **Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences**. *Protein Sci* 2005, **14**:2723–2734.

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

27. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV: **FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence**. *Bioinformatics* 2010, **26**:326–332.

28. Walsh I, Seno F, Tosatto SCE, Trovato A: **PASTA 2.0: an improved server for protein aggregation prediction**. *Nucleic Acids Res* 2014, **42**:W301–W307.

29. Tsolis AC, Papandreou NC, Iconomidou VA, Hamodrakas SJ: **A consensus method for the prediction of "aggregation-prone" peptides in globular proteins**. *PLoS One* 2013, **8**.

30. Emily M, Talvas A, Delamarche C: **MetAmyl: a METa-predictor for AMYLoid proteins**. *PLoS One* 2013, **8**, e79722.

31. • Louros N, Konstantoulea K, De Vleeschouwer M, Ramakers M, Schymkowitz J, Rousseau F: **WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides**. *Nucleic Acids Res* 2020, **48**:D389–D393.

An updated open-access database storing >1400 amyloid-forming hexapeptide entries.

32. Tsiolaki PL, Nastou KC, Hamodrakas SJ, Iconomidou VA: **Mining databases for protein aggregation: a review**. *Amyloid* 2017, **24**:143–152.

33. Szulc N, Burdukiewicz M, Gąsior-Głogowska M, Wojciechowski JW, Chilimonîuk J, Mackiewicz P, Šneideris T, Smirnovas V, Kotulska M: **Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data**. *Sci Rep* 2021, **11**:8934.

34. • Prabakaran R, Rawat P, Kumar S, Michael Gromiha M: **ANuPP: a versatile tool to predict aggregation nucleating regions in peptides and proteins**. *J Mol Biol* 2021, **433**:166707.

An ensemble-classifier trained to identify amyloid-fibril forming peptides and regions in protein sequences.

35. Wojciechowski JW, Kotulska M: **PATH - prediction of amyloidogenicity by threading and machine learning**. *Sci Rep* 2020, **10**:1–9. 2020 101.

36. Kim C, Choi J, Lee SJ, Welsh WJ, Yoon S: **NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation**. *Nucleic Acids Res* 2009, **37**:W469.

37. Gasior P, Kotulska M: **FISH Amyloid - a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids**. *BMC Bioinf* 2014, **15**:54.

38. Niu M, Li Y, Wang C, Han K: **RFAmyloid: a web server for predicting amyloid proteins**. *Int J Mol Sci* 2018, **19**.

39. Keresztes L, Szögi E, Varga B, Farkas V, Perczel A, Grolmusz V: **The budapest amyloid predictor and its applications**. *Biomolecules* 2021, **11**.

40. •• Louros N, Orlando G, De Vleeschouwer M, Rousseau F, Schymkowitz J: **Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities**. *Nat Commun* 2020, **11**.

Demonstration that the amyloid space is significantly larger than previously thought, including soluble protein sequences.

41. Sawaya MR, Sambashivan S, Nelson R, Ivanova MI, Sievers SA, Apostol MI, Thompson MJ, Balbirnie M, Wiltzius JJW, McFarlane HT, *et al.*: **Atomic structures of amyloid cross-β spines reveal varied steric zippers**. *Nature* 2007, **447**:453–457. 2006 4477143.

42. Xu Y, Verma D, Sheridan RP, Liaw A, Ma J, Marshall NM, McIntosh J, Sherer EC, Svetnik V, Johnston JM: **Deep dive into machine learning models for protein engineering**. *J Chem Inf Model* 2020, **60**:2773–2790.

43. Raimondi D, Orlando G, Fariselli P, Moreau Y: **Insight into the protein solubility driving forces with neural attention**. *PLoS Comput Biol* 2020, **16**.

44. • Yang W, Tan P, Fu X, Hong L: **Prediction of amyloid aggregation rates by machine learning and feature selection**. *J Chem Phys* 2019, **151**.

A neuronal network approach to predict amyloid aggregation rates from intrinsic and extrinsic factors.

45. Michelitsch MD, Weissman JS: **A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions**. *Proc Natl Acad Sci U S A* 2000, **97**:11910–11915.

46. Harrison PM, Gerstein M: **A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes**. *Genome Biol* 2003, **4**:R40.

47. Toombs Ja, Petri M, Paul KR, Kan GY, Ben-Hur A, Ross ED: **De novo design of synthetic prion domains**. *Proc Natl Acad Sci U S A* 2012, **109**:6519–6524.

48. Lancaster AK, Nutter-Upham A, Lindquist S, King OD: **PLAAC: a web and command-line application to identify proteins with Prion-Like Amino Acid Composition**. *Bioinformatics* 2014, **30**:2–3.

49. Espinosa Angarica V, Angulo A, Giner A, Losilla G, Ventura S, Sancho J: **PrionScan: an online database of predicted prion domains in complete proteomes**. *BMC Genom* 2014, **15**:102.

50. Afsar Minhas F ul A, Ross ED, Ben-Hur A: **Amino acid composition predicts prion activity**. *PLoS Comput Biol* 2017, **13**.

51. Sabate R, Rousseau F, Schymkowitz J, Ventura S: **What makes a protein sequence a prion?** *PLoS Comput Biol* 2015, **11**, e1004013.

52. Zambrano R, Conchillo-Sole O, Iglesias V, Illa R, Rousseau F, Schymkowitz J, Sabate R, Daura X, Ventura S: **PrionW: a server to identify proteins containing glutamine/asparagine rich prion-like domains and their amyloid cores**. *Nucleic Acids Res* 2015, Jul 1, **43**:W331–W337.

53. • Gil-Garcia M, Iglesias V, Pallarès S, Ventura S: **Prion-like proteins: from computational approaches to proteome-wide analysis**. *FEBS Open Bio* 2021, Sep, **11**:2400–2417.
State-of-the art computational approaches for the identification of prion-like proteins and their function in proteomes.

54. Chiti F, Dobson CM: **Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade**. *Annu Rev Biochem* 2017, **86**:27–68.

55. Van Durme J, De Baets G, Van Der Kant R, Ramakers M, Ganesan A, Wilkinson H, Gallardo R, Rousseau F, Schymkowitz J: **Solubis: a webserver to reduce protein aggregation through mutation**. *Protein Eng Des Sel* 2016, **29**:285–289.

56. Sankar K, Krystek SR, Carl SM, Day T, Maier JKX: **AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches**. *Proteins Struct Funct Bioinforma* 2018, **86**:1147–1156.

57. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL: **Design of therapeutic proteins with enhanced stability**. *Proc Natl Acad Sci U S A* 2009, **106**:11937–11942.

58. Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S: **AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures**. *Nucleic Acids Res* 2015, Jul 1, **43**:W306–W313.

59. • Kuriata A, Iglesias V, Pujols J, Kurcinski M, Kmiecik S, Ventura S: **Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility**. *Nucleic Acids Res* 2019, **47**:W300–W307.
An updated AGGRESCAN3D algorithm allowing automated mutation analysis and dynamic calculations of large multimeric proteins, such as antibodies.

60. Sormanni P, Aprile FA, Vendruscolo M: **The CamSol method of rational design of protein mutants with enhanced solubility**. *J Mol Biol* 2015, **427**:478–490.

61. Kuriata A, Gierut AM, Oleniecki T, Ciemny MP, Kolinski A,M, Kmiecik S: **CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures**. *Nucleic Acids Res* 2018, **46**: W338–W343.

62. Gil-Garcia M, Bañó-Polo M, Varejão N, Jamroz M, Kuriata A, Díaz-Caballero M, Lascorz J, Morel B, Navarro S, Reverter D, *et al.*: **Combining structural aggregation propensity and stability predictions to redesign protein solubility**. *Mol Pharm* 2018, **15**:3846–3859.

63. Service RF: **"The game has changed." AI triumphs at protein folding**. *Science* 2020, **370**:1144–1145.

64. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, *et al.*: **Accurate prediction of protein structures and interactions using a three-track neural network**. *Science* 2021, Aug 20, **373**: 871–876.

65. • Pinheiro F, Santos J, Ventura S: **AlphaFold and the amyloid landscape**. *J Mol Biol* 2021, https://doi.org/10.1016/J.JMB.2021.167059.
The authors debate how the emergence of artificial intelligence applications like AlphaFold would impact the study of amyloids.

66. Daskalov A, Mammeri N El, Lends A, Shenoy J, Lamon G, Fichou Y, Saad A, Martinez D, Morvan E, Berbon M, *et al.*: **Structures of pathological and functional amyloids and prions, a solid-state NMR perspective**. *Front Mol Neurosci* 2021, Jul 1, **14**:670513.

67. Ragonis-Bachar P, Landau M: **Functional and pathological amyloid structures in the eyes of 2020 cryo-EM**. *Curr Opin Struct Biol* 2021, **68**:184–193.

68. Tompa DR, Kadhirvel S: **Changes in hydrophobicity mainly promotes the aggregation tendency of ALS associated SOD1 mutants**. *Int J Biol Macromol* 2020, **145**:904–913.

69. Habibi N, Mohd Hashim SZ, Norouzi A, Samian MR: **A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli**. *BMC Bioinf* 2014, **15**.

70. Navarro S, Ventura S: **Computational re-design of protein structures to improve solubility**. *Expet Opin Drug Discov* 2019, **14**:1077–1088.

71. •• Santos J, Pallarès I, Iglesias V, Ventura S: **Cryptic amyloidogenic regions in intrinsically disordered proteins: function and disease association**. *Comput Struct Biotechnol J* 2021, **19**: 4192–4206.
Demonstration that, in contrast to what is usually assumed, intrinsically disordered proteins contain amyloidogenic sequences that are important for their funtion, with implications for the emergence of globular folds.

72. Ulamec SM, Brockwell DJ, Radford SE: **Looking beyond the core: the role of flanking regions in the aggregation of amyloidogenic peptides and proteins**. *Front Neurosci* 2020, **14**.

73. Goldschmidt L, Teng PK, Riek R, Eisenberg D: **Identifying the amylome, proteins capable of forming amyloid-like fibrils**. *Proc Natl Acad Sci U S A* 2010, **107**:3487–3492.

74. Moon SP, Balana AT, Pratt MR: **Consequences of post-translational modifications on amyloid proteins as revealed by protein semisynthesis**. *Curr Opin Chem Biol* 2021, **64**:76–89.

75. Brudar S, Hribar-Lee B: **Effect of buffer on protein stability in aqueous solutions: a simple protein aggregation model**. *J Phys Chem B* 2021, **125**:2504–2512.

76. • Holec SAM, Woerman AL: **Evidence of distinct α-synuclein strains underlying disease heterogeneity**. *Acta Neuropathol* 2021, **142**.
Evidence for the rol of α-synuclein amyloid polymorphs in disease manifestation.

77. Froula JM, Castellana-Cruz M, Anabtawi NM, Camino JD, Chen SW, Thrasher DR, Freire J, Yazdi AA, Fleming S, Dobson CM, *et al.*: **Defining α-synuclein species responsible for Parkinson's disease phenotypes in mice**. *J Biol Chem* 2019, Jul 5, **294**:10392–10406.

78. Santos J, Iglesias V, Santos-Suárez J, Mangiagalli M, Brocca S, Pallarès I, Ventura S: **pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity**. *Cells* 2020, **9**:145. 2020, 9:145.

79. • Pintado C, Santos J, Iglesias V, Salvador V: **SolupHred: a server to predict the pH-dependent aggregation of intrinsically disordered proteins**. *Bioinformatics* 2021, **37**.
A server that explicitly considers the impact of the environmental pH on the aggregation of intrinsically disordered proteins.

80. • Shi Y, Zhang W, Yang Y, Murzin AG, Falcon B, Kotecha A, van Beers M, Tarutani A, Kametani F, Garringer HJ, *et al.*: **Structure-based classification of tauopathies**. *Nature* 2021, **598**: 359–363.
Cryo-electron microscopy structures of tau filaments show that the fibrils folds differ between tauopathies and allow to classify diferent clinical manifestations.

81. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool T, Bates R, Žídek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**:583–589.

82. Hamodrakas SJ, Liappa C, Iconomidou VA: **Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins**. *Int J Biol Macromol* 2007, **41**:295–300.

83. Ahmed AB, Znassi N, Château M-T, Kajava AV: **A structure-based approach to predict predisposition to amyloidosis**. *Alzheimer's Dementia* 2015, **11**:681–690.

84. Bondarev SA, Bondareva OV, Zhouravleva GA, Kajava AV: **BetaSerpentine: a bioinformatics tool for reconstruction of amyloid structures**. *Bioinformatics* 2018, **34**:599–608.

85. Bryan AW, Menke M, Cowen LJ, Lindquist SL, Berger B: **BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis**. *PLoS Comput Biol* 2009, **5**, e1000333.

86. O'Donnell CW, Waldispühl J, Lis M, Halfmann R, Devadas S, Lindquist S, Berger B: **A method for probing the mutational landscape of amyloid structure**. *Bioinformatics* 2011, **27**: i34–i42.

87. Bryan AW, O'Donnell CW, Menke M, Cowen LJ, Lindquist S, Berger B: **STITCHER: dynamic assembly of likely amyloid and prion β-structures from secondary structure predictions**. *Proteins Struct Funct Bioinforma* 2012, **80**:410–420.

88. Thangakani AM, Kumar S, Nagarajan R, Velmurugan D, Gromiha MM: **GAP: towards almost 100 percent prediction for β-strand-mediated aggregating peptides with distinct morphologies**. *Bioinformatics* 2014, **30**:1983–1990.

89. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D: **The 3D profile method for identifying fibril-forming segments of proteins**. *Proc Natl Acad Sci U S A* 2006, **103**:4074.

90. Orlando G, Silva A, Macedo-Ribeiro S, Raimondi D, Vranken W: **Accurate prediction of protein beta-aggregation with generalized statistical potentials**. *Bioinformatics* 2020, **36**:2076–2081.

91. Li Y, Zhang Z, Teng Z, Liu X: **PredAmyl-MLP: prediction of amyloid proteins using multilayer perceptron**. *Comput Math Methods Med* 2020, Nov 20, **2020**:8845133.

92. Liaw C, Tung CW, Ho SY: **Prediction and analysis of antibody amyloidogenesis from sequences**. *PLoS One* 2013, **8**.

93. Tian J, Wu N, Guo J, Fan Y: **Prediction of amyloid fibril-forming segments based on a support vector machine**. *BMC Bioinf* 2009, **10**:S45.

94. Família C, Dennison SR, Quintas A, Phoenix DA: **Prediction of peptide and protein propensity for amyloid formation**. *PLoS One* 2015, **10**, e0134679.

95. Burdukiewicz M, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M: **Amyloidogenic motifs revealed by n-gram analysis**. *Sci Rep* 2017, **7**:1–10. 2017 71.