ADMET & DMPK 8(4) (2020) 391-399; doi: http://dx.doi.org/10.5599/admet.831

ADMET

Open Access : ISSN : 1848-7718

http://www.pub.iapchem.org/ojs/index.php/admet/index

Feature article

Solubility of proteins

Mauno Vihinen*

Protein Structure and Bioinformatics, Department of Experimental Medical Science, BMC B13, SE-22184 Lund, Sweden

*Corresponding Author: E-mail: mauno.vihinen@med.lu.se; Tel.: +46-72-5260022

Received: April 24, 2020; Revised: June 18, 2020; Published: June 28, 2020

Abstract

Solubility is a fundamental protein property that has important connotations for therapeutics and use in diagnosis. Solubility of many proteins is low and affect heterologous overexpression of proteins, formulation of products and their stability. Two processes are related to soluble and solid phase relations. Solubility refers to the process where proteins have correctly folded structure, whereas aggregation is related to the formation of fibrils, oligomers or amorphous particles. Both processes are related to some diseases. Amyloid fibril formation is one of the characteristic features in several neurodegenerative diseases, but it is related to many other diseases, including cancers. Severe complex V deficiency and cataract are examples of diseases due to reduced protein solubility. Methods and approaches are described for prediction of protein solubility and aggregation, as well as predictions of consequences of amino acid substitutions. Finally, protein engineering solutions are discussed. Protein solubility can be increased, although such alterations are relatively rare and can lead to trade-off with some other properties. The aggregation prediction methods mainly aim to detect aggregation-prone sequence patches and then making them more soluble. The solubility predictors utilize a wide spectrum of features.

Keywords

aggregation; protein engineering; biologic; solubility prediction

Introduction

Solubility is an important property for all drugs, including biologics. Many proteins and polypeptides are poorly soluble and those trespassing through cellular membranes, membrane proteins, are not in traditional sense soluble at all. Proteome-wide analysis of solubility in *Caenorhabditis elegans* indicated that about 75 % of proteins appear in cells in abundances close to their solubility limits [1]. We can distinguish two phenomena in relation to protein behaviour in solution. Solubility and aggregation are related but have different meanings. Solubility is defined here and in many other publications as the concentration in which intact protein is in equilibrium with solid phase [2-4]. In the case of aggregation, protein molecules bind together, often due to irreversibly altered conformation, and form insoluble high molecular weight forms (see Figure 1) [5].

Unlike aggregated forms, precipitated solubilizable protein in solid phase can be made soluble by dilution. Aggregated protein is in solid phase and typically undergoes irreversible structural changes, thus aggregated protein cannot be returned back to soluble form and original structure. Proteins in these amyloid fibrils have extensive β-strand secondary structures. When intrinsically disordered proteins

aggregate they form amorphous deposits, whereas native-like structures aggregate to native-like deposits (Figure 1) [5]. Both solubility and aggregation have biological and biotechnological consequences. Reduced solubility as well as increased aggregation are related to several diseases. Plaques formed by aggregated proteins are common in neurodegenerative diseases. Altered conformations of prion proteins can “infect” other proteins and cause their aggregation. β-sheet formation is often related to aggregation and prion formation. These secondary structural elements can stack intra- and intermolecularly to form large insoluble aggregates. Even protein crystallization is related to solubility. Ordered protein crystals are needed for X-ray crystallography to reveal protein structures. These crystals are grown slowly and contain typically large amounts of solvent. The goal of the crystallization is to keep the proteins in their native conformation, however certain packing effects affecting local conformation are common.

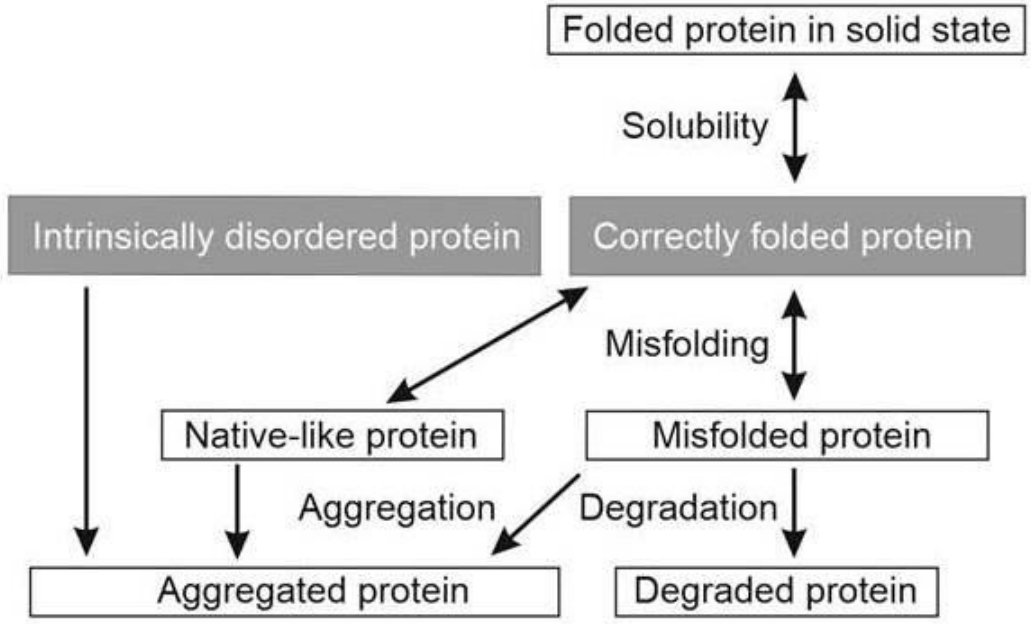

Figure 1. Relationships of protein solubility, aggregation and degradation. Insoluble protein is correctly folded, whereas the structure of aggregated protein is altered. Most irreversibly misfolded proteins are degraded as part of normal protein turnover. Intrinsically disordered proteins and native-like structures have their specific aggregation mechanisms.

Many factors affect solubility and aggregation, including intrinsic properties of the protein, solvent and additives as well as physical conditions. List of relevant protein factors is long and could be started with amino acid sequence and composition, three-dimensional structure and exposure of residues, intramolecular interactions within the protein (salt bridges, hydrogen bonds, electrostatic and hydrophobic interactions, multimeric status etc.), and protonation status. Solvent properties (polarity, bond and interaction forming ability, density etc.) and constituents (such as excipients, salts or organic solvents) and their concentrations have significant contributions to solubility. Further, factors like pH and temperature affect both the protein and solvent.

Proteins are large and relatively fragile molecules, thus not necessarily ideal as drugs. Lots of research has been devoted to find smaller structures having (some) functions of the biological proteins. Miniproteins or protein scaffolds have been tested as biologics [6]. Some functions can be retained even in short peptides as highlighted by 2018 Nobel prize in chemistry for Frances H. Arnold, George P. Smith and Sir Gregory P. Winter, the last two ones "for the phage display of peptides and antibodies" by developing small binding molecules. In this respect, cyclic peptides are of interest as they have more conformational restraints than linear molecules.

Quite successful solubility prediction methods have been developed for small molecules that are most common as drugs (see chapters in this volume). These methods, however, do not work for proteins. There are several reasons. Proteins are much larger, they have lots of groups and protein folding affects the solvent accessibility of the groups. Further, the protein structures are flexible and have a large number of slightly different conformations. Therefore, different approaches are needed for protein solubility prediction. Here, methods for protein solubility and aggregation prediction are introduced. Further, methods to investigate the effects on solubility due to amino acid substitutions (variations) are discussed. This topic is important for the design of changes for protein engineering e.g. to increase protein solubility, production etc.

Protein solubility is discussed here from the perspective of predictions for solubility, aggregation and for variation effects. The principles of the methods are discussed along with performances of the tools. It is apparent that the performances of the methods are not very high, which is due to the complexity of the phenomenon and thereby difficulty in finding good features that would reliably distinguish between the solubility states. One limiting factor has been the small number of experimentally verified cases. However, the field is making progress and there are some rather good methods available.

Solubility prediction

During the last decade, several computational methods have been developed to predict protein solubility, especially in the context of heterologous protein overexpression. These methods utilize different approaches, often in the field of machine learning. Solubility is a complex phenomenon and good predictive features are difficult to find, however, there are some trends such as residue-residue interactions [7] and structural flexibility [8].

Based on examples with known solubility, computer tools have been trained (see e.g. [9]). Protein related parameters such as hydropathy scales and amino acid compositions have been used as features. The best methods claim accuracy of over 80 % for two-state predictions of soluble/insoluble. Methods in this category include ccSOL omics [10], DeepSol [11], PaRSnIP [12], Protein-Sol [13], SODA [14], SOLart [15], SOLpro [16], SWI [8] and others.

Table 1. Protein solubility predictors

Method	URL
Trained with data from Structural Biology Knowledgebase	
ccSOL omics	http://s.tartaglialab.com/update_submission/45568/57e42bea38
DeepSol	https://zenodo.org/record/1162886#.Xoxw_EGxVEY
PaRSnIP	https://github.com/RedaRawi/PaRSnIP
SWI	https://tisigner.com/sodope
Trained with data from eSOL	
ProteinSol	https://protein-sol.manchester.ac.uk/
SOLart	http://babylone.ulb.ac.be/SOLART/
SOLpro	http://scratch.proteomics.ics.uci.edu/
Trained with data from PON-Sol	
SODA	http://old.protein.bio.unipd.it/soda/

Most methods are trained on one of two major datasets. Protein Structure Initiative was a large project to determine protein structures *en masse*. Their Structural Biology Knowledgebase [17] contains information for crystallization trials and for (heterologous) protein production and has been used for training many of the tools including ccSOL omics, PaRSnIP, different versions of DeepSol and SWI. eSOL

solubility database (http://tanpaku.org/tp-esol/index.php?lang=en) for almost all *Esherichia coli* proteins [18] has been used for some other methods, such as Protein-Sol, SOLart along with *Saccharomyces cerevisiae* protein solubility details [19]. The remaining methods are trained with PON-Sol data [4] or not trained at all.

Although many proteins are poorly soluble their solubility is biologically sufficient as many proteins have very low abundance in cells [1]. Data extracted by DeepSol developers from Structural Biology Knowledgebase indicated that 45.2 % out of 129.643 tested inherent and heterologously expressed proteins in *E. coli* were soluble [11].

DeepSol is a deep learning method, ccSOL omics is a support vector machine solution, PaRSnIP utilizes gradient boosting, SOLart random forest, SWI and Protein-Sol are based on weighted scores. Details for the algorithm used in SODA have not been released.

As with any prediction task, the choice of the tool(s) is important. The best comparisons are independent benchmark studies [20, 21], however such studies are not available for any of the prediction tasks discussed in here. Instead, there are some comparisons of methods along with predictor description, including comparisons of four [14], seven [12] and eight [11] tools. The first of these studies indicated SODA to be the best, accuracy 59.2, the second PaRSnIP (accuracy 74.11 and Matthews correlation coefficient (MCC) 0.48) and now defunct PROSO II (accuracy 64.35, MCC 0.31) [22], and in the third DeepSol, PaRSnIP and PROSO II were the best ones with accuracies and MCC values of 0.77/0.55, 0.74/0.48 and 0.64/0.34, respectively.

Analysis of predictions for 57 UDP-dependent glycosyltransferases with 11 predictors [23] is interesting, however, may be biased due to containing proteins just from a single family. The best performing methods in their analysis were SoluProt, DeepSol versions 3, 1 and 2, SolPro and PaRSnIP. Unfortunately, the results are provided only in a format of a figure, thus no numbers are available.

Aggregation prediction

Aggregation is largely mediated by short sequence stretches of consecutive residues, these regions are typically 15 residues or longer [24]. Almost all human proteins can form self-complementary and thus aggregation-prone segments [25]. However, many proteins do not aggregate and during evolution have adapted to prevent amyloid formation [26]. Chaperones that assist proteins to fold are in central role [27]. Aggregated proteins form either amyloid fibrils, amorphous or native-like deposits. CPAD [28], AmyLoad [29] and AmyPro [30] are databases dedicated for information about protein aggregation. CPAD includes data for amyloid peptides, aggregation prone peptides and aggregation rates while AmyLoad contains amylodogenic sequence information. Proteins containing experimentally verified amyloidogenic regions are collected to AmyPro.

Aggregation predictors can be grouped to two major categories: sequence-based and three-dimensional structure-based. The structure-based methods utilize calculations of free energy difference between solution and aggregation phases, β-structure formation propensity, residue exposure and so on. Since protein 3D structures are not always available, sequence-based methods are needed, as well. Machine learning methods use features such as amino acid composition and proportions of certain amino acid types for training. Aggregation predictors include for example AGGRESCAN [31], AGGRESCAN3D [32], AMYLPRED [33], ArchCandy [34], FoldAmyloid [35], MetAmyl [36], PASTA 2.0 [37], TANGO [38], and Waltz [39]. AGGRESCAN3D and ArchCandy are based on three dimensional structures, some of the others use some structural features, as well.

Table 2. Protein aggregation predictors

AGGRESCAN	http://bioinf.uab.es/aggrescan/
AGGRESCAN3D	https://bitbucket.org/lcbio/aggrescan3d/src/master/
AMYLPRED2	http://aias.biol.uoa.gr/AMYLPRED2/
ArchCandy	https://bioinfo.crbm.cnrs.fr/index.php?route=tools&tool=7&lang=uk
FoldAmyloid	http://bioinfo.protres.ru/fold-amyloid/
MetAmyl	http://metamyl.genouest.org/e107_plugins/metamyl_aggregation/db_prediction_meta.php
PASTA 2.0	http://protein.bio.unipd.it/pasta2/help.html
RFAmyloid	http://server.malab.cn/RFAmyloid/
Tango	http://tango.crg.es/
Waltz	https://waltz.switchlab.org/

AmylPred and MetAmyl are metapredictors, i.e. use predictions from other tools. RFamyloid is machine learning-based method, Waltz has a position specific scoring matrix, the others are based on physicochemical propensities and other features including information about secondary structural elements, amino acid composition, structural features, physicochemical propensities of amino acids, packing density, hydrogen bonds etc. Neural network tool for amyloid aggregation rate prediction is a related application [40]. Some of the methods can be used both for solubility and aggregation prediction, such as SOLart [15]. SODA, a solubility predictor, utilizes aggregation and disorder propensities [14].

Systematic independent method performance assessments are missing. Some recent studies contain comparisons of several methods. PASTA 2.0 was compared to eight methods, showing the best performance (MCC 0.24) along with AMYLPRED 2 (0.22) and MetAmyl (0.19) [37]. However, the performances are not very high, specificity being clearly better than sensitivity for all the tools, the best Matthews correlation coefficient being only 0.24. Developers of ArchCandy saw really big differences in performances for six predictors [41], their own tool (error rate for amyloid prediction 1.4 %) along with PASTA (4.2 %), TANGO (5.0 %) and Waltz (11.4 %) being the best. This test contained only soluble segments of 15 residues or longer. They were collected from three-dimensional protein structures determined with NMR, thus the proteins are soluble, however contained highly flexible regions without ordered structures, typically in the termini. Comprehensive performance assessment would require both positive and negative cases [20, 21], thereby we do not know if the best methods are overpredicting soluble regions and underpredicting aggregation-prone segments.

Prediction of solubility and aggregation affecting variants

The tools described above are for predictions on entire proteins or polypeptides. Much less effort has been put on predicting the effect of variations on solubility or aggregation. We are looking at tools predicting consequences of amino acid substitutions as there is not enough data for other types of variations. Single amino acid alterations can have profound effects of solubility and lead to diseases, including severe complex V deficiency [42] and cataract [43]. Protein solubility and aggregation are mechanisms in diseases, including cancers [44, 45].

Predictors for solubility affecting variants include CamSol [46], OptSolMut [47], PON-Sol [4], SODA [14] and SolubiS [48]. CamSol uses residue-specific solubility profile. The method is not available as a tool, only the algorithm has been described. OptSolMut has been trained with a small dataset that contains also aggregation cases. Weights for scoring function were optimized with linear programming for 137 cases of single and multiple variants. PON-Sol is a random forest-based machine learning method. It was trained and tested on 406 single amino acid substitutions for which solubility effect have been experimentally

determined. It predicts variants into three classes: solubility decreasing and increasing variants and those not affecting solubility. This is a more realistic scenario than binary prediction. In the three-state prediction PON-Sol had correct prediction ratio of 0.597 on cross validation and 0.488 for independent test set (note that random prediction has a score of 0.33). Thus, there is still place for substantial improvements. The method development has been hampered by small number of known solubility-affecting variants. PON-Sol is no more available; however, a new extended and improved version will be released soon. SODA has been recommended to predict variants decreasing solubility [14]. It was developed with PON-Sol data.

These methods can be used for numerous purposes including identification of disease related amino acid substitutions, predictions of solubility of heterologous recombinant protein expression and enhanced crystallizability. Of the aggregation prediction tools, PASTA 2.0 can predict also effects of amino acid substitutions. SoluBis has a somewhat different application, optimization of multiple variants to increase protein solubility [48]. It detects aggregation prone segments and then suggests variants to modify them. So, actually it is an aggregation prevention predictor. It combines predictions from interaction analysis tool FoldX [49], aggregation predictor TANGO [38] and structural analysis with YASARA [50].

Predictions for protein engineering

Protein properties have complex relations. Several approaches have been tried to improve solubility or prevent aggregation. Electrostatic interactions have been a starting point for one approach [51], stability and aggregation for another [52], and surface patches to avoid aggregation for a third one [53]. SolubiS tries to reduce aggregation propensity [48].

Structural changes designed by four tools have been reviewed in relation to structure-based predictions [54]. The discussed tools included Aggrescan3D, CamSol, Spatial Aggregation Propensity (SAP) and SolubiS. SAP was a proposal to apply molecular dynamics simulations, not an implemented tool [55].

Systematic performance assessments have not been made to these methods. Massively parallel reporter assay (MPRA) of two proteins, TEM-1 β-lactamase, a common antibiotic resistance protein in Gram positive bacteria, and *E. coli* levoglucosan kinase, indicated trade-offs between fitness and solubility [56]. Solubility in this paper was defined as properly folded protein. They used two analysis methods, which revealed whether the protein was folded. In yeast surface display screen the investigated protein had to be folded otherwise the fusion protein was degraded. In twin-arginine translocation-selective export the protein was exported to bacterial periplasm only if correctly folded. They generated >93 % of all possible single amino acid variants in the two proteins. Solubility increasing variants were rare, only 4 to 5 % had this effect. Many solubility increasing variants affected also some other property. Comparisons to fitness increasing variants revealed that these two features co-occurred very rarely and there were trade-offs between them.

Conclusions

Computational methods available for the prediction of protein solubility and aggregation were discussed along with tools for engineering solubility or aggregation by introducing amino acid substitutions. Many features and different algorithms have been applied to the available solutions. Although systematic performance assessments have not been performed, it is evident that the methods have widely varying performances. Solubility and avoidance of aggregation are crucial properties for any protein to be used for diagnosis or therapy.

***Acknowledgements**: Financial support from Vetenskapsrådet, Swedish Cancer Society and Alfred Österlunds Stiftelse is gratefully acknowledged.*

Conflict of interest*: None.*

References

[1] G. Vecchi, P. Sormanni, B. Mannini, A. Vandelli, G.G. Tartaglia, C.M. Dobson, F.U. Hartl, M. Vendruscolo. Proteome-wide observation of the phenomenon of life on the edge of solubility. *Proc. Natl. Acad. Sci. U. S. A.* **117** (2020) 1015-1020.

[2] T. Arakawa, S.N. Timasheff. Theory of protein solubility. *Methods Enzymol.* **114** (1985) 49-77.

[3] P. Garidel. Protein solubility from biochemical, physicochmical and colloidal perspective. *Am. Pharm. Rev.* **December 13** (2013).

[4] Y. Yang, A. Niroula, B. Shen, M. Vihinen. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics* **32** (2016) 2032-2034.

[5] F. Chiti, C.M. Dobson. Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annu. Rev. Biochem.* **86** (2017) 27-68.

[6] Z.R. Crook, N.W. Nairn, J.M. Olson. Miniproteins as a powerful modality in drug development. *Trends Biochem. Sci.* **45** (2020) 332-346.

[7] Q. Hou, R. Bourgeas, F. Pucci, M. Rooman. Computational analysis of the amino acid interactions that promote or decrease protein solubility. *Sci. Rep.* **8** (2018) 14661.

[8] B.K. Bhandari, P.P. Gardner, C.S. Lim. Solubility-Weighted Index: fast and accurate prediction of protein solubility. (2020). https://doi.org/10.1093/bioinformatics/btaa578.

[9] N. Habibi, S.Z. Mohd Hashim, A. Norouzi, M.R. Samian. A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. *BMC Bioinformatics* **15** (2014) 134.

[10] F. Agostini, D. Cirillo, C.M. Livi, R. Delli Ponti, G.G. Tartaglia. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. *Bioinformatics* **30** (2014) 2975-2977.

[11] S. Khurana, R. Rawi, K. Kunji, G.Y. Chuang, H. Bensmail, R. Mall. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **34** (2018) 2605-2613.

[12] R. Rawi, R. Mall, K. Kunji, C.H. Shen, P.D. Kwong, G.Y. Chuang. PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics* **34** (2018) 1092-1098.

[13] M. Hebditch, M.A. Carballo-Amador, S. Charonis, R. Curtis, J. Warwicker. Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* **33** (2017) 3098-3100.

[14] L. Paladin, D. Piovesan, S.C.E. Tosatto. SODA: prediction of protein solubility from disorder and aggregation propensity. *Nucleic Acids Res.* **45** (2017) W236-W240.

[15] Q. Hou, J.M. Kwasigroch, M. Rooman, F. Pucci. SOLart: a structure-based method to predict protein solubility and aggregation. *Bioinformatics* **36** (2020) 1445-1452.

[16] C.N. Magnan, A. Randall, P. Baldi. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* **25** (2009) 2200-2207.

[17] M.J. Gabanyi, P.D. Adams, K. Arnold, L. Bordoli, L.G. Carter, J. Flippen-Andersen, L. Gifford, J. Haas, A. Kouranov, W.A. McLaughlin, D.I. Micallef, W. Minor, R. Shah, T. Schwede, Y.P. Tao, J.D. Westbrook, M. Zimmerman, H.M. Berman. The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J. Struct. Funct. Genomics* **12** (2011) 45-54.

[18] T. Niwa, B.W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, H. Taguchi. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc. Natl. Acad. Sci. U. S. A.* **106** (2009) 4201-4206.

[19] E. Uemura, T. Niwa, S. Minami, K. Takemoto, S. Fukuchi, K. Machida, H. Imataka, T. Ueda, M. Ota, H. Taguchi. Large-scale aggregation analysis of eukaryotic proteins reveals an involvement of intrinsically disordered regions in protein folding. *Sci. Rep.* **8** (2018) 678.

[20] M. Vihinen. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13 Suppl 4** (2012) S2.

[21] M. Vihinen. Guidelines for reporting and using prediction tools for genetic variation analysis. *Hum. Mutat.* **34** (2013) 275-282.

[22] P. Smialowski, G. Doose, P. Torkler, S. Kaufmann, D. Frishman. PROSO II--a new method for protein solubility prediction. *Febs j.* **279** (2012) 2192-2200.

[23] F.A. Ghomi, T. Kittilä, D.H. Welner. A benchmark of protein solubility prediction methods on UDP-dependent glycosyltransferases. (2020). doi: https://doi.org/10.1101/2020.02.28.962894**.**

[24] A.B. Ahmed, A.V. Kajava. Breaking the amyloidogenicity code: methods to predict amyloids from amino acid sequence. *FEBS Lett.* **587** (2013) 1089-1095.

[25] L. Goldschmidt, P.K. Teng, R. Riek, D. Eisenberg. Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **107** (2010) 3487-3492.

[26] E. Monsellier, F. Chiti. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.* **8** (2007) 737-742.

[27] N.B. Nillegoda, A.S. Wentink, B. Bukau. Protein disaggregation in multicellular organisms. *Trends Biochem. Sci.* **43** (2018) 285-300.

[28] P. Rawat, R. Prabakaran, R. Sakthivel, A. Mary Thangakani, S. Kumar, M.M. Gromiha. CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides. *Amyloid* (2020) 1-6.

[29] P.P. Wozniak, M. Kotulska. AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics* **31** (2015) 3395-3397.

[30] M. Varadi, G. De Baets, W.F. Vranken, P. Tompa, R. Pancsa. AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res.* **46** (2018) D387-d392.

[31] O. Conchillo-Sole, N.S. de Groot, F.X. Aviles, J. Vendrell, X. Daura, S. Ventura. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* **8** (2007) 65.

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

[33] K.K. Frousios, V.A. Iconomidou, C.M. Karletidi, S.J. Hamodrakas. Amyloidogenic determinants are usually not buried. *BMC Struct Biol.* (2009) Jul 9; 9:44. doi: https://doi.org/10.1186/1472-6807-9-44.

[34] A.B. Ahmed, N. Znassi, M.T. Chateau, A.V. Kajava. A structure-based approach to predict predisposition to amyloidosis. *Alzheimers Dement.* **11** (2015) 681-690.

[35] S.O. Garbuzynskiy, M.Y. Lobanov, O.V. Galzitskaya. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics.* **26**(3) (2010) 326-332. doi: https://doi.org/10.1093/bioinformatics/btp691.

[36] M. Emily, A. Talvas, C. Delamarche. MetAmyl: a METa-predictor for AMYLoid proteins. *PLoS One.* 8(11) (2013) e79722. doi: https://doi.org/10.1371/journal.pone.0079722.

[37] I. Walsh, F. Seno, S.C. Tosatto, A. Trovato. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res* **42** (2014) W301-W307.

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

[39] K.L. Morris, A. Rodger, M.R. Hicks, M. Debulpaep, J. Schymkowitz, F. Rousseau, L.C. Serpell. Exploring the sequence-structure relationship for amyloid peptides. *Biochem J.* **450**(2) (2013) 275-283. doi: https://doi.org/10.1042/BJ20121773.

[40] W. Yang, P. Tan, X. Fu, L. Hong. Prediction of amyloid aggregation rates by machine learning and feature selection. *J. Chem. Phys.* **151** (2019) 084106.

[41] D.B. Roche, E. Villain, A.V. Kajava. Usage of a dataset of NMR resolved protein structures to test aggregation versus solubility prediction algorithms. *Protein Sci.* **26** (2017) 1864-1869.

[42] A. Meulemans, S. Seneca, T. Pribyl, J. Smet, V. Alderweirldt, A. Waeytens, W. Lissens, R. Van Coster, L. De Meirleir, J.P. di Rago, D.L. Gatti, S.H. Ackerman. Defining the pathogenesis of the human Atp12p W94R mutation using a Saccharomyces cerevisiae yeast model. *J. Biol. Chem.* **285** (2010) 4099-4109.

[43] U.P. Andley, M.A. Reilly. In vivo lens deficiency of the R49C alphaA-crystallin mutant. *Exp. Eye Res.* **90** (2010) 699-702.

[44] G.A.P. de Oliveira, Y. Cordeiro, J.L. Silva, T. Vieira. Liquid-liquid phase transitions and amyloid aggregation in proteins related to cancer and neurodegenerative diseases. *Adv. Protein Chem. Struct. Biol.* **118** (2019) 289-331.

[45] M. Kanapathipillai. Treating p53 mutant aggregation-associated cancer. *Cancers (Basel)* **10** (2018) 154.

[46] P. Sormanni, F.A. Aprile, M. Vendruscolo. The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **427** (2014) 478-490.

[47] Y. Tian, C. Deutsch, B. Krishnamoorthy. Scoring function to predict solubility mutagenesis. *Algorithms Mol. Biol.* **5** (2010) 33.

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

[49] R. Guerois, J.E. Nielsen, L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320** (2002) 369-387.

[50] H. Land, M.S. Humble. YASARA: A tool to obtain structural guidance in biocatalytic investigations. *Methods Mol. Biol.* **1685** (2018) 43-67.

[51] C.J. O'Brien, M.A. Blanco, J.A. Costanzo, M. Enterline, E.J. Fernandez, A.S. Robinson, C.J. Roberts. Modulating non-native aggregation and electrostatic protein-protein interactions with computationally designed single-point mutations. *Protein Eng. Des. Sel.* **29** (2016) 231-243.

[52] M. Gil-Garcia, M. Bano-Polo, N. Varejao, M. Jamroz, A. Kuriata, M. Diaz-Caballero, J. Lascorz, B. Morel, S. Navarro, D. Reverter, S. Kmiecik, S. Ventura. Combining structural aggregation propensity and stability predictions to redesign protein solubility. *Mol. Pharm.* **15** (2018) 3846-3859.

[53] M.A. Carballo-Amador, E.A. McKenzie, A.J. Dickson, J. Warwicker. Surface patches on recombinant erythropoietin predict protein solubility: engineering proteins to minimise aggregation. *BMC Biotechnol.* **19** (2019) 26.

[54] S. Navarro, S. Ventura. Computational re-design of protein structures to improve solubility. *Expert Opin. Drug Discov.* **14** (2019) 1077-1088.

[55] N. Chennamsetty, V. Voynov, V. Kayser, B. Helk, B.L. Trout. Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U. S. A.* **106** (2009) 11937-11942.

[56] J.R. Klesmith, J.P. Bacik, E.E. Wrenbeck, R. Michalczyk, T.A. Whitehead. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl. Acad. Sci. U. S. A.* **114** (2017) 2265-2270.