Abstract

Summary: Protein aggregation is associated with many human disorders and constitutes a major bottleneck for producing therapeutic proteins. Our knowledge of the human protein structures repertoire has dramatically increased with the recent development of the AlphaFold (AF) deep-learning method. This structural information can be used to understand better protein aggregation properties and the rational design of protein solubility. This article uses the Aggrescan3D (A3D) tool to compute the structure-based aggregation predictions for the human proteome and make the predictions available in a database form. In the A3D database, we analyze the AF-predicted human protein structures (for over 20.5 thousand unique Uniprot IDs) in terms of their aggregation properties using the A3D tool. Each entry of the A3D database provides a detailed analysis of the structure-based aggregation propensity computed with A3D. The A3D database implements simple but useful graphical tools for visualizing and interpreting protein structure datasets. It also enables testing the influence of user-selected mutations on protein solubility and stability, all integrated into a user-friendly interface.

Availability and implementation: A3D database is freely available at: http://biocomp.chem.uw.edu.pl/A3D2/hproteome. The data underlying this article are available in the article and in its online supplementary material.

Contact: sekmi@chem.uw.edu.pl and salvador.ventura@uab.cat

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

In July 2021, a database of highly accurate structure predictions for the human proteome was published (Tunyasuvunakool et al., 2021). The predictions computed using the newly developed neural network model AlphaFold (AF), were shown to be competitive with experimental structures (Jumper et al., 2021).

Here, we have constructed the AGGRESCAN3D (A3D) Database by computing the aggregation propensity of the human protein models from the AF database. The A3D is a structure-based predictor of surface-exposed aggregation-prone regions. The A3D algorithm exploits the information of 3D atomic models to compute the structurally corrected aggregation values (A3D score) for each amino acid (Kuriata et al., 2019a, b; Pujols et al., 2018; Zambrano et al., 2015). A3D can predict the effect of mutations on protein stability and aggregation propensity, as well as suggest solubility-enhancing mutations. This algorithm has been employed to study the constraints imposed by aggregation on protein evolution (Carija et al., 2019), to diagnose the functional impact of genetic mutations (Seaby and Ennis, 2020), to predict the aggregation of the SARS-CoV-2 proteome (Flores-León et al., 2021), to assist the design of novel nanomaterials (Gil-García and Ventura, 2021) or to engineer the solubility of therapeutic proteins (de Aguiar et al., 2021; Gil-García et al., 2018) among many other applications.

2 A3D database features

The A3D database integrates A3D analysis for 23391 predicted structures of the human proteome from the AF database. The content of the A3D database can be queried by UniProt ID, Gene or protein name (see Movie S1 in Supplementary Information for
which residues with pLDDT < A3D analysis using three different AF models for each protein were removed (see Supplementary Information). Access to these version of this figure appears in the online version of this article). 

Fig. 1. Examples of protein model visualizations from the A3D database. For each database entry, under the Structure tab, two protein copies are presented colored according to (i) the A3D score and (ii) the AlphaFold (AF) model confidence score. The A3D score is visualized in shades from dark blue (highly soluble residues, score < −2.5), through white (no predicted influence on aggregation properties), to dark red (aggregation-prone residues, score > 2.5). The AF per-residue confidence score (pLDDT) is presented in dark blue (very high confidence, pLDDT > 90), light blue (confident, 90 > pLDDT > 70), yellow (low confidence, 70 > pLDDT > 50) and orange (very low confidence, pLDDT < 50). Note that pLDDT < 50 is a reasonably strong predictor of disorder (Tunyasuvunakool et al., 2021), which suggests that a particular region may be unstructured as a linker between domains (see b) or as an inherently disordered domain (see c). (a) An example of a globular protein predicted with high confidence is shown (A color version of this figure appears in the online version of this article).

globular
AF-S4R460-F1
multidomain
AF-Q96MN5-F1
disordered
AF-Q8N9P0-F1

In summary, the A3D database can be helpful in the study and redesign of human proteins’ solubility (also in combination with other human proteome predictions; Prabakaran et al., 2021). It also allows investigating correlations between structural aggregation propensity and protein function, stability, architecture, location, abundance, lifetime or essentiality at the proteome level. In Supplementary Information, we illustrate and discuss the utility of the database with selected case reports.

Funding
A.E.B-D. received financial support from Roy J. Carver Charitable Trust and Iowa State University Foundation. This work was supported by the Spanish Ministry of Science and Innovation (MICINN) [PID2019-105017RB-I00 to S.V.], by ICREA, ICREA-Academia 2019, and by EU [PhasAge/H2020-WIDESPREAD-2020-3 to S.V.]; the Spanish Ministry of Science and Innovation with a Juan de la Cierva Incorporacion [JJC2019-041039-I to J.G.-P.]; S.K. acknowledges funding from the National Science Centre, Poland [2020/39/B/NZ2/01301].

Conflict of Interest: none declared.

References
Carija, A. et al. (2019) Computational assessment of bacterial protein structures indicates a selection against aggregation. Cells, 8, 856.
de Aguiar, R.B. et al. (2021) Generation and functional characterization of a single-chain variable fragment (scFv) of the anti-FGF2 3F12E7 monoclonal antibody. Sci. Rep., 11, 1432.
Flores-León, M. et al. (2021) In silico analysis of the aggregation propensity of the SARS-CoV-2 proteome: insight into possible cellular pathologies. Biochim. Biophys. Acta. Proteins Proteom., 1869, 140693.
Gil-Garcia, M. et al. (2018) Combining structural aggregation propensity and stability predictions to redesign protein solubility. Mol. Pharm., 15, 3846–3859.
Gil-Garcia, M., Ventura S. (2021) Multifunctional antibody-conjugated coil-led-coil protein nanoparticles for selective cell targeting. Acta Biomater., 131, 472–482.
Jumper, J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.
Kuriata, A. et al. (2019a) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res.*, 47, W300–W307.
Kuriata, A. et al. (2019b) Aggrescan3D standalone package for structure-based prediction of protein aggregation properties. *Bioinformatics*, 35, 3834–3835.
Prabakaran, R. et al. (2021) ANuPP: a versatile tool to predict aggregation nucleating regions in peptides and proteins. *J. Mol. Biol.*, 433, 166707.
Pujols, J. et al. (2018) AGGRESCAN3D: toward the prediction of the aggregation propensities of protein structures. *Methods Mol. Biol.*, 1762, 427–443.
Seaby E.G., Ennis S. (2020) Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Brief. Funct. Genomics*, 19, 243–258.
Tunyasuvunakool, K. et al. (2021) Highly accurate protein structure prediction for the human proteome. *Nature*, 596, 590–596.
Zambrano, R. et al. (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.*, 43, W306–13.