Nucleic Acids Research, 2024, **52**, W170–W175
https://doi.org/10.1093/nar/gkae382
Advance access publication date: 13 May 2024
Web Server issue

Aggrescan4D: structure-informed analysis of pH-dependent protein aggregation

Oriol Bárcenas [1], Aleksander Kuriata [2], Mateusz Zalewski [2], Valentín Iglesias [1,3], Carlos Pintado-Grima [1], Grzegorz Firlik [2], Michał Burdukiewicz [1,3], Sebastian Kmiecik [2,*] and Salvador Ventura [1,*]

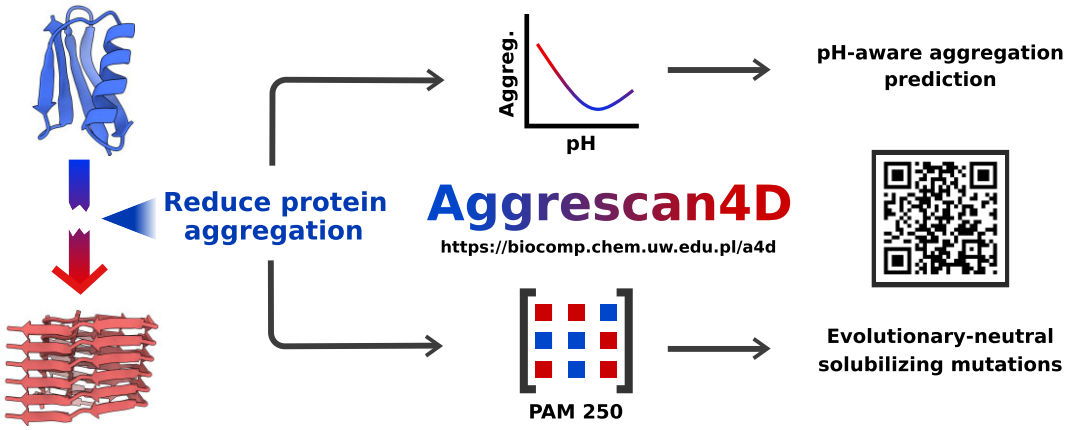
[1]Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
[2]Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland
[3]Clinical Research Centre, Medical University of Białystok, Kilińskiego 1, 15-369 Białystok, Poland

[*]To whom correspondence should be addressed. Tel: +34 93 586 8956; Fax: +34 93 581 2011; Email: salvador.ventura@uab.es
Correspondence may also be addressed to Sebastian Kmiecik. Tel: +48 22 8220211 (Ext. 310); Fax: +48 22 8220211 (Ext. 320); Email: sekmi@chem.uw.edu.pl

Abstract

Protein aggregation is behind the genesis of incurable diseases and imposes constraints on drug discovery and the industrial production and formulation of proteins. Over the years, we have been advancing the Aggresscan3D (A3D) method, aiming to deepen our comprehension of protein aggregation and assist the engineering of protein solubility. Since its inception, A3D has become one of the most popular structure-based aggregation predictors because of its performance, modular functionalities, RESTful service for extensive screenings, and intuitive user interface. Building on this foundation, we introduce Aggrescan4D (A4D), significantly extending A3D's functionality. A4D is aimed at predicting the pH-dependent aggregation of protein structures, and features an evolutionary-informed automatic mutation protocol to engineer protein solubility without compromising structure and stability. It also integrates precalculated results for the nearly 500,000 jobs in the A3D Model Organisms Database and structure retrieval from the AlphaFold database. Globally, A4D constitutes a comprehensive tool for understanding, predicting, and designing solutions for specific protein aggregation challenges. The A4D web server and extensive documentation are available at https://biocomp.chem.uw.edu.pl/a4d/. This website is free and open to all users without a login requirement.

Graphical abstract



Introduction

Protein aggregation is a multifactorial process, directed by the intrinsic properties of proteins (1) and heavily influenced by environmental factors (2). This phenomenon accounts for the onset of highly debilitating proteinopathies, posing a significant challenge to human health. Additionally, it hinders the development and implementation of protein-based biotechnological and biomedical applications (3).

Classical computational methods that predict protein aggregation rely on sequential information to assess the propensity of a protein to aggregate. These algorithms belong to four main categories: (i) heuristic detection of molecular determinants of aggregation (4–6), (ii) assessment of the conformational compatibility with amyloid-like architectures (7–9), (iii) consensus methods employing weighted predictions from other algorithms (10,11) and (iv) machine-learning

Received: March 11, 2024. **Revised:** April 11, 2024. **Editorial Decision:** April 27, 2024. **Accepted:** April 29, 2024

methods (12–15). While some of these algorithms provide a binary decision on whether a given sequence will aggregate, in most cases, they identify and weight the contribution of aggregation-prone regions (APRs), as is the case for Aggrescan (6). While highlighting APRs provides information on the regions driving aggregation, this strategy assumes that the proximity between residues is directly proportional to their sequential distance, which might apply for disordered or denatured protein states but is an imprecise approximation for the native state of globular proteins.

Being structure agnostic, these methods treat the aggregation contribution of amino acids residing in collapsed or inaccessible regions (such as the hydrophobic core) equally to exposed residues. To overcome these limitations, we developed a three-dimensional structure aggregation predictor, Aggrescan3D (A3D), which projects classical Aggrescan scoring to proteins' native structure, introducing the concept of STructural APRs (STAPs (16)), which correspond to the main drivers of protein aggregation in natively folded structures (17). To further enhance the biological relevance of its predictions, A3D incorporates CABS-flex (18,19). CABS-flex is a cost-efficient conformational sampling method designed to capture the inherent flexibility of proteins in solution. It accounts for the influence of multiple conformations on aggregation, moving beyond the limitations of a single, static structure description.

Building on these improvements, here we present Aggrescan4D (A4D), which adds a new dimension to protein aggregation prediction: environmental pH. This new variable contextualizes aggregation as a phenomenon modulated by environmental conditions. This paves the way to studying the relationship between pH and aggregation in biological contexts (20,21), while providing a tool to modulate protein aggregation in biotechnological setups. We have also added a new, evolutionary-based automatic mutation protocol to reduce protein aggregation while minimally impacting protein thermodynamic stability. Additionally, other user-centric improvements have been incorporated to streamline A4D use. Collectively, this update brings experimental results and computational predictions closer by providing a better description of the protein's biochemical context.

New features and updates

pH-dependent aggregation prediction

Proteins are, in general terms, evolutionarily selected to be soluble in their natural contexts (22), but the protein microenvironment (temperature, ions, or pH) can profoundly impact this capacity. This is highly relevant for engineered proteins (such as antibodies or enzymes), which have not undergone natural selection to ensure solubility under their intended conditions of use. Despite their importance, aggregation predictors have historically disregarded environmental conditions and focused solely on the proteins' physicochemical characteristics to model protein aggregation. In 2020, we inferred that environmental changes in pH impacted amino acids' charge and lipophilicity, and by applying a pH-dependent scale for proteinogenic amino acids (23), we fitted an empirical equation capable of capturing pH-dependent aggregation in intrinsically disordered proteins (IDPs) (24). This led us to develop SolupHred, the first disordered protein aggregation predictor that systematically incorporated the pH variable into its logic (25). Later in 2023, CamSol 3.0 adopted the same rationale to introduce pH-dependent solubility predictions for IDPs and globular proteins (26). Thus, we sought to adapt SolupHred's methodology to structured proteins by calculating structural pKa values with pKa-ANI, an algorithm based on deep representation learning (27), to predict the influence of pH on the aggregation of proteins in their folded configurations.

To model the influence of pH on aggregation propensity, we used two metrics: the maximum aggregation propensity ($A4D_{max}$) and the average aggregation propensity ($A4D_{avg}$). These two measures delineate two different aggregation regimes: $A4D_{max}$ represents the highest propensity observed across all residues and is better suited for proteins displaying defined STAPs. $A4D_{avg}$ represents the mean propensity across all residues and is more appropriate for polypeptides with an aggregation tendency that is more evenly distributed throughout the protein structure. Therefore, these two measures address the innate difference in which aggregation propensity is encoded in globular proteins and IDPs. $A4D_{max}$ is the most appropriate metric to describe pH-dependent aggregation for globular proteins, while $A4D_{avg}$ is better suited for IDPs. To aid in selecting the appropriate metric, we provide an objective criterion capturing the compactness of the protein structure for discriminating between ordered and disordered protein structures (Supplementary Information SI 1).

We assessed the performance of $A4D_{max}$ and $A4D_{avg}$ by analyzing data collected from 18 different experiments studying the impact of pH changes on the aggregation or solubility of 14 different proteins. Eleven of these experiments were conducted with globular proteins and 7 on IDPs (summarized in SI 2 and Supplementary Table S1; STs S2-S19 contain collected experimental data). These experiments used different metrics to measure the aggregation/solubility of the protein in the experimental setup, hereby referred to as experimental reporter. This umbrella term allows us to gather together aggregation and solubility measurements, such as Thioflavin T fluorescence (28) or solubility in PEG (29), respectively. We used linear regression to evaluate the relationship between the experimental reporters of aggregation/solubility and both A4D metrics.

The final benchmark covered $A4D_{max}$ and $A4D_{avg}$ performance. Although protein solubility cannot always be assumed to be the inverse of aggregation propensity, for the sake of completeness, we decided to include as well CamSol 3.0, a web server for the sequence-based prediction of pH-dependent protein solubility (26). A4D scores increase with a higher aggregation tendency. Conversely, as CamSol 3.0 predicts solubility, its scores should be lower for higher aggregation propensities.

To assess the algorithm's ability to discriminate the influence of pH on protein solubility/aggregation, we analyzed the concordance of the linear regression slope between the experimental reporters and the benchmarked metrics (summarized in SI3 and STs S20-S23).

The overall concordance of $A4D_{max}$, $A4D_{avg}$ and CamSol 3.0 were 78%, 89% and 78%, respectively. However, the degree of concordance varied depending on the proteins' conformation. Notably, for IDPs, both $A4D_{avg}$ and CamSol 3.0 demonstrated perfect concordance (100%); while, as expected, $A4D_{max}$ performed significantly worse, reaching only

43%. The situation is the opposite in the case of globular proteins. Here, $A4D_{max}$ reached a concordance of 100%, whereas $A4D_{avg}$ and CamSol 3.0 showed worse performance, 82% and 64%, respectively.

Thus, we have computed a novel structure-informed consensus measure, $A4D_{consensus}$, which considers the protein structure's compactness and based on that, selects the most appropriate metric for each query ($A4D_{max}$ or $A4D_{avg}$), reaching a 100% overall concordance in the evaluated experiments (Figure 1A). Even though $A4D_{max}$ is recommended for globular proteins, it is worth noting that $A4D_{avg}$ remains a valid metric across all scenarios, reporting an overall concordance of 89%. Details of the benchmark are provided in SI 3 (STs S24-S41 and Supplementary Figures S1- S18). A case study is presented in SI 4 and SF S19.

To cover all tools aimed at predicting pH-dependent protein aggregation, we integrated a benchmark of SolupHred with the disordered proteins used in the general benchmark. SolupHred, like $A4D_{avg}$ and CamSol 3.0, demonstrated 100% concordance on this dataset, something expected as it was specifically designed for IDPs.

Evolutionary-aware protein solubility improvement

The automated mutation mode to enhance solubility was introduced in A3D 2.0 (30). This method optimizes the input structure using FoldX (31) and ranks the most aggregation-prone residues according to their A3D score. Subsequently, the less soluble residues are mutated to charged amino acids (lysine, arginine, aspartic acid, or glutamic acid). While these mutations might turn optimal for reducing protein aggregation, they often induce significant changes in side chain physicochemical properties, which can negatively impact folding and stability, potentially disrupting protein function.

A4D provides an alternative to circumvent this limitation by applying a solubilizing strategy informed on the evolutionary conservation of point substitutions. Point accepted mutation (PAM) matrices store the sum of accepted substitutions in closely related sequences (identity ≥ 0.85), ensuring changes are a consequence of a single point mutation ($A \rightarrow B$) and not an accumulation of multiple mutations ($A \rightarrow C \rightarrow B$) (32,33). This translates to evolutionary favored changes that should have minimal impact on protein structure and stability. Despite this generalization, the server allows restraining the algorithm from mutating residues known by users as critical for maintaining protein function.

This protocol integrates evolutionary and aggregation information to suggest conservative mutations that reduce the aggregation propensity of the protein. First, conserved and neutral mutations are chosen based on the PAM250 substitution matrix (log-transformed PAM value ≥ 0). Out of these, only solubilizing mutations (ΔA4D score < 0) are considered. Finally, up to three mutations are suggested for each residue, prioritizing solubilization. We conducted a case study on bovine growth hormone (bGH) predicted structure (AF-A4GX95, residues 27–217). Despite sharing 90% sequential identity, whereas the human protein (hGH) is very soluble, bGH exhibits extreme aggregation across different experimental conditions (34). Notably, the analysis revealed different instances where the automated protocol's recommended mutations aligned with the residue identities found in hGH (V76T, Figure 1B). A comparison of this protocol and the protocol added in A3D 2.0 is provided in SI 5 and SF S20.

Other improvements

Beyond core functionalities, this update also contains numerous quality-of-life improvements. First, we expanded file compatibility to accept CIF/mmCIF file formats. Moreover, we have added seamless cross-database linking in the form of automatic AlphaFold structure fetching from UniProt ID (35) and A3D-Model Organisms DataBase (A3D-MODB) (36) entry retrieval. A3D-MODB includes precomputed predictions for over 164k proteins across 12 model organisms. Avoiding unnecessary calculations speeds up the tool and reduces its environmental impact, especially for recurrent jobs or training scenarios (37).

An essential consideration when using protein structures as inputs is the presence of long, disordered regions in proteins from the AlphaFold database, which are absent in most experimentally resolved protein structures. These disordered regions are occupying (pseudo)-random positions, generally close to the globular core (if it exists) (38). The static nature of these models leads to the generation of artificial surface regions stemming from the arbitrary and fixed position of disordered regions, which might lead to artifactual results, such as protecting or generating STAPs (36). To mitigate this, AlphaCutter is implemented as a pre-processing option, generating protein structures containing only globular regions (39).

In addition to these technical improvements, we have undertaken a thorough redesign of our user interface to improve the tool's usability and user experience. Moreover, A4D now includes a selection of example jobs showcasing the offered functionalities, complementing the detailed tutorial and FAQ sections already present in the server.

Limitations

We acknowledge that limitations associated with A4D pH-dependent aggregation predictions may impact the overall performance of the tool and subsequent benchmark.

First and foremost, the scarcity of well-annotated available data has been one of the most limiting factors for this work. Although we aimed to amend this limitation by enriching existing datasets with our unpublished data, we still cover only 14 different proteins, representing a considerably small fraction of the protein landscape. A second limiting factor was the high diversity of experimental reporters of protein aggregation provided by the authors. Given this problem, we interchangeably use measures of aggregation and solubility. Third, the heterogeneity of data sources has also contributed to the noise in the performance assessment (40).

Discussion and conclusions

Before the advent of bioinformatic tools, understanding and modifying the aggregation tendency of proteins was a very challenging and expensive task. For instance, to produce soluble, protein-based pharmaceuticals, techniques such as phage display (41) were often used. Services like A4D help users understand the causes behind aggregation in their proteins and assist in devising strategies to prevent it. While this behavior is primarily encoded in a protein's sequence and structure, other factors such as protein or crowding agent concentration, temperature, or pH can heavily influence the aggregation of polypeptides. Recognizing the importance of pH in

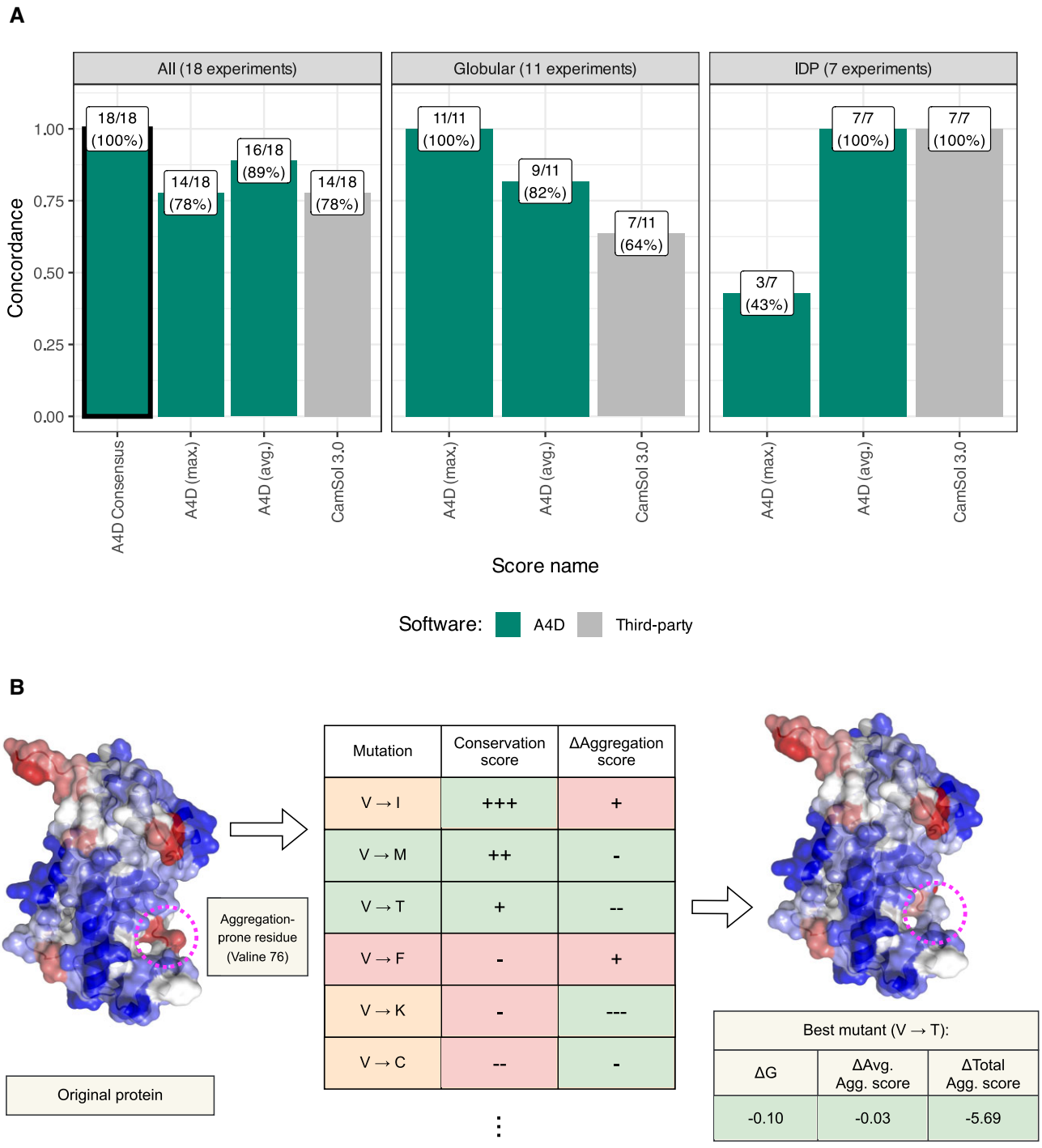


Figure 1. (**A**) Benchmark results for A4D (green) and CamSol 3.0 (gray) pH predictions. Results are presented jointly for all proteins and separately for globular proteins and IDPs. (**B**) Schematic representation of the evolutionary-aware solubility improvement protocol applied to bovine growth hormone. Mutations are suggested for the protein's most aggregation-prone residues, considering their relative conservation and aggregation score. If the new residue is predicted to be conserved or neutral and improve the solubility of the protein, it is considered for mutation (green). On the contrary, if the residue is not conserved or does not reduce the aggregation potential (orange) or both (red), it is not considered.

modulating protein aggregation, we updated our existing algorithm, A3D, to model this effect. This represents a categorical shift, moving beyond protein sequence (Aggrescan) and structure (Aggrescan3D) analyses, to provide a more comprehensive vision of protein aggregation. To our knowledge, SolupHred (2020) (25), and CamSol 3.0 (2023) (26) are the only existing resources that systematically model the effect of pH on solubility/aggregation of protein structures. This limited landscape reflects the difficulty of modeling the contribution of solvent conditions to protein aggregation, and the scarcity of available experimental data. We also introduce a novel protocol that suggests mild, evolutionary-neutral mutations to solubilize and, eventually, stabilize proteins without requiring a deep understanding of the protein at hand. A4D provides a platform for researchers to optimize their experimental conditions and generate more consistent data,

which can feed back the present algorithm and trigger the development of novel computational tools and standardized benchmarks.

A3D has been successfully applied in the redesign of less aggregation-prone variants of biotherapeutics, engineering novel self-assembled nanomaterials, understanding of pathological and physiological protein aggregation, and as a tool for teaching protein aggregation in university courses (42). While mutation protocols are the most straightforward and potent to prevent protein aggregation, they may face limitations such as the impossibility of mutating aggregation-prone but functionally important residues, like CDRs in antibodies or the existence of intellectual protection issues, like in the case of biosimilars. Thus, modifying the aggregation behavior of proteins without altering the protein sequence itself is key in many biotechnological and medical applications. Given the new functionalities implemented in this update, we anticipate that A4D will become a widely used tool within the protein science community.

Data availability

All data on the predicted aggregation propensity supporting the findings of this study are included within this paper and its Supplementary Information files.

Supplementary data

Supplementary Data are available at NAR Online.

Funding

O.B. was supported by the Spanish Ministry of Science and Innovation via a doctoral grant [FPU22/03656]; V.I. was supported by Spanish Ministry of Universities; European Union-Next Generation EU [ruling 02/07/2021, Universitat Autònoma de Barcelona]; 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]; M.B. was supported by the Maria Zambrano grant funded by the European Union-NextGenerationEU; S.V. was supported by Spanish Ministry of Science and Innovation [PID2022-137963OB-I00]; ICREA, ICREA-Academia 2020; A.K., M.Z. and S.K. acknowledge funding from the National Science Centre, Poland [Sheng 2021/40/Q/NZ2/00078]. Funding for open access charge: National Science Centre, Poland [Sheng 2021/40/Q/NZ2/00078 and PID2022-137963OB-I00].

Conflict of interest statement

None declared.

References

1. López de la Paz,M. and Serrano,L. (2004) Sequence determinants of amyloid fibril formation. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 87–92.
2. Owen,M.C., Gnutt,D., Gao,M., Wärmländer,S.K.T.S., Jarvet,J., Gräslund,A., Winter,R., Ebbinghaus,S. and Strodel,B. (2019) Effects of in vivo conditions on amyloid aggregation. *Chem. Soc. Rev.*, **48**, 3946–3996.
3. Chiti,F. and Dobson,C.M. (2017) Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.*, **86**, 27–68.
4. Fernandez-Escamilla,A.-M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.
5. Tartaglia,G.G. and Vendruscolo,M. (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.*, **37**, 1395–1401.
6. Conchillo-Solé,O., de Groot,N.S., Avilés,F.X., Vendrell,J., Daura,X. and Ventura,S. (2007) AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65.
7. Garbuzynskiy,S.O., Lobanov,M.Y. and Galzitskaya,O.V. (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics (England)*, **26**, 326–332.
8. Ahmed,A.B., Znassi,N., Château,M.-T. and Kajava,A.V. (2015) A structure-based approach to predict predisposition to amyloidosis. *Alzheimers Dementia*, **11**, 681–690.
9. Louros,N., Orlando,G., De Vleeschouwer,M., Rousseau,F. and Schymkowitz,J. (2020) Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nat. Commun.*, **11**, 3314.
10. Tsolis,A.C., Papandreou,N.C., Iconomidou,V.A. and Hamodrakas,S.J. (2013) A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. *PloS One*, **8**, e54175.
11. Emily,M., Talvas,A. and Delamarche,C. (2013) MetAmyl: a METa-predictor for AMYLoid proteins. *PloS One*, **8**, e79722.
12. Gasior,P. and Kotulska,M. (2014) FISH Amyloid - a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids. *BMC Bioinformatics*, **15**, 54.
13. Família,C., Dennison,S.R., Quintas,A. and Phoenix,D.A. (2015) Prediction of peptide and protein propensity for amyloid formation. *PloS One*, **10**, e0134679.
14. Burdukiewicz,M., Sobczyk,P., Rödiger,S., Duda-Madej,A., Mackiewicz,P. and Kotulska,M. (2017) Amyloidogenic motifs revealed by n-gram analysis. *Sci. Rep.*, **7**, 12961.
15. Charoenkwan,P., Ahmed,S., Nantasenamat,C., Quinn,J. M.W., Moni,M.A., Lio',P. and Shoombuatong,W. (2022) AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning. *Sci. Rep.*, **12**, 7697.
16. Santos,J., Pujols,J., Pallarès,I., Iglesias,V. and Ventura,S. (2020) Computational prediction of protein aggregation: advances in proteomics, conformation-specific algorithms and biotechnological applications. *Comput. Struct. Biotechnol. J.*, **18**, 1403–1413.
17. Zambrano,R., Jamroz,M., Szczasiuk,A., Pujols,J., Kmiecik,S. and Ventura,S. (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.*, **43**, W306–W313.
18. Jamroz,M., Orozco,M., Kolinski,A. and Kmiecik,S. (2013) Consistent view of protein fluctuations from all-atom molecular dynamics and coarse-grained dynamics with knowledge-based force-field. *J. Chem. Theory Comput.*, **9**, 119–125.
19. Kuriata,A., Gierut,A.M., Oleniecki,T., Ciemny,M.P., Kolinski,A., Kurcinski,M. and Kmiecik,S. (2018) CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res.*, **46**, W338–W343.
20. Horváth,D., Dürvanger,Z., K.Menyhárd,D., Sulyok-Eiler,M., Bencs,F., Gyulai,G., Horváth,P., Taricska,N. and Perczel,A. (2023) Polymorphic amyloid nanostructures of hormone peptides involved in glucose homeostasis display reversible amyloid formation. *Nat. Commun.*, **14**, 4621.
21. Raposo,G., Tenza,D., Murphy,D.M., Berson,J.F. and Marks,M.S. (2001) Distinct protein sorting and localization to premelanosomes, melanosomes, and lysosomes in pigmented melanocytic cells. *J. Cell Biol.*, **152**, 809–824.
22. Monsellier,E. and Chiti,F. (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.*, **8**, 737–742.

23. Zamora,W.J., Campanera,J.M. and Luque,F.J. (2019) Development of a structure-based, ph-dependent lipophilicity scale of amino acids from continuum solvation calculations. *J. Phys. Chem. Lett.*, **10**, 883–889.
24. Santos,J., Iglesias,V., Santos-Suárez,J., Mangiagalli,M., Brocca,S., Pallarès,I. and Ventura,S. (2020) pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity. *Cells*, **9**, 145.
25. Pintado,C., Santos,J., Iglesias,V. and Ventura,S. (2021) SolupHred: a server to predict the ph-dependent aggregation of intrinsically disordered proteins. *Bioinformatics*, **37**, 1602–1603.
26. Oeller,M., Kang,R., Bell,R., Ausserwöger,H., Sormanni,P. and Vendruscolo,M. (2023) Sequence-based prediction of pH-dependent protein solubility using camsol. *Brief. Bioinform.*, **24**, bbad004.
27. Gokcan,H. and Isayev,O. (2022) Prediction of protein pKa with representation learning. *Chem. Sci.*, **13**, 2462–2474.
28. Gade Malmos,K., Blancas-Mejia,L.M., Weber,B., Buchner,J., Ramirez-Alvarado,M., Naiki,H. and Otzen,D. (2017) ThT 101: a primer on the use of thioflavin T to investigate amyloid formation. *Amyloid*, **24**, 1–16.
29. Oeller,M., Sormanni,P. and Vendruscolo,M. (2021) An open-source automated PEG precipitation assay to measure the relative solubility of proteins with low material requirement. *Sci. Rep.*, **11**, 21932.
30. Kuriata,A., Iglesias,V., Pujols,J., Kurcinski,M., Kmiecik,S. and Ventura,S. (2019) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res.*, **47**, W300–W307.
31. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
32. Dayhoff,M.O. (1972) A model of evolutionary change in proteins. *Atlas Protein Sequence Struct.*, **5**, 89–99.
33. Liò,P. and Goldman,N. (1998) Models of molecular evolution and phylogeny. *Genome Res.*, **8**, 1233–1244.
34. Dellacha,J.M., Santomĕ,J.A. and Paladini,A.C. (1968) Physicochemical and structural studies of bovine growth hormone. *Ann. NY Acad. Sci. U.S.A.*, **148**, 313–327.
35. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
36. Badaczewska-Dawid,A.E., Kuriata,A., Pintado-Grima,C., Garcia-Pardo,J., Burdukiewicz,M., Iglesias,V., Kmiecik,S. and Ventura,S. (2024) A3D model organism database (A3D-MODB): a database for proteome aggregation predictions in model organisms. *Nucleic Acids Res.*, **52**, D360–D367.
37. Lucivero,F. (2020) Big data, big waste? a reflection on the environmental sustainability of big data initiatives. *Sci. Eng. Ethics*, **26**, 1009–1030.
38. Ruff,K.M. and Pappu,R.V. (2021) AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.*, **433**, 167208.
39. Tam,C. and Iwasaki,W. (2023) AlphaCutter: efficient removal of non-globular regions from predicted protein structures. *Proteomics*, **23**, e2300176.
40. Landrum,G.A. and Riniker,S. (2024) Combining IC50 or Ki values from different sources is a source of significant noise. *J. Chem. Inf. Model.*, **64**, 1560–1567.
41. Sidhu,S.S. (2000) Phage display in pharmaceutical biotechnology. *Curr. Opin. Biotechnol.*, **11**, 610–616.
42. Pintado-Grima,C., Bárcenas,O., Bartolomé-Nafría,A., Fornt-Suñé,M., Iglesias,V., Garcia-Pardo,J. and Ventura,S. (2023) A review of fifteen years developing computational tools to study protein aggregation. *Biophysica*, **3**, 1–20.

Received: March 11, 2024. **Revised:** April 11, 2024. **Editorial Decision:** April 27, 2024. **Accepted:** April 29, 2024