Computational and Structural Biotechnology Journal 23 (2024) 4011–4018

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Mini-review

Aggregating amyloid resources: A comprehensive review of databases on amyloid-like aggregation

Valentín Iglesias [a,1], Jarosław Chilimoniuk [a,1], Carlos Pintado-Grima [b], Oriol Bárcenas [b,c], Salvador Ventura [b,d], Michał Burdukiewicz [a,e,*]

[a] Clinical Research Centre, Medical University of Białystok, Białystok, Poland
[b] Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
[c] Institute of Advanced Chemistry of Catalonia (IQAC), CSIC, Barcelona, Spain
[d] Hospital Universitari Parc Taulí, Institut d'Investigació i Innovació Parc Taulí (I3PT-CERCA), Universitat Autònoma de Barcelona, Sabadell, Spain
[e] Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania

ARTICLE INFO

Keywords:
Amyloid
Bioinformatics
Biotherapeutics
Database
Protein aggregation
Protein solubility
Prion

ABSTRACT

Protein aggregation is responsible for several degenerative conditions in humans, and it is also a bottleneck in industrial protein production and storage of biotherapeutics. Bioinformatics tools have been developed to predict and redesign protein solubility more efficiently by understanding the underlying principles behind aggregation. As more experimental data become available, dedicated resources for storing, indexing, classifying and consolidating experimental results have emerged. These resources vary in focus, including aggregation-prone regions, 3D patches or protein stretches capable of forming amyloid fibrils. Some of these resources also consider the experimental conditions that cause protein aggregation and how they affect the process. This review article explores how protein aggregation databases have evolved and surveys state-of-the-art resources. We highlight their applications, complementarity and existing limitations. Moreover, we showcase the existing symbiosis between amyloid-related databases and predictive tools. To increase the usefulness of our review, we supplement it with a comprehensive list of present and past amyloid databases: https://biogenies.info/amyloid-database-list/.

1. Introduction

Protein aggregation is a second-order reaction in which soluble monomeric species transit towards multimeric architectures forming protein deposits, usually highly ordered fibrillar structures named amyloids. The amyloid conformation consists of an intermolecular in-register stacking of β-stranded proteins in parallel or antiparallel form, which runs perpendicular to the fiber axis [1]. Amyloid fibrils can be detected by the binding of specific dyes such as Thioflavin-T or Congo Red, detergent and proteolytic resistance, and the presence of cross-β signals on X-ray diffraction patterns, typically at 4.7 and 10.2 Å [2]. The ability to form amyloid structures appears to be a generic property of proteins and is not tied to specific sequences of amino acids [3].

This fibrillization of proteins is widely recognized as a key factor in the onset of a myriad of different debilitating human conditions known as protein amyloidoses. This category covers disorders such as Parkinson's disease (PD), Alzheimer's disease (AD), Creutzfeldt-Jakob's, Amyotrophic Lateral Sclerosis (ALS) or Transthyretin (ATTR) amyloidosis [4]. Structural determinations of different amyloids from biopsies reveal the same core amyloid-forming protein can achieve multiple arrangements *in vivo* [5]. Specific amyloid polymorphs are thought to be associated with distinct disease manifestations [6]. The International Society of Amyloidosis recommends notating proteins constituting amyloid fibrils which deposition causes these disorders with an initial A (or amyloid) [7].

Recently, amyloidoses has garnered attention in the context of diabetes and various cancers, as well as viral, parasitic and bacterial infections [8–11]. Moreover, amyloids can also indirectly accelerate the onset of other diseases through cross-seeding, where one type of amyloid aggregate promotes the self-assembly of a different type of

Abbreviations: APRs, aggregation-prone regions; CARs, Cryptic amyloidogenic regions; PDB, Protein Data Bank; PrLD, Prion-like domain.

* Corresponding author at: Clinical Research Centre, Medical University of Białystok, Białystok, Poland.
E-mail address: michalburdukiewicz@gmail.com (M. Burdukiewicz).

[1] Valentın Iglesias and Jarosław Chilimoniuk contributed equally to this article

https://doi.org/10.1016/j.csbj.2024.10.047
Received 1 August 2024; Received in revised form 24 October 2024; Accepted 27 October 2024
Available online 29 October 2024

amyloidogenic protein [12].

The aggregation of protein-based products is a significant concern in the biopharmaceutical industry. While not always amyloid-like, such aggregation can result in substantial economic losses, leading to production bottlenecks, particularly in developing biotherapeutics and other protein-based products [13]. Consequently, considerable effort has been devoted to minimizing aberrant self-assembly during the development, production, and formulation of these proteins [14].

The same properties of amyloids that lead to their pathological aggregation in cells, in turn, are beneficial for organisms to develop specific biological functions such as scaffolding bacterial biofilms, eukaryotic eggshells or serving as amino acid storage in plant seeds [15–17]. This phenomenon prompted efforts to generate amyloids of functionalized self-assembled nanostructures [18,19]. Moreover, disturbing protein homeostasis by inducing protein aggregation has been shown as a viable antimicrobial strategy [20,21].

A protein's aggregation propensity is primarily determined by its amino acid sequence and the spatial arrangement of such residues. Consecutive amino acids with high aggregation propensity constitute aggregation-prone regions (APRs). High hydrophobicity, low local or net charge and a favorable propensity to form a β-sheet structure are considered the primary factors contributing to amyloid aggregation of linear amyloid sequences [2,6,22]. Cryptic Amyloidogenic Regions (CARs) are sequential stretches found in disordered proteins with mild amyloidogenic character [23]. CARs are widespread in IDRs and other low-complexity regions such as PrLDs [24]. This is due to their lower risk of undergoing pathogenic aggregation while maintaining a high prevalence to establish functional protein-protein interactions.

On the other hand, folded proteins may display spatially clustered APRs (STAPs), including non-consecutive amino acids [5]. However, clustered hydrophobic amino acids protected from the solvent in the hydrophobic core or transmembrane segments have negligible effect on protein aggregation. Thus, knowledge of their 3D conformation is required to correctly assess folded proteins' aggregation potential in their native state. These sequential and structural elements, among others, dictate the amyloid-forming capabilities of proteins.

The protein aggregation propensity of a given polypeptide can be heavily modulated by environmental factors impacting the reaction's kinetics, thermodynamics or structural properties. Protein concentration, incubation temperature, pH, identity and osmolarity of salts, reducing/oxidizing compounds, post-translational modifications (PTMs), presence of lipids, presence or absence of pre-formed fibrils or other additives, as well as stirring the samples have a particular effect on the deposition process [25,26].

One of the specific examples of amyloids is prion proteins. Prusiner first coined the term "prion" to refer to the proteinaceous particle capable of inducing neurodegenerative conditions in mammals [27]. This protein could post-translationally convert the soluble native state into an infectious, self-templating and self-propagating cytotoxic conformation between cells, individuals and even species [28]. Expansion of the prion phenomenon beyond mammalian diseases allowed the identification of novel prions and also of prion-like proteins and their prion-like domains (PrLDs): proteins capable of prion conversion but unable of transmission between individuals or species [29].

Due to the multifaceted complexity of amyloid aggregation, involving genetic, biochemical, and physicochemical factors, researchers attempted to gather knowledge on that topic in dedicated databases [30]. This review aims to list the currently available databases on that topic, present their scope and discuss their co-evolution with dedicated predictive algorithms. Moreover, we showcase how the field of amyloid self-assembly impacts the prediction and annotation of non-amyloid aggregation.

2. Amyloid aggregation databases

In our analysis of amyloid databases, we categorize the available resources based on two primary criteria. Firstly, we differentiate between databases containing only information on amyloid sequences and those focusing on detailed structural data. Secondly, we classify these databases by the nature of their data, distinguishing between experimentally confirmed data and emerging datasets of predicted amyloid-related properties (**see** Fig. 1 and Table 1).

2.1. AmyLoad

AmyLoad (http://comprec-lin.iiar.pwr.edu.pl/amyload/)[31] collects peptides and proteins with experimentally verified amyloid propensity. The database stores 1400 entries with annotations on self-assembly potential and the experimental method and conditions employed to measure the aggregation.

2.2. AmyloBase

AmyloBase (http://bioserver2.sbsc.unifi.it/AmyloBase.html)[32] collects experimental data on the self-assembly kinetics of point mutants of three unique proteins and peptides. The database contains the studied fragment position in protein, its length and mass, protein origin, mutation type, number of hotspots, experimental conditions (pH, protein concentration, ionic strength, temperature), study method, the kinetics of aggregation (e.g., lag phase) and if the end product is amyloid fibrils.

2.3. AmyloGraph

AmyloGraph (http://amylograph.com/)[33] explores the concept of how different amyloid precursor proteins impact aggregation and amyloid formation of one another. The database stores manually curated data from 562 manuscripts, resulting in 896 records of experimentally derived data on amyloid-amyloid interactions for 46 amyloidogenic proteins. Each entry represents one interaction and can be studied as a graph or table. In the table format, users can find both interacting proteins' sequences, their length, impact on aggregation speed and fibril morphology (homo- or heterogeneous).

2.4. Amyloid atlas

Amyloid Atlas (https://people.mbi.ucla.edu/sawaya/amyloidatlas/) [1] provides a manually curated list of structures of human pathological amyloids based on cryogenic electron microscopy (cryo-EM), solid-state nuclear magnetic resonance (ssNMR) and microcrystal electron diffraction amyloid structures derived from PDB [42]. Storing 506 fibril entries at the moment of writing, Amyloid Atlas is the most comprehensive resource dedicated to the 3D structures of amyloids. Each entry contains the protein's name, origin, and the 3D structure colored according to the residue polarity and estimated solvation energy. Energetic stabilization is also itemized per chain, layer, and residue.

2.5. AmyPro

The AmyPro database (http://amypro.net)[34] covers 125 records categorized using four labels: functional amyloid, functional prion, pathogenic, biologically not relevant or not known. Each entry provides protein sequence with highlighted APRs. In the case of proteins with solved 3D structures, amyloid-forming stretches are marked on top. Moreover, AmyPro provides links to relevant publications describing amyloidogenicity of proteins.

2.6. Cryptic amyloidogenic regions database (CARs DB)

CARs-DB (http://carsdb.ppmclab.com/)[35] collects CARs within intrinsically disordered regions (IDRs), which are more polar and have milder aggregation potential than amyloidogenic stretches found in globular proteins. CARs-DB offers more than 8900 unique CARs across

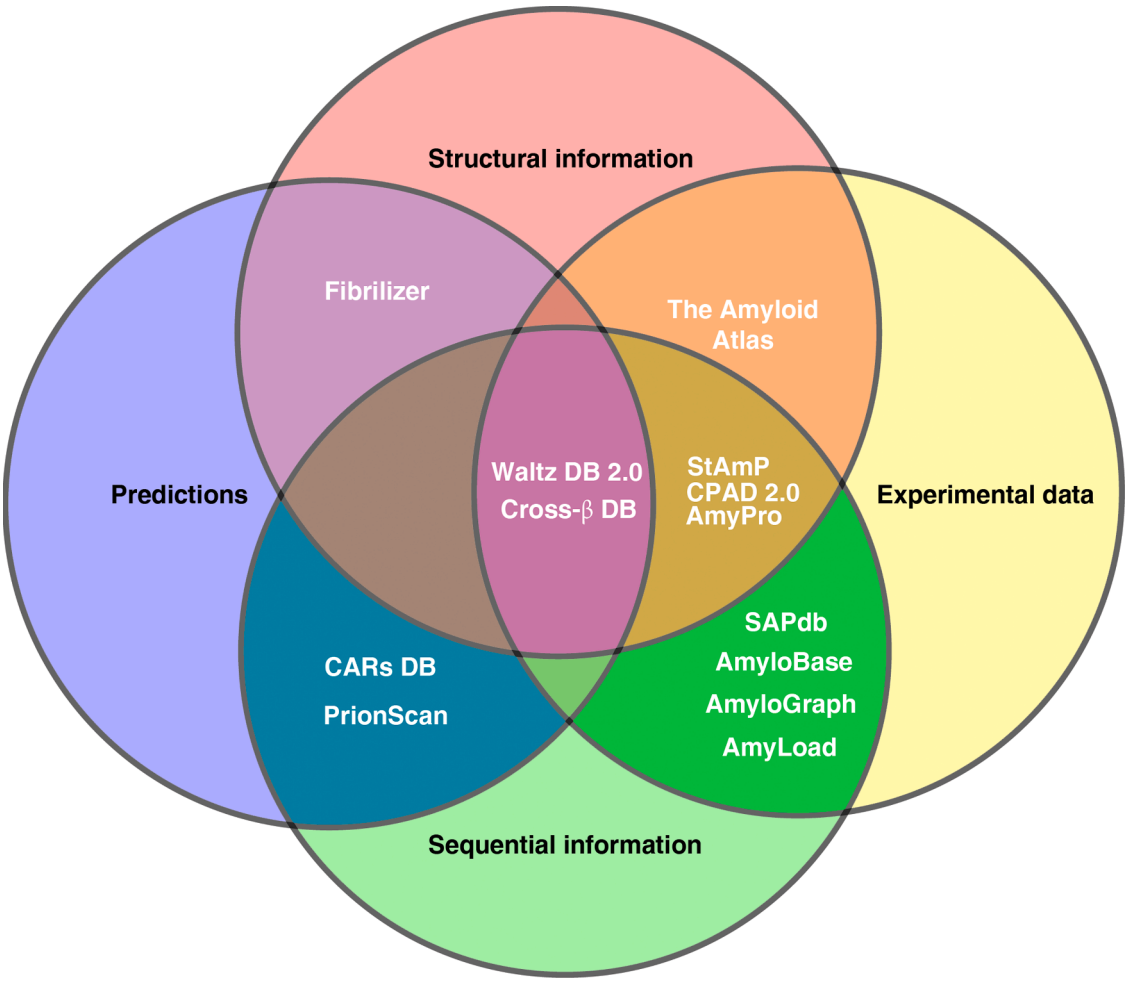

Fig. 1. Summary of active amyloid databases. A distinction is made on the source of their information (experimental or predicted) and the level of structural complexity considered (sequences or structures).

1711 IDRs derived from the DisProt database [43]. Each CAR is described by its sequence, protein of origin, CAR position in the IDR and the Waltz score.

2.7. Curated protein aggregation database (CPAD) 2.0

CPAD 2.0 (https://web.iitm.ac.in/bioinfo2/cpad2/index.html)[36,44] stores manually curated entries related to protein aggregation and is enriched with kinetic and structural information. The collection is divided into four databases: amyloidogenic peptides, APRs, aggregation kinetics, and structures of amyloid fibrils and amyloid-precursor proteins. The first one, peptide-centric, contains peptide sequence, its length, position in a sequence, origin, class (amyloid or non-amyloid), net and absolute charge, and results from NuAPRpred, TANGO [45], AGGRESCAN [46] and PASTA 2.0 [47]. The APR collection describes experimentally validated APRs, including peptide origin, category, mutation type, prion properties (if observed), APRs' protein position, length and sequence. The database of self-assembly kinetics gathers protein aggregation kinetic data, sequence type (wild or mutated), experimental conditions and measurement method. The database also features protein and peptide 3D structures, including the monomer length, origin, mutation type, class (amyloid or non-amyloid precursor), structure determination method, and resolution, if applicable.

2.8. Cross-beta DB

Cross-Beta DB (https://crossbetadb.crbm.cnrs.fr/)[37] gathers amyloid-forming regions of the naturally occurring cross-β structures within amyloid fibrils. The database is the result of careful manual curation on experimentally tested cross-β amyloid forming stretches. It contains 115 individual entries from 44 different amyloid-precursor proteins. In each entry, protein's origin, sequence, length, Archcandy2.0 [48] prediction results, experimental conditions (protein concentration, pH, temperature, buffer), measurement method and fibril state can be found. Moreover, the APR sequence, position, molecular weight and mutations, among others, are displayed. The experimentally obtained 3D structure can also be viewed alongside the amino acid composition graph.

2.9. Fibrillizer

Fibrillizer (https://amyloid.cs.mcgill.ca/database/index.html) collects results from CreateFibril [38], a tool that builds fibril atomic resolution models. The database focuses on energetically favorable amyloid fibril polymorphism, storing potential multiple combinations of energetically possible supra-fibrillar assemblies in the form of single, stack, ring and polygon structures for three proteins: Aβ42, Aamylin and HET-s.

2.10. Prionscan

PrionScan (http://webapps.bifi.es/prionscan/)[39] stores predicted prion-like proteins from complete proteomes. It stores approximately 28000 PrLDs for over 3200 organisms covering major taxonomic

Table 1
Amyloid databases described in this manuscript. The interactive table with extended database descriptions is available online at: https://biogenies.info/amyloid-database-list/. Abbreviations used in the table: ENA: European Nucleotide Archive; CPAD: Curated Protein Aggregation Database; KEGG: Kyoto Encyclopedia of Genes and Genomes.

Database	Link to database	Reference	Data sources and links to other databases and software
AmyLoad	http://comprec-lin.iiar.pwr.edu.pl/amyload/	Wozniak and Kotulska [31]	TANGO; WALTZ; AmylFrag; AGGRESCAN; AmylHex; PubMed
AmyloBase	http://bioserver2.sbsc.unifi.it/AmyloBase.html	Belli, Ramazzotti, and Chiti [32]	PubMed; UniProt
AmyloGraph	http://amylograph.com/	Burdukiewicz et al. [33]	PubMed; UniProt
Amyloid Atlas	https://people.mbi.ucla.edu/sawaya/amyloidatlas/	Sawaya et al. [1]	PDB; PubMed
AmyPro	http://amypro.net	Varadi et al. [34]	PDB; UniProt; PubMed
CARs DB	http://carsdb.ppmclab.com/	Pintado-Grima, et al. [35]	DisProt; UniProt
CPAD 2.0	https://web.iitm.ac.in/bioinfo2/cpad2/index.html	Rawat et al. [36]	CPAD; Waltz-DB 2.0; AmyLoad; AmyPro; UniProt; PDB; PubMed
Cross-Beta DB	https://crossbetadb.crbm.cnrs.fr/	Gonay et al. [37]	PDB; AmyPro; UniProt; PubMed
Fibrilizer	https://amyloid.cs.mcgill.ca/database/index.html	Smaoui et al. [38]	PDB
PrionScan	http://webapps.bifi.es/prionscan/	Espinosa Angarica et al. [39]	UniProt; ENA; Protein; KEGG; Pfam; QuickGO
StAmP	https://stamp.switchlab.org/	Louros et al. [40]	WALTZ-DB; AmyPro; CPAD; PDB; UniProt
Waltz DB 2.0	http://waltzdb.switchlab.org/	Louros, Konstantoulea, et al. [41]	PubMed; UniProt; PDB

divisions. Each entry displays the protein's sequence and origin, and the predicted prion domain can be highlighted.

2.11. Structural analysis of Amyloid polymorphs (StAmP)

The StAmP (https://stamp.switchlab.org/)[40] database focuses on the structural diversity of amyloid fibril polymorphisms. The database, which results from manual curation and is enriched by automatic bioinformatic analyses, gathers 133 experimentally solved fibril structures using ssNMR, solution NMR and cryo-EM from amyloidogenic proteins. STAmP combines all available polymorph structures for a given protein with their APRs. Each entry has a short description indicating mutation type, studied fragment position and method, 3D model and thermodynamic profile. If applicable, it includes annotations about the specific amyloidosis it is found in, the tissue it was derived from and whether it was of human origin. Mutants can be compared using a correlation matrix.

2.12. Waltz-DB 2.0

The original Waltz-DB gathered experimentally verified hexapeptides [49]. The newest database iteration, Waltz-DB 2.0 (http://waltzdb.switchlab.org/), expands the original concept with structural information [41]. Currently, it stores 1416 peptide records, 512 of them have amyloid-forming properties and 904 self-assemble into amorphous aggregates. Each entry contains peptide sequence, information on its ability to form amyloid fibrils, source protein identifier and the position of a peptide in its sequence. Moreover, it reports experimental (TEM, ThT assay, FTIR, Proteostat assay) and predicted results (WALTZ, TANGO, PASTA). It also provides computed hydrophobicity and propensity to form amyloid structures, energy calculations and 3D structure predicted by CORDAX [50]. If applicable, Waltz-DB includes a microscopic image of fibril and aggregation kinetics.

2.13. Merits and shortcomings of amyloid and aggregation databases

Despite described efforts to provide the community with curated resources on the amyloid formation phenomenon, they all present caveats and limitations given their scope or chosen architecture (Table 2). One of the most prevalent problems is having a limited search engine that hinders finding the desired entries. Another widespread and closely related technical limitation is the limited exporting capabilities of the data into established formats (like CSV or JSON). In addition, some databases limit the amount and type of downloadable information. For instance, AmyLoad allows obtaining entry names, polypeptide sequences, and amyloid propensities in bulk. However, users must access each entry individually if they are interested in the information on experimental procedures or the associated references. Some databases do not provide full dataset download but restrict the data obtention to a single entry, as for CPAD 2.0. The last technical aspect is the difficulty of the database usage, mostly related to the user interface or a way of presenting the data.

The database usability is secondary to the data quality provided by each source. Some databases offer limited information on their entries, either because the individual records are described using very few details or contain missing items. Another important consideration involves databases that provide predicted results. While these resources enable immediate access to the results of predictive algorithms, it is essential to remain mindful of the limitations tied to each predictive algorithm. Finally, several databases gather a low number of entries, often reflecting a focus on a particular topic, like naturally occurring cross-β forming amyloid databases (Cross-Beta DB).

The diversity of data stored in these databases allows for exploring mechanisms of protein aggregation and amyloidosis from different angles. However, this diversity hinders the compilation of available resources into a single knowledge base and the subsequent development of the unified benchmark dataset for predictors of amyloidogenicity. Instead, the tools solve a problem best described by a single available dataset.

One major limitation affecting efforts at predicting amyloidogenicity of proteins and peptides is the unanimous focus on sequences as its sole determinant. Although this process is directly tied to the properties of protein sequences, it is heavily influenced by environmental conditions. Therefore, while it is intractable to perform all experiments *in vivo*, more emphasis should be put on reporting the exact conditions where amyloidogenicity is observed. Recently, the MIRRAGGE initiative (Minimum Information Required for Reproducible AGGregation Experiments) established a standardized framework for reporting protein aggregation experiments [51], aiming to increase the consistency and reproducibility of experimental data and subsequently harmonize descriptions of reported experimental conditions. It is paramount to develop tools to assess whether aggregation occurs in physiologically compatible conditions and predict the influence of environmental factors.

3. Other amyloid- and aggregation-related databases

The importance of amyloid self-assembly has led to the creation of several databases that compile extensive information on this complex process. While the databases discussed in Section 2 focus on the biophysical aspects of amyloid aggregation, other resources address different aspects of this issue. Databases such as ALZGENE [52], PDGENE [53], or ALSGENE [54] explore the genetic patterns

Table 2
Main limitations of described databases.

Database	Limited filtering	Limited exports	Hard to navigate or use	Limited entry information	Prediction database	Low number of entries
AmyLoad	X	X		X		
AmyloBase	X	X				X
AmyloGraph			X			
Amyloid Atlas	X	X				
AmyPro	X					X
CARs DB	X				X	
CPAD 2.0			X			
Cross-Beta DB						X
Fibrilizer	X	X		X	X	X
PrionScan					X	
StAmP	X					
Waltz DB 2.0						X

influencing the pathological amyloid accumulation in AD, PD, and ALS, respectively, reporting genetic association of gene variants or non-synonymous single nucleotide polymorphisms (SNPs) to these disorders. By covering meta-analyses from multiple Genome-Wide Association Studies (GWAS) or association studies, novel genes with roles in these amyloidosis (other than the aggregating amyloid protein) can be established. For instance, an increase of the cleavage of amyloid-beta precursor protein (APP) into the Aβ peptide can be observed due to a SNP in Calcium homeostasis 1 (CALHM1) that causes dysregulation of Ca^{2+} homeostasis [55] and by a combination of low expression levels of Oxysterol-binding protein-1 (OSBP1) and high intracellular cholesterol levels [56].

AL-Base [57] takes a more protein-centric approach while maintaining a clear focus on disease. It is a database of antibody light chain sequences associated with plasma cell dyscrasias, especially immunoglobulin light chain amyloidosis. The database contains almost 5000 nucleotide and protein sequences, categorized by germline (κ, λ) and clinical status.

The αSynPEP-DB [58] was based on the discovery of naturally occurring LL-37 human peptide that was observed to inhibit the aggregation of alpha-synuclein protein (αSyn) [59,60]. The database gathers 123 biogenic peptides found in PD-relevant tissues predicted to have similar inhibitory potential. These peptides with unique structural information are predicted to bind only to the toxic species of αSyn and hold promising therapeutic potential for PD. Each record has peptide name, inhibitory sequence and length, type (neuropeptide, antimicrobial, food-derived, gut-microbiome), helical score, hydrophobic score, dipole moment, and net charge per residue. Expanded information on the peptide such as predicted cytotoxicity, blood-brain barrier permeability or expression levels can be found in each entry.

Amyloid aggregation could also be considered one of the subfields of general protein aggregation. The Aggrescan3D Model Organism Database (A3D-MODB) [61], built upon predictions of Aggrescan3D 2.0 [62], focuses on protein aggregation. A3D-MODB provides proteome-wide predictions for protein solubility and aggregation properties from the native state for 12 model organisms. Each entry includes a detailed description of the protein's structure and aggregation propensity.

The self-assembly of amyloid fibrils has a low thermodynamic cost, making them useful for nanomaterial development. As a result, some databases focus on collecting data related to amyloid aggregation from a nanotechnology perspective. For example, SAPdb [63] contains 1049 entries of experimentally validated nanostructures formed by tripeptides, dipeptides, and single amino acids. It also provides detailed information about their chemical modifications and experimental conditions. While the data primarily comes from amyloid-related resources like AmyLoad and Waltz-DB, this database presents the information uniquely by filtering based on the size of the self-assembled nanostructure.

4. Co-evolution of amyloid databases and prediction of amyloid propensity

Amyloid self-assembly datasets and databases have been pivotal for pushing forward the understanding of protein aggregation. In part, these resources have paved the way for the development of predictive tools (Fig. 2). Initially, the collection of sequences capable of amyloid self-assembly was motivated by an attempt to disentangle the underlying mechanisms behind this process. In an early study, Chiti *et al.* conducted multiple mutations on the acylphosphatase protein, measuring the changes in aggregation rate *in vitro*, and gathered bibliographical data for seven other polypeptides [64]. The expansion of this initial dataset [65,66], led to the development of the Zyggregator prediction method [67].

Similarly, López de la Paz and Serrano performed saturation mutagenesis on all positions of an amyloid-forming peptide [68]. The findings of this study and the community-generated AmylHex database [69] spurred the development of the Waltz algorithm, which utilizes position-specific scoring matrices to predict amyloid propensity [70]. The data used to develop Waltz was expanded by an order of magnitude with 1089 experimentally and bibliographically obtained hexapeptides, leading to the development of Waltz-DB. Anew, this expanded dataset facilitated the development of new tools for predicting amyloid aggregation, including CORDAX [50]. The idea of predicting APRs was further refined by the IMPAcT [71], which leverages data from ZipperDB [72] to define the effect of point mutations on the amyloid propensity of proteins.

Correspondingly, the dataset of point mutations of Aβ obtained *in cellulo* [73], later included in AmyloBase, led to the development of AGGRESCAN [46]. Despite this starting point, the newest iteration of this algorithm, Aggrescan4D, predicts the general aggregation propensity of globular proteins considering their 3D structure and the pH of the solution [74,75].

Even very narrowly specialized datasets enable the development of predictive models. For example, AmyloGraph, a database of amyloid interactions, was used to develop AmyloComp, an algorithm to estimate the structural potential of two sequences to form heterogeneous amyloid fibrils [76] and PACT, dedicated to predicting cross-interactions between amyloid proteins [77].

5. Conclusions

The complexity of amyloid aggregation has led to a large information influx, which soon matured into databases. Experimental data and predictions, organized into structured resources, have triggered the development of bioinformatic algorithms for predicting amyloidogenicity and amorphous protein aggregation. These tools have, in turn, accelerated new experimental studies, which have provided data for expanding existing databases and enabling the development of new ones. This cyclic process has significantly improved the understanding of the physicochemical determinants that drive soluble proteins into

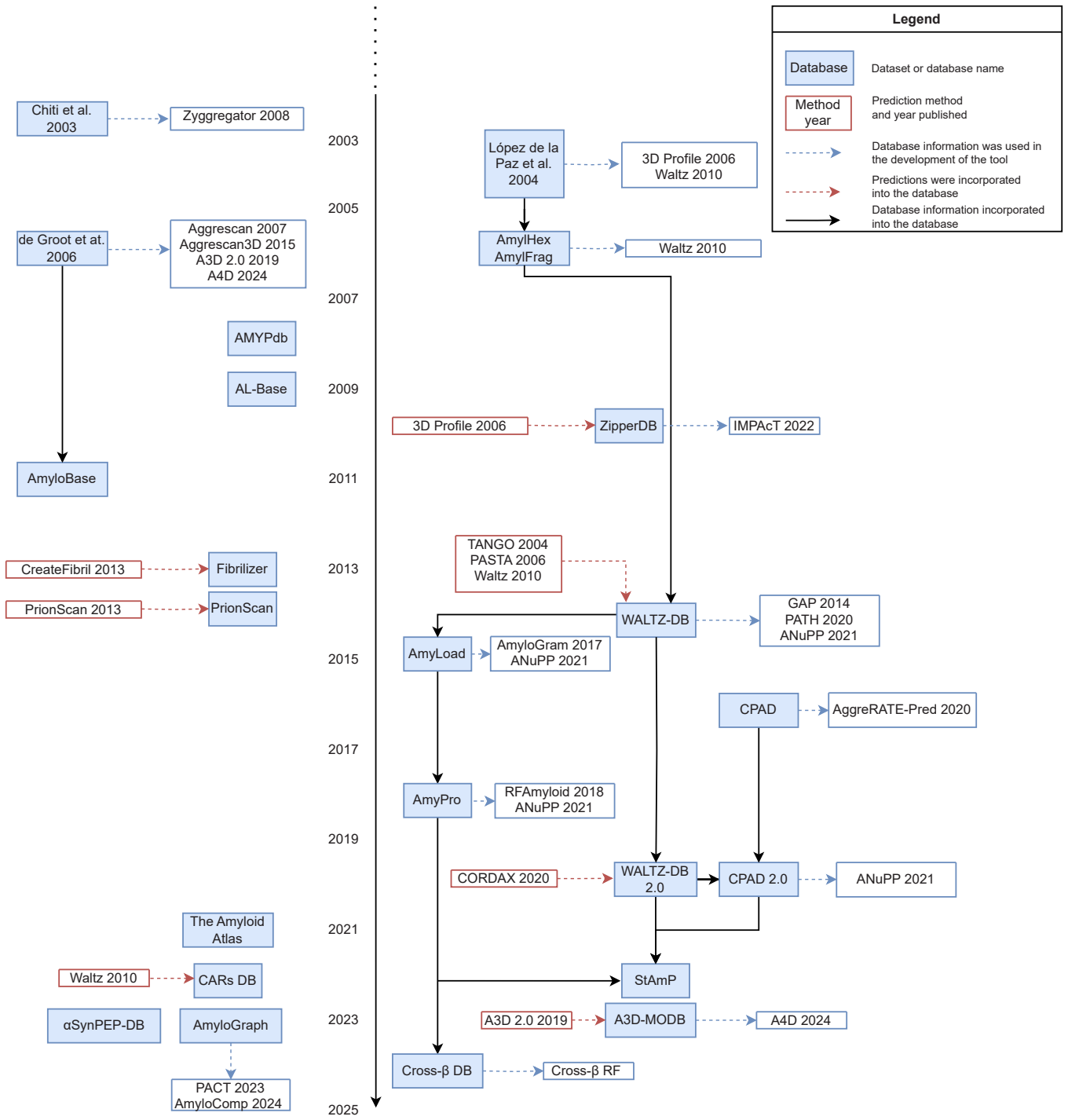

Fig. 2. Timeline of the intertwined relationship between aggregation datasets and databases and predictive tools. Protein aggregation datasets are differentiated from databases by including the first author and year of publication.

aggregates, ultimately leading to increased success in developing and formulating protein-based products and therapeutics, anti-aggregation therapies for amyloidosis, and the creation of novel technological applications. Acknowledging the challenges of identifying variables involved in each experiment, we believe that integrating the data stored in these resources will allow the development of highly accurate machine learning-assisted predictive methods, expanding our understanding of the physicochemical determinants that drive proteins into amyloid aggregates.

CRediT authorship contribution statement

Valentín Iglesias: Writing – review & editing, Writing – original draft, Visualization, Data curation, Conceptualization. **Salvador Ventura:** Writing – review & editing, Supervision, Conceptualization. **Michał Burdukiewicz:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Funding acquisition, Conceptualization. **Carlos Pintado-Grima:** Writing – review & editing, Writing – original draft, Conceptualization. **Oriol Bárcenas:** Writing – review & editing, Writing – original draft, Conceptualization. **Jarosław Chilimoniuk:** Writing – review & editing, Writing – original draft,

Visualization, Software, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Valentin Iglesias reports financial support was provided by Polish National Agency for Academic Exchange NAWA. Oriol Barcenas reports financial support was provided by Spanish Ministry of Science and Innovation. Carlos Pintado-Grima reports financial support was provided by Secretariat of Universities and Research of the Catalan Government. Carlos Pintado-Grima reports financial support was provided by European Social Fund. Salvador Ventura reports financial support was provided by Spanish Ministry of Science and Innovation, Catalan Institution for Research and Advanced Studies and Generalitat of Catalunya. Michal Burdukiewicz reports financial support was provided by National Science Centre Poland.

Acknowledgments

V.I. was supported by the Polish National Agency for Academic Exchange NAWA under the ULAM NAWA Programme [BPN/ULM/2023/1/00189/U/00001]; C.P.G. was supported by the Secretariat of Universities and Research of the Catalan Government and the European Social Fund [2023 FI_3 00018]; O.B. was supported by the Spanish Ministry of Science and Innovation via a doctoral grant [FPU22/03656]; S.V has been supported by the Spanish Ministry of Science and Innovation [PID2022–137963OB-I00], Generalitat de Catalunya [2021-SGR-00635 AGAUR] and by ICREA, [ICREA-Academia 2020]; M.B. acknowledges funding from the National Science Centre Poland SONATA 19 grant [2023/51/D/NZ7/02847]. Funding for open access charge: Medical University of Białystok [B.SUB.24.592].

References

[1] Sawaya MR, Hughes MP, Rodriguez JA, Riek R, Eisenberg DS. The expanding amyloid family: structure, stability, function, and pathogenesis. Cell Sep. 2021;184(19):4857–73. https://doi.org/10.1016/j.cell.2021.08.013.
[2] Rousseau F, Schymkowitz J, Serrano L. Protein aggregation and amyloidosis: confusion of the kinds? Curr Opin Struct Biol Feb. 2006;16(1):118–26. https://doi.org/10.1016/j.sbi.2006.01.011.
[3] Dobson CM. Protein folding and misfolding. Nature Dec. 2003;426(6968):884–90. https://doi.org/10.1038/nature02261.
[4] Chiti F, Dobson CM. Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. Annu Rev Biochem Jun. 2017;86:27–68. https://doi.org/10.1146/annurev-biochem-061516-045115.
[5] Santos J, Iglesias V, Ventura S. Computational prediction and redesign of aberrant protein oligomerization. Prog Mol Biol Transl Sci 2020;169:43–83. https://doi.org/10.1016/bs.pmbts.2019.11.002.
[6] Ventura S, et al. Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. Proc Natl Acad Sci May 2004;101(19):7258–63. https://doi.org/10.1073/pnas.0308249101.
[7] Buxbaum JN, et al. Amyloid nomenclature 2022: update, novel proteins, and recommendations by the International Society of Amyloidosis (ISA) Nomenclature Committee. Amyloid Oct. 2022;29(4):213–9. https://doi.org/10.1080/13506129.2022.2147636.
[8] Tilk S, Frydman J, Curtis C, Petrov D. Cancers adapt to their mutational load by buffering protein misfolding stress. eLife May 2023;12. https://doi.org/10.7554/eLife.87301.1.
[9] Liu S, et al. Highly efficient intercellular spreading of protein misfolding mediated by viral ligand-receptor interactions. Nat Commun Oct. 2021;12(1):5739. https://doi.org/10.1038/s41467-021-25855-2.
[10] Michiels E, Rousseau F, Schymkowitz J. Mechanisms and therapeutic potential of interactions between human amyloids and viruses. Cell Mol Life Sci CMLS Mar. 2021;78(6):2485–501. https://doi.org/10.1007/s00018-020-03711-8.
[11] Mukherjee A, Morales-Scheihing D, Butler PC, Soto C. Type 2 diabetes as a protein misfolding disease. Trends Mol Med Jul. 2015;21(7):439–49. https://doi.org/10.1016/j.molmed.2015.04.005.
[12] Moreno-Gonzalez I, Edwards Iii G, Salvadores N, Shahnawaz M, Diaz-Espinoza R, Soto C. Molecular interaction between type 2 diabetes and Alzheimer's disease through cross-seeding of protein misfolding. Mol Psychiatry 2017;22(9):1327–34. https://doi.org/10.1038/mp.2016.230.
[13] Chi EY, Krishnan S, Randolph TW, Carpenter JF. Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. Pharm Res Sep. 2003;20(9):1325–36. https://doi.org/10.1023/A:1025771421906.
[14] Roberts CJ. Protein aggregation and its impact on product quality. Curr Opin Biotechnol Dec. 2014;30:211–7. https://doi.org/10.1016/j.copbio.2014.08.001.
[15] Peña-Díaz S, Olsen WP, Wang H, Otzen DE. Functional amyloids: the biomaterials of tomorrow? Adv Mater Deerfield Beach Fla May 2024;36(18):e2312823. https://doi.org/10.1002/adma.202312823.
[16] Iconomidou VA, Vriend G, Hamodrakas SJ. Amyloids protect the silkmoth oocyte and embryo. FEBS Lett Aug. 2000;479(3):141–5. https://doi.org/10.1016/s0014-5793(00)01888-3.
[17] Antonets KS, et al. Accumulation of storage proteins in plant seeds is mediated by amyloid formation. PLoS Biol Jul. 2020;18(7):e3000564. https://doi.org/10.1371/journal.pbio.3000564.
[18] Otzen D, Riek R. Functional amyloids. Cold Spring Harb Perspect Biol Dec. 2019;11(12):a033860. https://doi.org/10.1101/cshperspect.a033860.
[19] Díaz-Caballero M, Navarro S, Fuentes I, Teixidor F, Ventura S. Minimalist prion-inspired polar self-assembling peptides. ACS Nano Jun. 2018;12(6):5394–407. https://doi.org/10.1021/acsnano.8b00417.
[20] Román-Álamo L, et al. Effect of the aggregated protein dye YAT2150 on Leishmania parasite viability. Antimicrob Agents Chemother Mar. 2024;68(3):e0112723. https://doi.org/10.1128/aac.01127-23.
[21] Wu G, et al. Enhanced therapeutic window for antimicrobial Pept-ins by investigating their structure-activity relationship. PloS One 2023;18(3):e0283674. https://doi.org/10.1371/journal.pone.0283674.
[22] Graña-Montes R, Pujols-Pujol J, Gómez-Picanyol C, Ventura S. Prediction of protein aggregation and amyloid formation. In: Rigden DJ, editor. in From Protein Structure to Function with Bioinformatics. Dordrecht: Springer Netherlands; 2017. p. 205–63. https://doi.org/10.1007/978-94-024-1069-3_7.
[23] Santos J, Pallarès I, Iglesias V, Ventura S. Cryptic amyloidogenic regions in intrinsically disordered proteins: function and disease association. Comput Struct Biotechnol J 2021;19:4192–206. https://doi.org/10.1016/j.csbj.2021.07.019.
[24] Pintado-Grima C, Santos J, Iglesias V, Manglano-Artuñedo Z, Pallarès I, Ventura S. Exploring cryptic amyloidogenic regions in prion-like proteins from plants. Front Plant Sci 2022;13:1060410. https://doi.org/10.3389/fpls.2022.1060410.
[25] Santos J, et al. pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity. Cells Jan. 2020;9(1):145. https://doi.org/10.3390/cells9010145.
[26] Roeters SJ, et al. Elevated concentrations cause upright alpha-synuclein conformation at lipid interfaces. Nat Commun Sep. 2023;14(1):5731. https://doi.org/10.1038/s41467-023-39843-1.
[27] Prusiner SB. Novel proteinaceous infectious particles cause scrapie. Science Apr. 1982;216(4542):136–44. https://doi.org/10.1126/science.6801762.
[28] Kraus A, Groveman BR, Caughey B. Prions and the potential transmissibility of protein misfolding diseases. Annu Rev Microbiol 2013;67:543–64. https://doi.org/10.1146/annurev-micro-092412-155735.
[29] Gil-Garcia M, Iglesias V, Pallarès I, Ventura S. Prion-like proteins: from computational approaches to proteome-wide analysis. FEBS Open Bio Sep. 2021;11(9):2400–17. https://doi.org/10.1002/2211-5463.13213.
[30] Tsiolaki PL, Nastou KC, Hamodrakas SJ, Iconomidou VA. Mining databases for protein aggregation: a review. Amyloid Int J Exp Clin Investig J Int Soc Amyloidosis Sep. 2017;24(3):143–52. https://doi.org/10.1080/13506129.2017.1353966.
[31] Wozniak PP, Kotulska M. AmyLoad: website dedicated to amyloidogenic protein fragments. Bioinformation Oct. 2015;31(20):3395–7. https://doi.org/10.1093/bioinformatics/btv375.
[32] Belli M, Ramazzotti M, Chiti F. Prediction of amyloid aggregation in vivo. EMBO Rep Jul. 2011;12(7):657–63. https://doi.org/10.1038/embor.2011.116.
[33] Burdukiewicz M, et al. AmyloGraph: a comprehensive database of amyloid-amyloid interactions. Nucleic Acids Res Jan. 2023;51(D1):D352–7. https://doi.org/10.1093/nar/gkac882.
[34] Varadi M, De Baets G, Vranken WF, Tompa P, Pancsa R. AmyPro: a database of proteins with validated amyloidogenic regions. Nucleic Acids Res Jan. 2018;46(D1):D387–92. https://doi.org/10.1093/nar/gkx950.
[35] Pintado-Grima C, et al. CARs-DB: a database of cryptic amyloidogenic regions in intrinsically disordered proteins. Front Mol Biosci May 2022;9. https://doi.org/10.3389/fmolb.2022.882160.
[36] Rawat P, Prabakaran R, Sakthivel R, Mary Thangakani A, Kumar S, Gromiha MM. CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides. Amyloid Apr. 2020;27(2):128–33. https://doi.org/10.1080/13506129.2020.1715363.
[37] Gonay V, Dunne MP, Caceres-Delpiano J, Kajava AV. BioRxiv. Dev Mach-Learn-Based Amyloid Predict Cross-Beta DB Feb. 14, 2024. https://doi.org/10.1101/2024.02.12.579644.
[38] Smaoui MR, Poitevin F, Delarue M, Koehl P, Orland H, Waldispühl J. Computational assembly of polymorphic amyloid fibrils reveals stable aggregates. Biophys J Feb. 2013;104(3):683–93. https://doi.org/10.1016/j.bpj.2012.12.037.
[39] 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 Feb. 2014;15(1):102. https://doi.org/10.1186/1471-2164-15-102.
[40] Louros N, van der Kant R, Schymkowitz J, Rousseau F. StAmP-DB: a platform for structures of polymorphic amyloid fibril cores. Bioinformation Apr. 2022;38(9):2636–8. https://doi.org/10.1093/bioinformatics/btac126.
[41] 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 Jan. 2020;48(D1):D389–93. https://doi.org/10.1093/nar/gkz758.

[42] PDB consortium. Protein data bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res Jan. 2019;47(D1):D520–8. https://doi.org/10.1093/nar/gky949.
[43] Aspromonte MC, et al. DisProt in 2024: improving function annotation of intrinsically disordered proteins. Nucleic Acids Res Jan. 2024;52(D1):D434–41. https://doi.org/10.1093/nar/gkad928.
[44] Thangakani AM, Nagarajan R, Kumar S, Sakthivel R, Velmurugan D, Gromiha MM. CPAD, curated protein aggregation database: a repository of manually curated experimental data on protein and peptide aggregation. PLoS ONE Apr. 2016;11(4): e0152949. https://doi.org/10.1371/journal.pone.0152949.
[45] 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 Oct. 2004;22(10):1302–6. https://doi.org/10.1038/nbt1012.
[46] 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 Feb. 2007;8:65. https://doi.org/10.1186/1471-2105-8-65.
[47] Walsh I, Seno F, Tosatto SCE, Trovato A. PASTA 2.0: an improved server for protein aggregation prediction (no. Web Server issue) Nucleic Acids Res Jul. 2014;42: W301–7. https://doi.org/10.1093/nar/gku399.
[48] Ahmed AB, Znassi N, Château M-T, Kajava AV. A structure-based approach to predict predisposition to amyloidosis. Alzheimers Dement Jun. 2015;11(6): 681–90. https://doi.org/10.1016/j.jalz.2014.06.007.
[49] Beerten J, et al. WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. Bioinformation May 2015;31(10):1698–700. https://doi.org/10.1093/bioinformatics/btv027.
[50] Louros N, Rousseau F, Schymkowitz J. CORDAX web server: an online platform for the prediction and 3D visualization of aggregation motifs in protein sequences. Bioinformation May 2024;40(5):btae279. https://doi.org/10.1093/bioinformatics/btae279.
[51] Martins PM, et al. MIRRAGGE - Minimum information required for reproducible aggregation experiments. Front Mol Neurosci 2020;13:582488. https://doi.org/10.3389/fnmol.2020.582488.
[52] Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet Jan. 2007;39(1):17–23. https://doi.org/10.1038/ng1934.
[53] Lill CM, et al. Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: the PDGene database. PLoS Genet 2012;8(3): e1002548. https://doi.org/10.1371/journal.pgen.1002548.
[54] Lill C, et al. Comprehensive research synopsis and systematic meta-analyses in ALS genetics: the ALSGene database (P01.095). P01.095-P01.095 Neurology Apr. 2012; 78(1_ement). https://doi.org/10.1212/WNL.78.1_supplement.P01.095.
[55] Dreses-Werringloer U, et al. A polymorphism in CALHM1 influences Ca2+ homeostasis, Abeta levels, and Alzheimer's disease risk. Cell Jun. 2008;133(7): 1149–61. https://doi.org/10.1016/j.cell.2008.05.048.
[56] Zerbinatti CV, et al. Oxysterol-binding protein-1 (OSBP1) modulates processing and trafficking of the amyloid precursor protein. Mol Neurodegener Mar. 2008;3:5. https://doi.org/10.1186/1750-1326-3-5.
[57] Bodi K, Prokaeva T, Spencer B, Eberhard M, Connors LH, Seldin DC. AL-Base: a visual platform analysis tool for the study of amyloidogenic immunoglobulin light chain sequences. Amyloid Int J Exp Clin Investig J Int Soc Amyloidosis Mar. 2009; 16(1):1–8. https://doi.org/10.1080/13506120802676781.
[58] Pintado-Grima C, et al. aSynPEP-DB: a database of biogenic peptides for inhibiting α-synuclein aggregation. Database Jan. 2023;2023:baad084. https://doi.org/10.1093/database/baad084.
[59] Santos J, Pallarès I, Ventura S. Is a cure for Parkinson's disease hiding inside us? Trends Biochem Sci Aug. 2022;47(8):641–4. https://doi.org/10.1016/j.tibs.2022.02.001.
[60] Santos J, Gracia P, Navarro S, Peña-Díaz S, Pujols J, Cremades N, Pallarès I, Ventura S. α-Helical peptidic scaffolds to target α-synuclein toxic species with nanomolar affinity. Nat Commun 2021;12(1):3752. https://doi.org/10.1038/s41467-021-24039-2.
[61] Badaczewska-Dawid AE, et al. A3D Model organism database (A3D-MODB): a database for proteome aggregation predictions in model organisms. Nucleic Acids Res Jan. 2024;52(D1):D360–7. https://doi.org/10.1093/nar/gkad942.
[62] 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 Jul. 2019;47(W1):W300–7. https://doi.org/10.1093/nar/gkz321.
[63] Mathur D, Kaur H, Dhall A, Sharma N, Raghava GPS. SAPdb: A database of short peptides and the corresponding nanostructures formed by self-assembly. Comput Biol Med Jun. 2021;133:104391. https://doi.org/10.1016/j.compbiomed.2021.104391.
[64] Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature Aug. 2003;424 (6950):805–8. https://doi.org/10.1038/nature01891.
[65] DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M. Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. J Mol Biol Aug. 2004;341(5):1317–26. https://doi.org/10.1016/j.jmb.2004.06.043.
[66] Bravard A, Sabatier L, Hoffschir F, Ricoul M, Luccioni C, Dutrillaux B. SOD2: a new type of tumor-suppressor gene? Int J Cancer May 1992;51(3):476–80. https://doi.org/10.1002/ijc.2910510323.
[67] Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. Chem Soc Rev Jul. 2008;37(7):1395–401. https://doi.org/10.1039/b706784b.
[68] López de la Paz M, Serrano L. Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci USA Jan. 2004;101(1):87–92. https://doi.org/10.1073/pnas.2634884100.
[69] 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 Mar. 2006;103(11):4074–8. https://doi.org/10.1073/pnas.0511295103.
[70] Maurer-Stroh S, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods Mar. 2010;7(3):237–42. https://doi.org/10.1038/nmeth.1432.
[71] Rosenberg GM, Murray KA, Salwinski L, Hughes MP, Abskharon R, Eisenberg DS. Bioinformatic identification of previously unrecognized amyloidogenic proteins. J Biol Chem May 2022;298(5). https://doi.org/10.1016/j.jbc.2022.101920.
[72] Sawaya MR, et al. Atomic structures of amyloid Cross-β Spines reveal varied steric zippers. Nature May 2007;447(7143):453–7. https://doi.org/10.1038/nature05695.
[73] de Groot NS, Aviles FX, Vendrell J, Ventura S. Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. FEBS J Feb. 2006;273(3):658–68. https://doi.org/10.1111/j.1742-4658.2005.05102.x.
[74] Bárcenas O, et al. Aggrescan4D: structure-informed analysis of pH-dependent protein aggregation. Nucleic Acids Res Jul. 2024;52(W1):W170–5. https://doi.org/10.1093/nar/gkae382.
[75] Zalewski M, Iglesias V, Bárcenas O, Ventura S, Kmiecik S. Aggrescan4D: a comprehensive tool for pH-dependent analysis and engineering of protein aggregation propensity. Protein Sci 2024;33(10):e5180. https://doi.org/10.1002/pro.5180.
[76] Bondarev SA, Uspenskaya MV, Leclercq J, Falgarone T, Zhouravleva GA, Kajava AV. AmyloComp: a bioinformatic tool for prediction of amyloid Co-aggregation. J Mol Biol Jan. 2024:168437. https://doi.org/10.1016/j.jmb.2024.168437.
[77] Wojciechowski JW, Szczurek W, Szulc N, Szefczyk M, Kotulska M. PACT - prediction of amyloid cross-interaction by threading. Sci Rep Dec. 2023;13(1): 22268. https://doi.org/10.1038/s41598-023-48886-9.