Nucleic Acids Research, 2024, **52**, D360–D367
https://doi.org/10.1093/nar/gkad942
Advance access publication date: 28 October 2023
Database issue

A3D Model Organism Database (A3D-MODB): a database for proteome aggregation predictions in model organisms

Aleksandra E. Badaczewska-Dawid[1], Aleksander Kuriata[2], Carlos Pintado-Grima[3], Javier Garcia-Pardo[3], Michał Burdukiewicz[3,4], Valentín Iglesias[3,*], Sebastian Kmiecik[2,*] and Salvador Ventura[3,*]

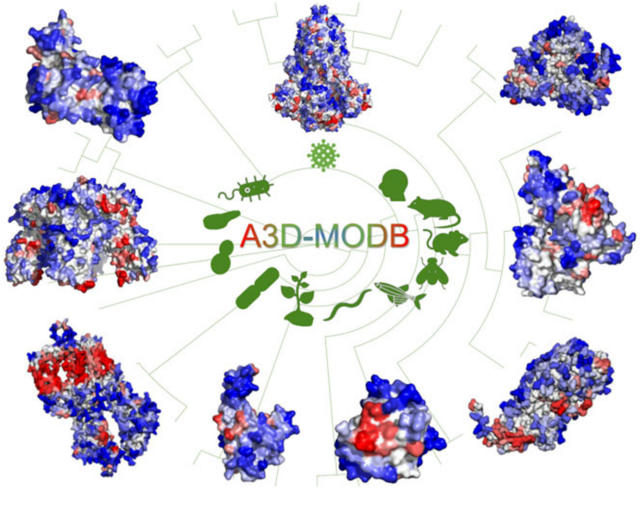
[1]Genome Informatics Facility, Office of Biotechnology, Iowa State University, Ames 50011 IA, USA
[2]Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland
[3]Institut de Biotecnologia i de Biomedicina (IBB) and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
[4]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
Correspondence may also be addressed to Valentín Iglesias. Tel: +34 93 581 2154; Fax: +34 93 581 2011; Email: valentin.iglesias@uab.cat

Abstract

Protein aggregation has been associated with aging and different pathologies and represents a bottleneck in the industrial production of biotherapeutics. Numerous past studies performed in *Escherichia coli* and other model organisms have allowed to dissect the biophysical principles underlying this process. This knowledge fuelled the development of computational tools, such as Aggrescan 3D (A3D) to forecast and re-design protein aggregation. Here, we present the A3D Model Organism Database (A3D-MODB) http://biocomp.chem.uw.edu.pl/A3D2/MODB, a comprehensive resource for the study of structural protein aggregation in the proteomes of 12 key model species spanning distant biological clades. In addition to A3D predictions, this resource incorporates information useful for contextualizing protein aggregation, including membrane protein topology and structural model confidence, as an indirect reporter of protein disorder. The database is openly accessible without any need for registration. We foresee A3D-MOBD evolving into a central hub for conducting comprehensive, multi-species analyses of protein aggregation, fostering the development of protein-based solutions for medical, biotechnological, agricultural and industrial applications.

Graphical abstract



Introduction

Protein aggregation has moved from an anecdotal laboratory phenomenon to its recognition as a key contributing factor in various chronic disorders (1). Moreover, protein aggregation is a common industrial bottleneck for the production, purification and storage of protein-based biotherapeutics (2,3), including antibody production and manufacturing of protein-based nanoparticles (4). In this context, decades of studying protein aggregation in *Escherichia coli*, *Saccharomyces cerevisiae* and other model organisms (5–8) contributed to a

Received: August 14, 2023. **Revised:** September 27, 2023. **Editorial Decision:** October 11, 2023. **Accepted:** October 11, 2023

robust knowledge of its underlying mechanisms and has spurred the development of computational algorithms capable of predicting protein aggregation *in silico* (9).

Computational tools designed to study protein aggregation had remarkable success in anticipating aggregation-prone regions (APRs) in intrinsically disordered proteins (IDPs), where aggregation is intricately tied to their sequence (9). However, for native globular proteins, these methods presented two primary limitations, as they (i) often overpredict the contribution of collapsed hydrophobic regions that are not exposed to solvents or transmembrane stretches, both of which have minimal impact on protein aggregation in native states/environments and (ii) base their predictions in sequential contiguity, overlooking the effects of residues that are spatially close, but sequentially remote. In response to these challenges, prediction methods that use the spatial information to identify structural aggregation-prone regions (STAPs) in proteins were introduced (10–13). In 2015, we developed Aggrescan 3D (A3D) (11) by spatially correcting the *in vivo*-derived Aggrescan aggregation propensity scale to focus on solvent-accessible residues (14,15). A3D features i) a dynamic mode to capture protein flexibility fluctuations using coarse-grained MD-simulations from CABS-Flex (16–18), ii) stability evaluations using the FoldX force field (19), (iii) a RESTful mode which allows programmatic access intended for large studies and iv) an automated tool to facilitate protein engineering for increased solubility, preserving or improving thermodynamic stability. A3D has allowed the redesign of more soluble and stable variants of protein chromophores (20), enzymes (21) as well as human antibodies (20). However, the limited availability of experimental protein structures constrained the potential applications of A3D.

The recent AlphaFold revolution (22,23) and subsequent release of the AlphaFold database (24) have facilitated access to highly accurate computed protein structures on a proteomic scale. Leveraging these data, we previously performed proteome-wide aggregation analysis for the entire human and yeast structuromes, providing a wealth of practical examples and releasing a repository of our findings (25,26). These resources represented a first step towards studying protein aggregation in large structural datasets. Moreover, these precomputed A3D calculations allowed diminishing the carbon footprint associated with redundant individual computational queries (27).

Herein, we constructed the A3D Database for Model Organisms (A3D-MODB) (http://biocomp.chem.uw.edu.pl/A3D2/MODB). This database encompasses over 500 000 structural predictions for over 160 000 proteins covering 12 model species including *Arabidopsis thaliana, Caenorhabditis elegans, Danio rerio, Escherichia coli, Mycoplasma genitalium, Mus musculus, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Homo sapiens, Rattus norvegicus, Drosophila melanogaster* and SARS-CoV-2. Besides providing A3D results on solubility and stability, the A3D-MODB incorporates an additional wealth of information on membrane proteins' topology by incorporating TOPCONS (28) and the per-residue model confidence metric (pLDDT) provided by AlphaFold. Moreover, each A3D-MODB entry has been linked to UniProt Protein Knowledgebase (UniProt KB) (29) and organism-specific gold standard databases such as the Human Protein Atlas, EchoBASE, MGD, SGD, Wormbase, FlyBase, RGD, TAIR, ZFIN, PomBase, WholeCellKB-MG (30–39). To our knowledge, A3D-MODB represents the most comprehensive resource available for structure-based aggregation predictions and is designed to significantly advance protein aggregation research at the proteomic level.

Materials and methods

Species selection and data collection

The model organisms included in the present study were selected based on the frequency of the predictions submitted to A3D (identifiable by UniProt Accession numbers (UniProt Acc.)) since the release of AlphaFold database, while ensuring inclusion of representatives from distant clades. All available protein models for each species (Figure 1 and Supplementary Table S1) were downloaded from the AlphaFold database (5 June 2023; structural model version v4), and previously existing resources for human and yeast were updated to the latest available models (5 June 2023; structural model version v4). AlphaFold database models cover most of each species' proteome except for proteins shorter than 16 or longer than 2700 residues. For human proteins longer than 2700 residues, the AlphaFold database provides overlapping fragments that can be traced in A3D-MODB. SARS-CoV-2 models were not available in the AlphaFold database. To ensure consistent results across all species, we employed the AlphaFold2 tool (alphafold/2.3.1 standalone) to predict the complete SARS-CoV-2 protein sequences. We then used the relaxed structure of the highest-ranked AlphaFold prediction for the subsequent analysis.

A3D analysis

Structural aggregation predictions were performed with the latest implementations of A3D (A3D 2.0) (40,41) in static mode with default settings: distance of aggregation of 10Å and FoldX force field energy minimization preceding the analysis (19). Custom analyses 'c50' and 'c70' were run for all predicted structures with defined AlphaFold pLDDT cutoffs of 50 and 70, respectively, to focus the study on globular domains. In these cases, all the residues with pLDDT $\leq$ 50 (indicating very low confidence and likely disorder) or $\leq$ 70 (indicating low model confidence) were removed before performing the A3D predictions.

Database construction

The A3D Model Organism online database's user interface was created using HTML and enhanced with custom JavaScript functions for increased interactivity. The websites visual design combines standard Bootstrap components with custom CSS styles. The website is hosted on an Apache2 web server, which uses MySQL integration for data storage, retrieval and querying of pre-calculated A3D entries for each organism. Interactive plots are dynamically created using the D3.js library, while molecular visualization is handled by the PyMOL tool (42). The Python library Bokeh is used for interactive data visualizations. A3D-MODB entries are seamlessly integrated with the A3D 2.0 server, allowing users direct submission of custom mutation analysis.

Transmembrane proteins domain topologies

A3D-MODB incorporates prediction of transmembrane regions for membrane proteins annotated as trans- or

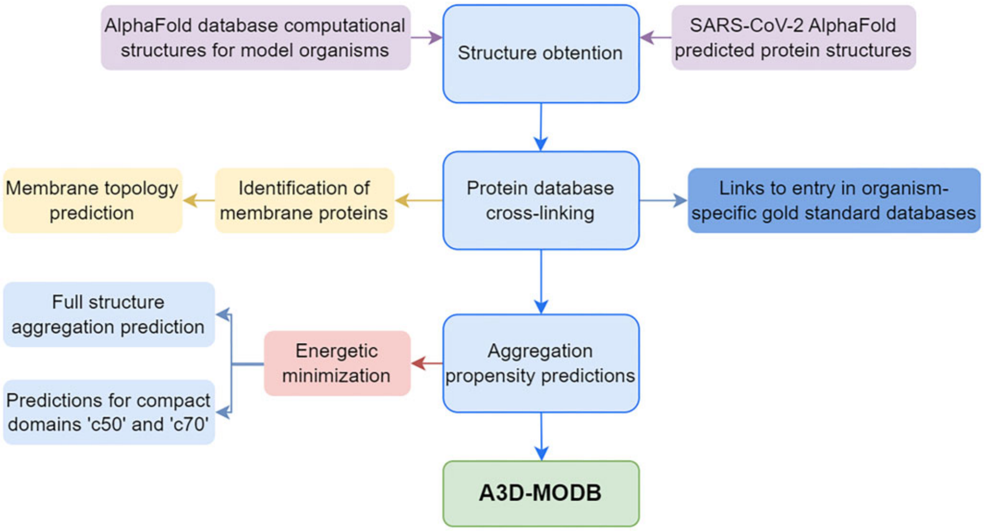

AlphaFold database computational structures for model organisms
Structure obtention
SARS-CoV-2 AlphaFold predicted protein structures
Membrane topology prediction
Identification of membrane proteins
Protein database cross-linking
Links to entry in organism-specific gold standard databases
Full structure aggregation prediction
Energetic minimization
Aggregation propensity predictions
Predictions for compact domains 'c50' and 'c70'
A3D-MODB

Figure 1. A3D-MODB construction pipeline. AlphaFold version 4 structures were obtained from AlphaFold database (24) for all organisms except SARS-CoV-2 for which AlphaFold2 standalone tool (22) was used to predict the complete SARS-CoV-2 protein sequences. Protein structural models were next linked to their entries on organism-specific gold standard databases: the Human Protein Atlas, EchoBASE, MGD, SGD, Wormbase, FlyBase, RGD, TAIR, ZFIN, PomBase, WholeCellKB-MG (30–38). UniProt Protein Knowledgebase (29) annotated membrane proteins were subject to topological prediction applying TOPCONS consensus algorithm (28) which integrates output from OCTOPUS (43), Philius (44), Polyphobius (45), SCAMPI (46) and SPOCTOPUS (47) predictive approaches. Protein models were energetically minimized using FoldX force field (19) before running the aggregation prediction on the latest version of A3D (40,41) for the whole protein and based on AlphaFold's pLDDT information, on the 'c50' and 'c70' models.

intra- membrane in UniProt KB using the TOPCONS consensus algorithm (28) and lists predicted topologies from OCTOPUS (43), Philius (44), Polyphobius (45), SCAMPI (46) and SPOCTOPUS (47) bioinformatics tools.

Case study

Receptor identities were obtained from the original publication where available (48). Non-reported orthologs were obtained by BLASTing UniProt accession against the reference proteome in UniProt KB. All blast-obtained proteins correspond to Swiss-Prot entries. Human protein α-helices mediating dimerization for SR1 and CTE-binding for SR2 were sourced from Tanenbaum's study (49). The corresponding structural motifs were mapped in orthologues by first performing a sequence alignment against human receptors in the UniProt KB and then manually curating helix stretches. Structural alignments were performed with Pymol.

Results

Data content

A3D Model Organism Database includes over 160 000 individual entries for 12 organisms covering distant clades and selected for their actual or potential relevance in protein aggregation studies. The inclusion of custom jobs 'c50' and 'c70' increases the precalculated A3D jobs available in the A3D-MODB to over 500 000. The database provides different integrated tools to facilitate studying how aggregation has moulded individual protein evolution or constrained the functionalities and structures of different protein families.

A3D-MODB features

Understanding structural aggregation

A3D provides rapid and detailed protein aggregation propensity predictions by transferring Aggrescan's *E. coli in vivo* derived scale (14,15) into the spatial information of experimentally-determined protein structures (reviewed in (50,51)). Switching from the primary sequence to the atomic 3D coordinates allows A3D to detect structural aggregation prone regions (STAPs) even when residues are not contiguous in sequence and to exclude solvent-protected hydrophobic regions that make a negligible contribution to the aggregation process (50,51). The web server includes additional features for rationalizing and re-engineering protein solubility through a user-friendly graphical interface or programmatic access through its REST API. Computed structural models require a more meticulous analysis as they might include intrinsically disordered regions (IDRs), stretches which are flexible and therefore mostly absent from crystal structures. IDRs occupy space almost randomly, and this dynamic nature is not represented in AlphaFold models as it depicts a single static frame of the conformational ensembles. Employing these models without accounting for this aspect could skew the prediction outcome, which became evident after manually curating 100 models from the human proteome (25). To mitigate this impact, A3D-MODB incorporates pre-run custom jobs 'c50' for residues with model confidence scores 'pLDDT' > 50 and 'c70' for those with pLDDT > 70, focusing on the protein globular domains and leaving IDRs aside. Moreover, the platform allows users to set their preferred model confidence threshold and perform the prediction on top of the resulting protein regions.

Estimating mutational impact on protein stability

Mutations that result in thermodynamically unfavourable variants usually provoke non-functional, non-expressed or short-lived polypeptides (20) which can lead to pathological conditions (1). A3D-MODB incorporates the FoldX force field to i) minimize the energetic state of the initial protein model by repairing bad torsion angles or Van der Waals' clashes before the aggregation prediction and ii) evaluate energetic variation upon user-designed mutations, which can be easily run through the ‘Custom jobs’ tab.

Membrane protein characterisation

A3D-MODB identifies proteins annotated in UniProt KB as membranous and performs topological predictions using TOPCONS consensus tool (28) along with the individual results from OCTOPUS (43), Philius (44), Polyphobius (45), SCAMPI (46) and SPOCTOPUS (47).

Programmatic access

A3D Model Organism Database includes the RESTful application program interface (API) already present in A3D. It allows 60 different queries per second per IP address and retrieves results in a machine-interpretable JSON file. Example usage in Bash and Python programming language are provided in the ‘Tutorial’ and described elsewhere (51). Programmatic access is most suitable for large-scale projects where retrieving results case-by-case would be tedious and time-consuming. Its application in a resource such as A3D-MODB is intended to aid complete proteome analyses or extensive evolutionary studies.

Usage of A3D-MODB

The A3D-MODB main page presents an upper panel which links to the latest version of A3D web server, its queue and a redirection to the database home screen. On the right margin, clickable links direct users to the ‘About’ section which holds documentation on the database usage, the A3D method and a brief explanatory video. ‘Cite’ redirects to a repository with related publications and ‘Download’ links a page with pre-computed comma-separated (CSV) files for each species, containing job ids for all processed UniProt Accession numbers (UniProt Acc.). In the central part, clickable links to model organism-specific databases are shown.

Searching the databases

Each species home page conserves the aforementioned general options and provides a per-organism search bar. A3D-MODB accepts UniProt Acc., gene or protein name along with organism-specific identifiers as query ID. Five example queries are provided below the search bar. Search results are retrieved as a list of records matching different query categories. Clicking the desired identifier will redirect users to the individual entry record.

Individual entry record

A3D-MODB organizes the results and features for each record in specific tabs:

- ‘Project details’ contains fundamental information about the entry and link to the external resources: it’s UniProt Acc., Gene identifiers, protein name, taxonomy and link to organism-specific database and if available, deposited experimental structures in PDB. Moreover, it includes details regarding the processed structure: the region considered, its sequence and the distance of analysis. Custom jobs with pLDDT >50 or >70 display the criteria-meeting sequence only, and for greater clarity will display a text reminding users it is not the full protein and provide a link to the full protein job.
- ‘Aggrescan3D plot’ portrays two per-residue interactive plots depicting (i) the A3D score distribution in which values above 0 are predicted to contribute towards protein aggregation and lower scores minimizing it and (ii) the Alphafold reported pLDDT along the sequence, a measure highly correlated with protein order/disorder (52). Plots can be locally saved in PNG format.
- ‘Aggrescan3D score’ reports the A3D and pLDDT values for each residue. Yellow masked rows correspond to aggregation contributing amino acids. On the uppermost part the database outlines descriptors useful to compare different proteins or protein variants: the protein minimal, maximal, average and total scores. The table can be downloaded in CSV format.
- ‘Structure’ shows visualization of the protein structure coloured by A3D and pLDDTs scores. A3D Model Organism allows tagging residues and taking snapshots which will be saved under the ‘Gallery’ tab. Protein structures in PDB format can be downloaded with A3D score in B-factor position.
- ‘Transmembrane regions’ which is displayed only for proteins annotated as membrane in UniProt KB. TOPCONS consensus as well as five different topological predictions (see Materials and methods) are provided in the upper plot. Furthermore, a second representation shows the consensus prediction reliability along the sequence. Topological prediction results can be saved as a TXT file and both plots as PNG figures.
- ‘Custom jobs’ lists user-derived analyses for each entry along with pre-computed ‘c50’ and ‘c70’ for which low-confidence regions were eliminated from the predictions using a >50 and >70 pLDDT score cutoff respectively. Of note, jobs run from ‘Custom jobs’ will be publicly listed, allowing other users to access and utilize them.
- ‘Gallery’ for which user-generated snapshots from ‘Structure’ can be saved as JPEG files.

Case study—evolutionary solutions to confront deleterious protein aggregation

It is assumed that the propensity to misfold towards β-sheet enriched, aggregated assemblies constitutes a common characteristic of proteins. Aggregation has therefore exerted evolutionary pressure on organisms’ proteins, conditioning hydrophobic residues’ spatial distribution (53–55), folding velocity (55,56), expression levels (55,57,58), length (55,58–59), quaternary structure stability (55,58,60) or the incorporation of solubilizing regions (55,61). A nice example is illustrated in the recent reconstruction of the evolutionary history of steroid hormone receptors (SR) by Thornton and collaborators (48). They demonstrated that the ancestral estrogen-receptor exposed a hydrophobic patch which compromised protein stability and facilitated aggregation. The receptor formed dimers as a way to allocate this STAP in the water-excluding interface, albeit no benefit was derived from this assembly in terms of protein activity. A gene duplication event led to the emergence of modern SR variants: the SR1 lineage which retained dimerization, and the SR2s which incorporated a C-terminal extension (CTE) that binds the exposed STAP therefore ensuring that the monomeric protein remains functional, stable and soluble (Figure 2). By integrating human, mouse, rat and zebrafish proteomes, A3D-MODB is able to recapitulate this lineage’s fate. Figure 2 shows A3D-MODB predictions for SR1-estrogen receptors and SR2-progesterone receptors. SRs are large proteins with a high degree of protein disorder, which

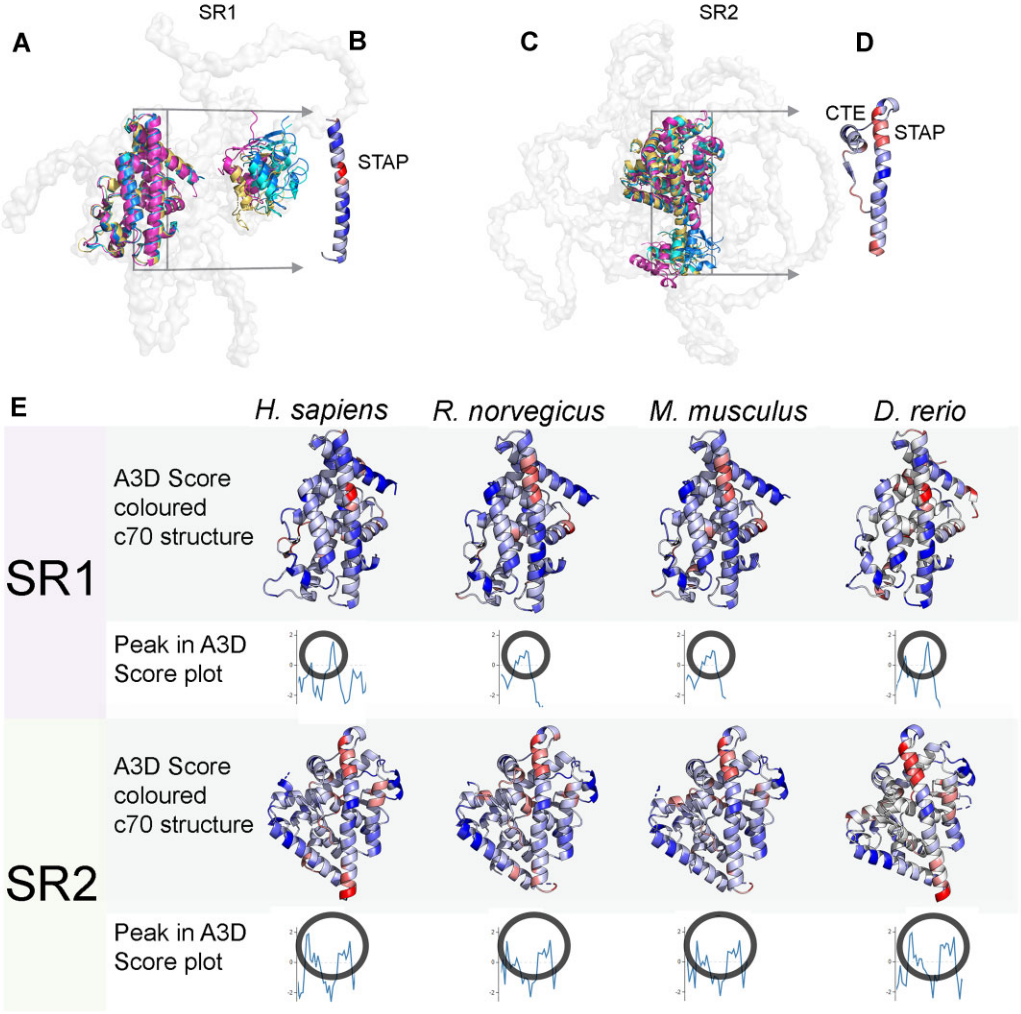


Figure 2. A3D-MODB reconstruction of the evolutionary solutions to confront exposed structural aggregation-prone regions (STAPs) in steroid hormone receptors of humans, mouse, rat and zebrafish. (**A**) Structural alignment for human, mouse, rat and zebrafish estrogen receptors (SR1) coloured yellow, cyan, blue and magenta respectively. For clarity purposes highly flexible stretches with pLDDT ≤ 70 were removed (c70) albeit the semi-transparent surface of human full protein is shown. (**B**) Dimerization helix in estrogen receptor monomers where the conserved hydrophobic region is water accessible, represented in A3D colour scheme which goes from blue (less aggregation propensity), white (neutral) and red (more aggregation prone regions). Protein dimerizes to avoid exposing this STAP. (**C**) Structural alignment of the c70 structures for human, mouse, rat and zebrafish progesterone receptors (SR2) applying colour code as in (A). The rectangle highlights the interaction between the C-terminal extension (CTE) and the exposed STAP region. (**D**) Representation of human progesterone STAP-containing helix and CTE using A3D colour scale. (**E**) A3D-MODB reports STAP conservation for human, mouse, rat and zebrafish SR1 and SR2 as seen in c70 structures and as peaks in the A3D Score distribution. Encircled peaks represent conserved STAP. For clarity purposes, only the globular domain of the SR1 containing the STAP is shown on (E). Human, rat, mouse and zebrafish UniProt Acc. for estrogen receptors SR1: P03372, Q62986, O08537, P57717 and progesterone receptors SR2: P06401, Q63449, Q00175 and C9V3N7.

is reflected in the superposition of AlphaFold models for SR1 and SR2 (Figure 2A and C). We used the ‘c70’ analysis to focus on the conserved globular domains highlighted in the Thornton study. For all SR1 and SR2 proteins we could identify the lineage-conserved STAP (Figure 2B and D), whose presence was evident in both graphical and numerical representations of individual proteins (Figure 2E).

Discussion

Protein aggregation has been an unavoidable constraint that has shaped the course of protein evolution. The use of sequence-based protein aggregation predicting software across different model organisms have revealed the impact of aggregation propensities on proteins’ folding time, size, half-life, cellular concentration, subcellular localization or translation rate (55–59,62,63). The largest computational screening for an individual species employing available, non-redundant, protein structures (∼600, roughly 14% of the proteome) was carried out in *E. coli* (58). Interestingly, this study suggested that evolutionary pressures to mitigate protein aggregation also impacted structural elements, including surfaces and interfaces.

Following the public release of the AlphaFold database, we have witnessed a sharp increase in A3D usage. This trend was particularly pronounced for proteins lacking previously determined experimental structures, primarily from model organisms (64,65). Notably, the 2020 pandemic caused by the emergence and spreading of SARS-COV-2 led to a rapid resolution of most of its structures and generated a peak of A3D analyses (66,67). To cater to this demand, we developed A3D-MODB offering pre-computed aggregation propensity analysis and tools for the study of this phenomenon on a proteome level for a diverse set of well-characterized model species. These organisms were selected from distant clades with the aim of facilitating comparative studies, such as the one exemplified here for steroid receptors.

Importantly, A3D-MODB provides users access to A3Ds' RESTful API, facilitating automatic computational data retrieval. For guidance, multiple use-case scenarios and scripts can be found in the server's tutorial or in a previous revision (51). This programmatic access not only streamlines data acquisition but also paves the way for comprehensive large-scale evolutionary studies on both general and specific factors influencing protein aggregation across organisms of varying complexity.

Despite recent effort by our research group and others to introduce environmental factors, like solution pH, into aggregation prediction models (68–70), these advancements have primarily been applied to IDPs and a limited set of experimentally validated globular proteins (70). Therefore, even if we see a clear need to include the protein microenvironment in aggregation predictions, validation and parametrization on larger protein benchmarks is still required. Additionally, there is a need to develop computationally efficient methods for modeling the thermodynamic and dynamic impact of these environmental factors on protein conformation before they can be systematically integrated into large-scale predictive frameworks.

Since its recent publication an estimate of 7000 individual users have made use of the human database. We anticipate that A3D-MODB will provide solutions for a much wider audience of researchers, given not only its extensive collection of structuromes but also its integration with databases in different biological domains. Given its adaptable architecture, the database is poised to accommodate future additions of organisms relevant to the medical, biotechnological, agricultural and industrial sectors.

In conclusion, we are confident that the A3D-MODB introduced in this study sets a new standard in protein aggregation research, which we expect to become a core resource in the field.

Data availability

A3D Model Organism Database is freely available to academic and non-profit research institutions for research purposes only; a commercial licence must be obtained for any other A3D-MODB use. The different results are publicly available through multiple data-access mechanisms. Website-generated figures and coordinate files with A3D results in the B-factor or AlphaFold database (24) provided pLDDT can be retrieved through the browser interaction. All individual protein's numerical results (including UniProt Acc.) can also be reached via a RESTful API. For large-scale analysis, the A3D-MODB provides a pre-computed comma-separated file (CSV) for each model organism containing job ids for all processed UniProt Acc. under the 'Download' tab. Latest version of A3D (40) can be installed locally following instructions in Kuriata *et al.* (41).

Supplementary data

Supplementary Data are available at NAR Online.

Funding

European Union's Horizon 2020 research and innovation programme [GA 952334 (PhasAGE)]; Spanish Ministry of Science and Innovation (MICINN) [PID2019-105017RB-I00 to S.V.]; ICREA [ICREA-Academia 2020 to S.V.]; Spanish Ministry of Science and Innovation (to J.G.-P.) with a Juan de la Cierva Incorporación [IJC2019-041039-I to J.G.-P.]; M.B. was supported by the Maria Zambrano grant funded by the European Union-NextGenerationEU; Spanish Ministry of Science and Innovation and the European Union-NextGenerationEU [Universitat Autònoma de Barcelona 02/07/2021 to V.I.]; Secretariat of Universities and Research of the Catalan Government and the European Social Fund [2023 FI_3 00018 to C.P.-G.]; National Science Centre, Sheng [2021/40/Q/NZ2/00078 to S.K.]. Funding for open access charge: European Union's Horizon 2020 research and innovation programme [GA 952334 (PhasAGE)].

Conflict of interest statement

None declared.

References

1. 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.
2. Shire,S.J., Shahrokh,Z. and Liu,J. (2004) Challenges in the development of high protein concentration formulations. *J. Pharm. Sci.*, **93**, 1390–1402.
3. Perchiacca,J.M. and Tessier,P.M. (2012) Engineering aggregation-resistant antibodies. *Annu. Rev. Chem. Biomol. Eng.*, **3**, 263–286.
4. Gil-Garcia,M. and Ventura,S. (2021) Multifunctional antibody-conjugated coiled-coil protein nanoparticles for selective cell targeting. *Acta Biomater.*, **131**, 472–482.
5. Braun,A., Kwee,L., Labow,M.A. and Alsenz,J. (1997) Protein aggregates seem to play a key role among the parameters influencing the antigenicity of interferon alpha (IFN-alpha) in normal and transgenic mice. *Pharm. Res.*, **14**, 1472–1478.
6. Auluck,P.K., Chan,H.Y., Trojanowski,J.Q., Lee,V.M. and Bonini,N.M. (2002) Chaperone suppression of alpha-synuclein toxicity in a Drosophila model for Parkinson's disease. *Science*, **295**, 865–868.
7. Arndt,V., Dick,N., Tawo,R., Dreiseidler,M., Wenzel,D., Hesse,M., Furst,D.O., Saftig,P., Saint,R., Fleischmann,B.K., *et al.* (2010) Chaperone-assisted selective autophagy is essential for muscle maintenance. *Curr. Biol.*, **20**, 143–148.
8. de Groot,N.S., Aviles,F.X., Vendrell,J. and Ventura,S. (2006) Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *FEBS J.*, **273**, 658–668.
9. Belli,M., Ramazzotti,M. and Chiti,F. (2011) Prediction of amyloid aggregation in vivo. *EMBO Rep.*, **12**, 657–663.
10. Sormanni,P., Aprile,F.A. and Vendruscolo,M. (2015) The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.*, **427**, 478–490.
11. 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.
12. Chennamsetty,N., Voynov,V., Kayser,V., Helk,B. and Trout,B.L. (2009) Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 11937–11942.

13. De Baets,G., Van Durme,J., van der Kant,R., Schymkowitz,J. and Rousseau,F. (2015) Solubis: optimize your protein. *Bioinformatics*, **31**, 2580–2582.
14. Conchillo-Sole,O., de Groot,N.S., Aviles,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 Bioinf.*, **8**, 65.
15. Sanchez de Groot,N., Pallares,I., Aviles,F.X., Vendrell,J. and Ventura,S. (2005) Prediction of “hot spots” of aggregation in disease-linked polypeptides. *BMC Struct. Biol.*, **5**, 18.
16. Jamroz,M., Kolinski,A. and Kmiecik,S. (2013) CABS-flex: server for fast simulation of protein structure fluctuations. *Nucleic Acids Res.*, **41**, W427–W431.
17. Kurcinski,M., Oleniecki,T., Ciemny,M.P., Kuriata,A., Kolinski,A. and Kmiecik,S. (2018) CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics***35**,694–695.
18. 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.
19. 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.
20. Gil-Garcia,M., Bano-Polo,M., Varejao,N., Jamroz,M., Kuriata,A., Diaz Caballero,M., Lascorz,J., Morel,B., Navarro,S., Reverter,D., *et al.* (2018) Combining structural aggregation propensity and stability predictions to re-design protein solubility. *Mol. Pharmaceutics.* ,**15** , 3846–3859.
21. Minich,A., Sarkanova,J., Levarski,Z. and Stuchlik,S. (2022) Enhancement of solubility of recombinant alcohol dehydrogenase from rhodococcus ruber using predictive tool. *World J. Microbiol. Biotechnol.*, **38**, 214.
22. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Zidek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
23. Pereira,J., Simpkin,A.J., Hartmann,M.D., Rigden,D.J., Keegan,R.M. and Lupas,A.N. (2021) High-accuracy protein structure prediction in CASP14. *Proteins*, **89**, 1687–1699.
24. 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.
25. Badaczewska-Dawid,A.E., Garcia-Pardo,J., Kuriata,A., Pujols,J., Ventura,S. and Kmiecik,S. (2022) A3D database: structure-based predictions of protein aggregation for the human proteome. *Bioinformatics*, **38**, 3121–3123.
26. Garcia-Pardo,J., Badaczewska-Dawid,A.E., Pintado-Grima,C., Iglesias,V., Kuriata,A., Kmiecik,S. and Ventura,S. (2023) A3DyDB: exploring structural aggregation propensities in the yeast proteome. *Microb. Cell Fact.*, **22**, 186.
27. Lucivero,F. (2020) Big data, Big waste? A reflection on the environmental sustainability of Big data initiatives. *Sci. Eng. Ethics*, **26**, 1009–1030.
28. Tsirigos,K.D., Peters,C., Shu,N., Kall,L. and Elofsson,A. (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.*, **43**, W401–W407.
29. UniProt, C. (2023) UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
30. Thul,P.J., Akesson,L., Wiking,M., Mahdessian,D., Geladaki,A., Ait Blal,H., Alm,T., Asplund,A., Bjork,L., Breckels,L.M., *et al.* (2017) A subcellular map of the human proteome. *Science*, **356**, eaal3321.
31. Misra,R.V., Horler,R.S., Reindl,W., Goryanin,I.I. and Thomas,G.H. (2005) EchoBASE: an integrated post-genomic database for Escherichia coli. *Nucleic Acids Res.*, **33**, D329–D333.
32. Blake,J.A., Baldarelli,R., Kadin,J.A., Richardson,J.E., Smith,C.L., Bult,C.J. and Mouse Genome Database,G. (2021) Mouse Genome Database (MGD): knowledgebase for mouse-human comparative biology. *Nucleic Acids Res.*, **49**, D981–D987.
33. Davis,P., Zarowiecki,M., Arnaboldi,V., Becerra,A., Cain,S., Chan,J., Chen,W.J., Cho,J., da Veiga Beltrame,E., Diamantakis,S., *et al.* (2022) WormBase in 2022-data, processes, and tools for analyzing Caenorhabditis elegans. *Genetics*, **220**, iyac003 .
34. Gramates,L.S., Agapite,J., Attrill,H., Calvi,B.R., Crosby,M.A., Dos Santos,G., Goodman,J.L., Goutte-Gattat,D., Jenkins,V.K., Kaufman,T., *et al.* (2022) FlyBase: a guided tour of highlighted features. *Genetics*, **220**, iyac035.
35. Vedi,M., Smith,J.R., Thomas Hayman,G., Tutaj,M., Brodie,K.C., De Pons,J.L., Demos,W.M., Gibson,A.C., Kaldunski,M.L., Lamers,L., *et al.* (2023) 2022 updates to the Rat Genome Database: a findable, accessible, interoperable, and reusable (FAIR) resource. *Genetics*, **224**, iyad042.
36. Berardini,T.Z., Reiser,L., Li,D., Mezheritsky,Y., Muller,R., Strait,E. and Huala,E. (2015) The Arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. *Genesis*, **53**, 474–485.
37. Bradford,Y.M., Van Slyke,C.E., Ruzicka,L., Singer,A., Eagle,A., Fashena,D., Howe,D.G., Frazer,K., Martin,R., Paddock,H., *et al.* (2022) Zebrafish information network, the knowledgebase for Danio rerio research. *Genetics*, **220**, iyac016.
38. Harris,M.A., Rutherford,K.M., Hayles,J., Lock,A., Bahler,J., Oliver,S.G., Mata,J. and Wood,V. (2022) Fission stories: using PomBase to understand schizosaccharomyces pombe biology. *Genetics*, **220**, iyab222.
39. Karr,J.R., Sanghvi,J.C., Macklin,D.N., Arora,A. and Covert,M.W. (2013) WholeCellKB: model organism databases for comprehensive whole-cell models. *Nucleic Acids Res.*, **41**, D787–D792.
40. 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.
41. Kuriata,A., Iglesias,V., Kurcinski,M., Ventura,S. and Kmiecik,S. (2019) Aggrescan3D standalone package for structure-based prediction of protein aggregation properties. *Bioinformatics*, **35**, 3834–3835.
42. Schrodinger, LLC. (2015) The PyMOL Molecular Graphics System, Version 1.8.
43. Viklund,H. and Elofsson,A. (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, **24**, 1662–1668.
44. Reynolds,S.M., Kall,L., Riffle,M.E., Bilmes,J.A. and Noble,W.S. (2008) Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput. Biol.*, **4**, e1000213.
45. Kall,L., Krogh,A. and Sonnhammer,E.L. (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, **21**(Suppl. 1), i251–i257.
46. Bernsel,A., Viklund,H., Falk,J., Lindahl,E., von Heijne,G. and Elofsson,A. (2008) Prediction of membrane-protein topology from first principles. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 7177–7181.
47. Viklund,H., Bernsel,A., Skwark,M. and Elofsson,A. (2008) SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics*, **24**, 2928–2929.
48. Hochberg,G.K.A., Liu,Y., Marklund,E.G., Metzger,B.P.H., Laganowsky,A. and Thornton,J.W. (2020) A hydrophobic ratchet entrenches molecular complexes. *Nature*, **588**, 503–508.
49. Tanenbaum,D.M., Wang,Y., Williams,S.P. and Sigler,P.B. (1998) Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 5998–6003.

50. Pujols,J., Pena-Diaz,S. and Ventura,S. (2018) AGGRESCAN3D: toward the prediction of the aggregation propensities of protein structures. *Methods Mol. Biol.*, **1762**, 427–443.
51. Pujols,J., Iglesias,V., Santos,J., Kuriata,A., Kmiecik,S. and Ventura,S. (2022) A3D 2.0 Update for the prediction and optimization of protein solubility. *Methods Mol. Biol.*, **2406**, 65–84.
52. Piovesan,D., Monzon,A.M. and Tosatto,S.C.E. (2022) Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci.*, **31**, e4466.
53. Monsellier,E., Ramazzotti,M., de Laureto,P.P., Tartaglia,G.G., Taddei,N., Fontana,A., Vendruscolo,M. and Chiti,F. (2007) The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution. *Biophys. J.*, **93**, 4382–4391.
54. Rousseau,F., Serrano,L. and Schymkowitz,J.W. (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity. *J. Mol. Biol.*, **355**, 1037–1047.
55. Santos,J., Iglesias,V. and Ventura,S. (2020) Computational prediction and redesign of aberrant protein oligomerization. *Prog. Mol. Biol. Transl. Sci.*, **169**, 43–83.
56. Monsellier,E. and Chiti,F. (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.*, **8**, 737–742.
57. Ciryam,P., Tartaglia,G.G., Morimoto,R.I., Dobson,C.M. and Vendruscolo,M. (2013) Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins. *Cell Rep.*, **5**, 781–790.
58. Carija,A., Pinheiro,F., Iglesias,V. and Ventura,S. (2019) Computational assessment of bacterial protein structures indicates a selection against aggregation. *Cells*, **8**, 856.
59. Monsellier,E., Ramazzotti,M., Taddei,N. and Chiti,F. (2008) Aggregation propensity of the human proteome. *PLoS Comput. Biol.*, **4**, e1000199.
60. Yee,A.W., Aldeghi,M., Blakeley,M.P., Ostermann,A., Mas,P.J., Moulin,M., de Sanctis,D., Bowler,M.W., Mueller-Dieckmann,C., Mitchell,E.P., *et al.* (2019) A molecular mechanism for transthyretin amyloidogenesis. *Nat. Commun.*, **10**, 925.
61. Grana-Montes,R., Marinelli,P., Reverter,D. and Ventura,S. (2014) N-terminal protein tails act as aggregation protective entropic bristles: the SUMO case. *Biomacromolecules*, **15**, 1194–1203.
62. Chen,Y. and Dokholyan,N.V. (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol. Biol. Evol.*, **25**, 1530–1533.
63. de Groot,N.S. and Ventura,S. (2010) Protein aggregation profile of the bacterial cytosol. *PLoS One*, **5**, e9383.
64. Mizuno,H., Hoshino,J., So,M., Kogure,Y., Fujii,T., Ubara,Y., Takaichi,K., Nakaniwa,T., Tanaka,H., Kurisu,G., *et al.* (2021) Dialysis-related amyloidosis associated with a novel beta(2)-microglobulin variant. *Amyloid*, **28**, 42–49.
65. Ruiz-Solani,N., Salguero-Linares,J., Armengot,L., Santos,J., Pallares,I., van Midden,K.P., Phukkan,U.J., Koyuncu,S., Borras-Bisa,J., Li,L., *et al.* (2023) Arabidopsis metacaspase MC1 localizes in stress granules, clears protein aggregates and delays senescence. *Plant Cell.*, **35**, 3325–3344.
66. Petrlova,J., Samsudin,F., Bond,P.J. and Schmidtchen,A. (2022) SARS-CoV-2 spike protein aggregation is triggered by bacterial lipopolysaccharide. *FEBS Lett.*, **596**, 2566–2575.
67. Abduljaleel,Z., Melebari,S., Athar,M., Dehlawi,S., Udhaya Kumar,S., Aziz,S.A., Dannoun,A.I., Malik,S.M., Thasleem,J. and George Priya Doss,C. (2023) SARS-CoV-2 vaccine breakthrough infections (VBI) by Omicron variant (B.1.1.529) and consequences in structural and functional impact. *Cell. Signal.***109**, 110798.
68. 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**, E145.
69. 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.
70. Oeller,M., Kang,R., Bell,R., Ausserwoger,H., Sormanni,P. and Vendruscolo,M. (2023) Sequence-based prediction of pH-dependent protein solubility using CamSol. *Brief. Bioinf.*, **24**, bbad004.

Received: August 14, 2023. **Revised:** September 27, 2023. **Editorial Decision:** October 11, 2023. **Accepted:** October 11, 2023