Article

The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility

Pietro Sormanni, Francesco A. Aprile and Michele Vendruscolo

Department of Chemistry, *University of Cambridge, Cambridge CB2 1EW, UK*

Correspondence to Michele Vendruscolo: *mv245@cam.ac.uk.*
http://dx.doi.org/10.1016/j.jmb.2014.09.026
Edited by S. Radford

Abstract

Protein solubility is often an essential requirement in biotechnological and biomedical applications. Great advances in understanding the principles that determine this specific property of proteins have been made during the past decade, in particular concerning the physicochemical characteristics of their constituent amino acids. By exploiting these advances, we present the CamSol method for the rational design of protein variants with enhanced solubility. The method works by performing a rapid computational screening of tens of thousand of mutations to identify those with the greatest impact on the solubility of the target protein while maintaining its native state and biological activity. The application to a single-domain antibody that targets the Alzheimer's Aβ peptide demonstrates that the method predicts with great accuracy solubility changes upon mutation, thus offering a cost-effective strategy to help the production of soluble proteins for academic and industrial purposes.

Introduction

Proteins are attractive diagnostic and therapeutic molecules because of their functional versatility and specificity, as well as their inherently low toxicity [1–4]. Antibodies, in particular, can be obtained with well-established methods, including immunization or phage and associated display methods, against virtually any target of therapeutic interest, which they bind with high affinity and specificity [5–9]. The importance of protein drugs is rapidly increasing, as they can be used to replace or supplement endogenous proteins (e.g. insulin, growth hormone, interleukins) and to treat a wide range of diseases, including cancer and autoimmune disorders [1–4,10]. Since protein drugs are generally not orally active, their preferred delivery method is subcutaneous delivery, which requires that a large amount of proteins is stored in small volumes (<2 ml), corresponding to highly concentrated formulations (≥ 50 mg/ml) that favor aggregation.

The maintenance of proteins in a soluble state is indeed an essential aspect in diagnostic and therapeutic applications [9,11–15], as well as being a fundamental requirement for protein homeostasis in living organisms [16–18]. Many proteins, however, have a strong tendency to aggregate, and therefore to lose their activity, in particular if brought under conditions that differ from those in their native cellular environments [19]. This problem affects particularly the recombinant expression of proteins, resulting in insoluble protein aggregates in many cases, such as inclusion bodies [20,21]. Thus, protein aggregation represents also a major biotechnological issue, preventing many proteins to be produced at economically convenient yields [13,20,22]. Effective experimental approaches to improve protein solubility during recombinant expression include the use of weak promoters, modified growth media, low temperatures, and solubility-enhancing tags [23–25] or large-scale screening and random mutagenesis [26,27].

More generally, insufficient solubility represents a major bottleneck for the development of protein-based drugs, as protein aggregates not only are non-functional but also can be toxic and may elicit an immune response in the patient [14,28]. In particular, as antibodies can be poorly soluble, there is a need of developing methods to increase their solubility in order to exploit their full potential in therapeutic applications. The maintenance of

solubility is particularly challenging in the case of these molecules because they should bind strongly their molecular targets and, in order to do so, they must expose on their surface aggregation-prone patches of amino acids. The goal is thus to find ways to maintain the high binding affinity and specificity properties of antibodies while minimizing their tendency to aggregate. Standard approaches to achieve this objective are based on the optimization of the solubility by phage display and heat denaturation [7,12,15,29,30]. In this way, a great number of variants are produced by random mutagenesis and the most soluble forms are selected. In order to reduce costs and time, it would be desirable to develop alternative methods in which the screening is performed by rational design. Strategies on this type based on expert analysis of antibody structures have been proposed [11,31–33]. In this context, because of the combinatorial nature of the problem, the use of computational methods is particularly convenient as the number of mutational variants that can be screened in this way is very large, as demonstrated by recent studies [34].

In this work, we build on recent advances in understanding the fundamental principles of protein aggregation [35–37] and of protein solubility [38–41], in particular of antibodies [9,33,34,42,43], to develop the CamSol method to design rationally protein variants with enhanced solubility. We illustrate this method in the case of a recently described single-domain antibody that binds the Aβ peptide [32]. The use of single-domain antibodies is attracting attention because these molecules can exhibit high affinity and specificity to their targets without the complications associated with the complex architecture of full-length antibodies [12,44]. We show that predicted and measured solubility values are highly correlated, thus demonstrating that the CamSol method offers a powerful alternative to experimental strategies in selecting soluble variants, as it can screen tens of thousands of candidate mutations in just a few minutes on a standard laptop[‡].

Results

The CamSol method

In this work, we describe the CamSol method of structure-based design of soluble protein variants. The method exploits recent advances in understanding the physicochemical properties of amino acids most directly responsible for the solubility of proteins [35–38,41,45,46], including the hydrophobicity, the electrostatic charges, and the interplay in their spatial patterning. In essence, by defining a solubility score through a phenomenological combination of these properties, the method performs a rapid and systematic computational screening of tens of thousands of amino acid substitutions or insertions to identify specific mutations that are predicted to maximally increase the solubility of a protein while preserving its fundamental properties, including its functional structure and binding affinity.

The method requires the knowledge of the native structure of the target protein, which could be available by experimental or by computational (e.g. homology modeling) techniques (see Materials and Methods). From the structure, one can distinguish, among the residues that are classified as poorly soluble, those that are required for functional reasons (e.g. the residues that form the hydrophobic core) from those that remain exposed to the solvent and are not strictly necessary. One can also provide a list of residues important for function or that cannot be otherwise mutated and the maximum number of mutations that the algorithm is allowed to perform so that the wild-type sequence is largely conserved.

More in detail, the CamSol method comprises the following four steps: (i) Calculation of the residue-specific intrinsic solubility profile, (ii) calculation of the structural correction to the intrinsic solubility profile, (iii) identification of suitable mutation sites using the structurally corrected solubility profile, and (iv) screening of all possible variants to identify the most soluble one using an overall intrinsic solubility score.

(i) In the calculation of the intrinsic solubility profiles, which relies solely on the knowledge of the amino acid sequence of the protein to be solubilized, we exploit the connection between protein aggregation propensity and protein solubility. From the point of view of thermodynamics, these concepts are distinct, as the solubility depends on the free energy difference between the monomeric and aggregated states, whereas the aggregation rate depends on the free energy barrier between these states (Fig. S1). In practice, however, methods of predicting aggregation rates have been used quite effectively in predicting also the solubility of proteins [41]. In order to obtain a "solubility profile", that is, a function that associates to each residue in an amino acid sequence a number that reflects its impact on the overall solubility, in this work, we used a strategy similar to that of the Zyggregator method of predicting protein aggregation propensity [36,45]. In the current implementation of the CamSol method, we employed a linear combination of specific physicochemical properties of amino acids, such as hydrophobicity and

electrostatic charge, which we smoothed over a window of seven residues to account for the effect of the neighboring residues [see Materials and Methods and Eqs. (1) and (2)]. The major difference with the Zyggregator method is that, in the CamSol calculations, we used a different set of parameters in order to remove the bias toward predicting amyloid-like aggregation (see Materials and Methods). Moreover, since with CamSol we aim to compute sequence-based solubility profiles, we changed the sign convention (with respect to Zyggregator) so that increasingly negative profiles represent increasingly insoluble regions, while increasingly positive profiles represent increasingly soluble ones. This residue-specific intrinsic solubility profile is not only the starting point for the calculation of the structural corrections in step (ii) but it is also used to calculate a single intrinsic solubility score for each protein variant (see Calculation of the CamSol solubility score). It is the variation of the latter that is expected to be proportional to the solubility change upon mutation, provided that the amino acid substitution and insertion sites have been selected as described in step (iii).

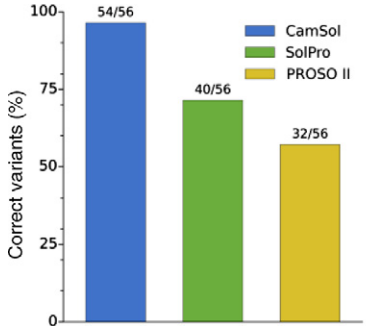
(ii) In the structural corrections, the intrinsic solubility profiles [calculated at step (i)] are modified to account for the proximity of the amino acids in the three-dimensional structure and for their solvent exposure [47] (see Calculation of the structurally corrected solubility profiles). These modified profiles are used to distinguish the poorly soluble residues that are required for fast and correct folding, such as the residues that form the hydrophobic cores of the native states, from the ones that are exposed to the solvent in the native state and might elicit the aggregation process.

(iii) In the identification of the most suitable sites for amino acid substitution or insertion, the regions containing poorly soluble solvent-exposed residues are analyzed using the structurally corrected solubility scores [calculated at step (ii)]. Such sites are selected using three criteria, as they should be the following: (a) within or close to poorly soluble regions, (b) exposed to the solvent, and (c) far from residues important for activity (see Selection of the sites for mutation).

(iv) Once suitable sites are selected, all possible amino acid substitutions and/or insertions are screened systematically using the intrinsic solubility score, and the most soluble variant is identified. In this step, since the selection of the most suitable sites for mutations [step (iii)] is carried out using the structurally corrected solubility score, it is sufficient, and computationally more convenient, to use the intrinsic solubility score.

Test of the CamSol solubility predictions on a database of protein variants

We initially tested the accuracy of the CamSol method by predicting the effects of different types of mutations on the solubility of a variety of proteins (Fig. 1 and Table S1). We considered one set of different proteins [48] and three sets of antibodies [43,49,50] collected from the literature, for a total of 56 protein variants whose solubility—or a closely



Reference	Proteins	Variants	Correct CamSol	Correct SolPro	Correct PROSO II	Antibody scaffolds
Trevino	15	22	22	15	16	no
Miklos	1	3	3	3	3	yes
Tan	1	1	1	1	1	yes
Dudgeon	2	30	28	21	12	yes

Fig. 1. Test of the CamSol solubility score. In the test, we considered a database composed of four different sets of protein variants for which solubility measures are available ("Trevino" [48], "Miklos" [49], "Tan" [50], and "Dudgeon" [43]). The bar plot reports the fraction of correctly predicted solubility changes upon mutation using CamSol, SOLpro, and PROSO II.

related quantity—had been measured (Fig. 1 and Table S1). These variants are derived from 19 different wild-type proteins of which at least the sequence is available. Given the variety of techniques employed in the different studies to assess the impact of the mutations on the solubility, we restricted our predictions to whether a mutation increases or decreases the solubility. Moreover, as mutation sites had been selected by the authors of the different works, we did not employ any structural correction in this analysis, which assesses the goodness of the CamSol solubility score (see Materials and Methods) in predicting solubility changes upon mutation.

Our results indicate that the CamSol method is highly accurate in predicting the effects of mutations on protein solubility (Fig. 1). When compared with other existing methods aimed at predicting solubility upon overexpression, such as SOLpro [39] and PROSO II [40], CamSol predicts correctly the change in solubility upon mutation for 54 out of 56 variants, compared with 40 and 32 of SOLpro and PROSO II, respectively (Fig. 1 and Table S1).

Test of the CamSol method for the rational design of soluble gammabody variants

In order to illustrate the use of the CamSol method for predicting mutations that enhance protein solubility, we applied it to a single-domain antibody in which the peptide corresponding to residues 33–42 of the Aβ42 peptide is grafted into the complementarity-determining region (CDR) 3. This single-domain antibody has been referred to as a "grafted amyloid-motif antibody" or "gammabody" Aβ(33-42) (Table S2) [32] and its solubility has been reported in the low-micromolar regime [32].

Following the CamSol procedure, we selected the optimal sites for the mutations, in this case, giving a maximum of three simultaneous mutations as a constraint in order to perform only minimal changes to the wild-type gammabody. This specific procedure identified a total of two sites for substitution (Q118 and

Table 1. List of the seven rationally designed gammabody variants derived in this work to illustrate the performance of the CamSol method

Variant name	Description	Solubility score	Monomer concentration (μM)
WT	Wild-type gammabody	−0.569	10.31 ± 0.04
QK	Q118K mutant	−0.524	14.2 ± 0.3
LE	L121E mutant	−0.531	15.8 ± 0.2
QKLE	Q118K-L121E mutant	−0.497	18.7 ± 0.3
NNK	Three-residue (NNK) insertion at A112	−0.496	21 ± 1
EEP	Three-residue (EEP) insertion at A112	−0.447	25.0 ± 0.8
DED	Three-residue (DED) insertion at A112	−0.432	29.9 ± 0.9
EEE	Three-residue (EEE) insertion at A112	−0.399	32.1 ± 0.8

We calculated the solubility score (see Materials and Methods) and measured the monomer concentration after 1 h of incubation at room temperature starting at 70 μM (see Materials and Methods).

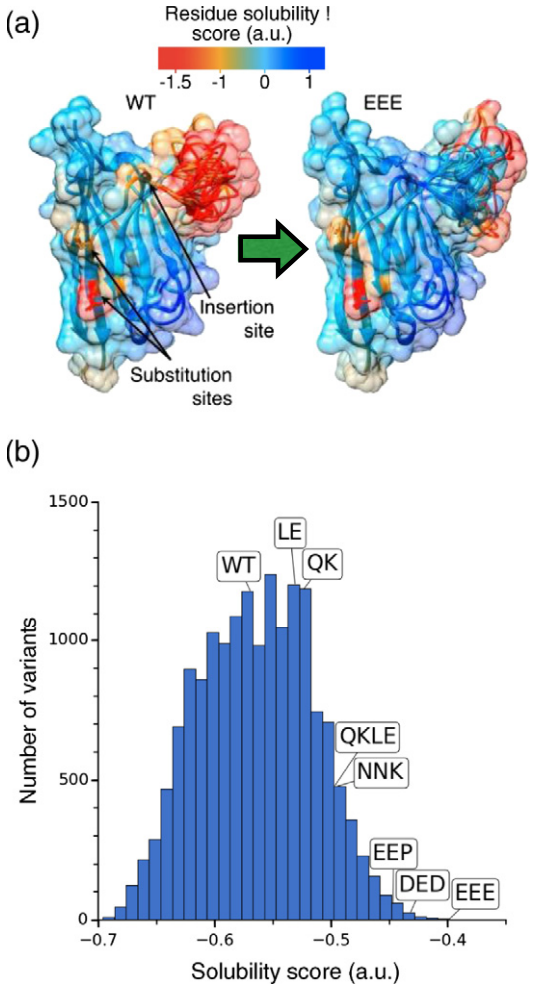


Fig. 2. Rational design of soluble gammabody Aβ(33-42) variants. (a) Schematic illustration of the CamSol method as applied to the gammabody Aβ(33-42). The structurally corrected solubility profile [Eq. (S1)] is color coded on the surface of the wild-type gammabody (left) and on the variant predicted to be the most soluble (EEE right, see Table 1). Arrows on the wild-type structure point to the sites selected for amino acid substitutions and insertions. Both structures are obtained with homology modeling (see Materials and Methods). (b) Distribution of the protein solubility scores for all possible combinations of mutations and/or insertions at the selected sites. A maximum of three residues was simultaneously changed. The seven variants chosen for experimental validation are flagged (Table 1).

L121) and one site for insertion (A112; Table 1 and Fig. 2a). We then scanned systematically the solubility scores of 16,440 possible combinations of mutations and/or insertions at these selected sites (8000 for a triple insertion at A112, 8000 for a double insertion at A112 plus a mutation in L121, 400 for a double mutation in Q118 and L121, and 20 + 20 for individual mutations in Q118 and in L121; Fig. 2b). The screening of all these mutational variants required less than 1 min on a laptop computer. The distribution in Fig. 2b is not symmetric about the wild-type sequence because the amino acid substitution and insertion sites are chosen as close as possible to the poorly soluble regions so to maximize the impact on the solubility of the protein. Thus, we identified a set of seven gammabody mutants with a predicted increased solubility with respect to the wild-type form (Table 1 and Table S1). No particular rule was applied in selecting these seven gammabody variants. We decided to include in the set the wild type (WT in Table 1), the variant predicted to be the most soluble (EEE), one variant for each of the identified amino acid substitution sites (QK and LE), the double mutant (QKLE), and three additional variants with a predicted solubility that was approximately equally spaced in the gaps between the already selected variants (NNK, EEP, DED).

Structural and functional characterization of the rationally designed gammabodies

The purity of all the seven gammabody mutational variants described above was characterized by NuPAGE analysis (see Materials and Methods and Fig. S2a) and their structural integrity by far-ultraviolet (far-UV) circular dichroism (CD) spectroscopy at 25 °C (see Materials and Methods and Fig. S2b). None of the mutational variants showed significant differences in the CD spectra with respect to the wild-type gammabody, indicating that they are all able to populate the wild-type scaffold structure.

In order to evaluate the influence of the mutations on antigen binding, we first characterized by dot blot assay the ability of all the mutational variants to bind the Aβ42 peptide (see Materials and Methods). Also in this case, we did not observe significant differences in the binding capability of the mutational variants (Fig. S3a). To prove that insertions flanking the CDR3 do not affect significantly the affinity and thus efficacy of the designed antibodies, we perform an ELISA test for the binding to Aβ42 on the wild-type and EEE gammabodies by titrating increasing quantities of Aβ42 into solutions containing the gammabody variants (see Materials and Methods). The two variants show a very similar concentration-dependent protein-binding curve, proving that the binding affinity of the engineered antibody has not been altered substantially (Fig. S3b).

Taken together, these results imply that the mutations introduced in the wild-type gammabody, at least in the cases that we tested, do not affect its structure and functionality.

Protein solubility measurements

In order to characterize the solubility of the gammabody variants, we estimated their saturation (or critical) concentration, that is, the concentration above which monomeric gammabody addition does not result in an increase of the concentration of the soluble species. To this end, we prepared gammabody samples at different concentrations and plotted them against the concentrations of the supernatant measured after centrifugation at 90,000 rpm for 1 h (Fig. 3). At variance with what found for most inorganic molecules or for other proteins [51], no clear plateau was reached in our measurements, even at concentrations corresponding to large amounts of precipitation (Fig. 3a). This observation

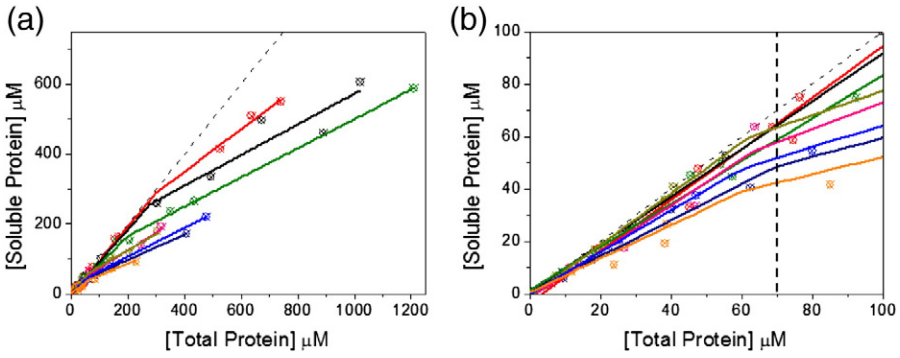


Fig. 3. Saturation concentration analysis for the gammabody variants. (a) Supernatant concentration upon centrifugation at 90,000 rpm as a function of the total protein concentration for all the seven gammabody variants tested in this work: EEE (red), DED (black), EEP (dark green), NNK (pink), QK/LE (light blue), LE (orange), QK (light green), and WT (dark blue). (b) Detail from (a) to describe the supernatant concentration at 70 μM total protein concentration (vertical broken line); lines are a guide to the eye.

arises from the formation of different oligomeric species during the oligomerization process (see below).

At first, we considered whether we could define as the saturation concentration the value of the total protein concentration at which the supernatant concentration deviates from the line of slope 1 (Fig. 3). However, this approach would yield results dependent on the speed of centrifugation and affected by large errors, which we estimate to be about 20% from five independent measurements. As a consequence, we measured the solubility as the amount of monomeric species in solution. We measured this quantity by analytical size-exclusion chromatography (SEC; see Materials and Methods), which represents an ideal technique to analyze the distribution of populations in solution [52]. Since the SEC analysis can be inaccurate in the presence of large aggregates, we selected an initial protein concentration of 70 μM, as no significant precipitation was observed for any of the variants at this concentration (Fig. 3b).

As the relative populations of monomeric and oligomeric species depend on the total concentration of the protein, we carried out a dynamic light scattering analysis (see Materials and Methods) of the two mutational variants with the weakest tendency to precipitate (i.e., EEE and DED; Table 1) as determined in the saturation concentration analysis (Fig. 3a). The apparent hydrodynamic radius, R_H, measured by dynamic light scattering does not correspond to individual protein species (Fig. S4). As a consequence, the variation of the apparent R_H reflects variations in the relative populations of the different species present in solution. We found that above a total protein concentration of about 70–100 μM the apparent R_H values reach a plateau, indicating that above this value the distribution of all the species in solution is no longer dependent on the total protein concentration (Fig. S4). Accordingly, we found that above this concentration the monomer population does not decrease further (Fig. S4). Moreover, the variation of the R_H for the wild-type gammabody was observed to be below the detection limit of the instrument that we used (at concentrations lower than 10 μM; Fig. S4). Since the assay presented in Fig. 3 suggests that, within experimental errors, all other variants have a solubility value between the one of the wild type and the one of the DED/EEE variants, they are expected to reach the R_H plateau at concentrations between 10 and 70 μM. Therefore, we concluded that, at a concentration of 70 μM, no significant protein precipitation is present and the distribution of soluble species has reached a plateau with respect to the total protein concentration.

Determination of the solubility of the rationally designed gammabody variants

Given the results described in the previous section, we used analytical SEC to determine the solubility of the different gammabodies by estimating the monomer populations in samples at 70 μM total gammabody concentration. The SEC analysis was performed at 4 °C to minimize the dissociation of the assemblies and to estimate the populations in the samples before the injection into the column.

In order to identify the peak corresponding to the native monomer, we compared the apparent R_H values derived from the retention volumes of the gel filtration with the one obtained with the program HydroPRO [53]. The latter was estimated to be 26.3 ± 1.6 Å, using 23 homology models to account for the different conformations of the C-terminal tag and disordered CDR loops (see Materials and Methods). This value was compared with the ones calculated from the positions of the peaks in the chromatograms from the SEC analysis. For example, the chromatogram of the DED gammabody variant shows four peaks (colored circles and numbers in Fig. 4). The major peak is at 9.3 ml (green circle labeled as "2") and corresponds to the retention volume of a protein with an apparent R_H of 32 Å, compatible with the value of the monomeric gammabody estimated with HydroPRO. The other three peaks are at 10.7 ml (R_H of 24 Å, blue circle labeled as "3"), also compatible with the R_H of the monomeric protein, at 8 ml (R_H of 43 Å, yellow circle labeled as "1"), and at 15.5 ml (R_H 13 Å, red circle labeled as "4").

Native PAGE analysis on the corresponding eluted fractions revealed the presence of monomers and dimers in peaks 1, 2, and 3. The highest amount of monomeric protein is found in peak 2, while peak 3

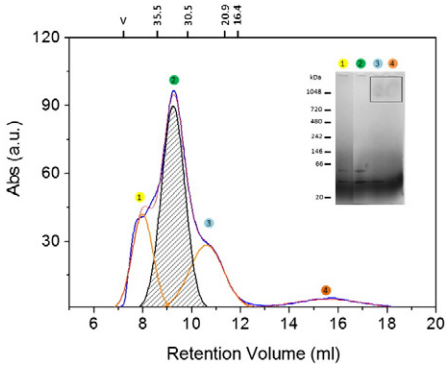


Fig. 4. SEC elution profile of the DED gammabody. The SEC elution of the DED gammabody profile (blue line) was fitted to a multi-peak Gaussian function (red line) to evaluate the relative fractions of monomeric and oligomeric gammabody species (see Materials and Methods). The elution peak derived from this analysis as corresponding to monomeric gammabodies is shown as a shaded area. Eluted fractions analyzed by native PAGE (shown in inset) are marked with colored dots: 8 ml (yellow circle labeled as "1"), 9.3 ml (green circle labeled as "2"), 10.5 ml (blue circle labeled as "3"), and 15.5 ml (red circle labeled as "4"). The R_H values of the standard proteins used to calibrate the SEC column (see Materials and Methods) are reported at the top of the chromatogram.

also contains high-molecular-weight protein species with abnormal migration on the gel. These species are present as well in peak 4 and the fact that they show higher retention volumes than the monomer indicates that they are assemblies able to interact with the matrix of the SEC column. Similarly, peak 1 is close to the void volume (7.3 ml) and may contain protein oligomers and aggregates too large to enter into the pores of the acrylamide gel. The presence of dimers and oligomers alongside monomers can be explained with a fast dynamic equilibrium between these species that leads to the formation of the corresponding populations on the experimental timescale of the native PAGE analysis [54]. For these reasons, since our scope is to determine the relative distributions of the monomeric gammabody populations, we considered only the main peak at 9.3 ml (Fig. 4). To this end, we incubated all the gammabody variants at 25 °C for 3 h at a concentration of 70 μM and then injected them into the chromatography column. For each variant, at least three independent analyses were performed and all the replicates were globally fitted to a multi-peak Gaussian function with only the height of the peaks as a floating parameter (Fig. S5). The chromatograms of different variants show peaks at slightly different retentions volumes and resolution, including the peak corresponding to the monomer (Fig. S6). This is most likely the consequence of the interaction of the proteins with the matrix of the SEC column, as previously reported for similar constructs [32], since all variants are very similar in structure and size (Fig. S2). Therefore, in order to account for the fact that different variants may interact differently with the matrix of the gel-filtration column, we fitted each variant independently.

Correlation between predicted and measured solubility values

To illustrate the overall results of the application of the CamSol procedure to increase the solubility of gammabody Aβ(33-42), in Fig. 5a, we show the structures of the mutational variants calculated by homology modeling using Modeller [55,56] (see Materials and Methods). For each gammabody variant, the surface of the model with the lowest discrete optimized protein energy (DOPE) score (see Materials and Methods) is color coded with the structurally corrected solubility profile (see Materials and Methods), while it is transparent for all other models. The increasing solubility of the gammabody variants is reflected by the decreasing amount of poorly soluble surface regions (shown in red in Fig. 5a). On a more quantitative level, we found an excellent correlation (coefficient of correlation R^2 = 0.98) between predicted and measured solubility values (Fig. 5b), indicating that the CamSol prediction offers an effective alternative to costly and time-consuming experimental measurements of protein solubility.

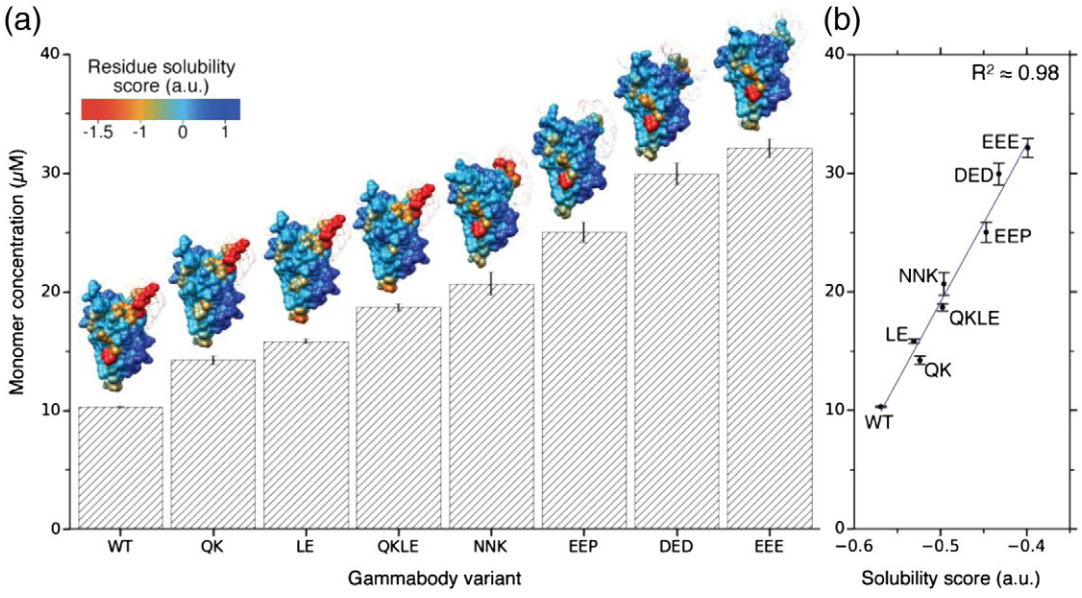


Fig. 5. Correlation between predicted and measured solubility values. (a) Bar plot of the measured monomer concentrations after incubation at a total concentration of 70 μM. Homology-derived structures of the variants are represented on top of the corresponding bars. The surface of the model with the best homology modeling score (the DOPE score [55,56]; see Materials and Methods) is color coded with the structurally corrected solubility profile, while it is transparent for all other models. (b) Correlation of the CamSol solubility scores (*x*-axis) as a function of the measured monomer concentrations (*y*-axis) of the different gammabody variants.

Conclusions

We have described the CamSol method of performing a rational design of protein mutational variants with enhanced solubility, and validated its predictions on a dataset of solubility changes upon mutation obtained from the literature. Through the application of this method to a single-domain antibody against the Aβ peptide, we have shown that it can readily provide highly soluble mutational variants.

The solubility score provided by the CamSol method can be exploited to rank libraries of protein variants according to their solubility. For example, *in vitro* antibody discovery techniques (e.g. phage display) usually yield a large number of antibody variants that bind to the desired epitope with high affinity. Since these variants share a high degree of sequence similarity, we expect that the CamSol method will produce accurate solubility rankings, reducing the need for experiments and helping the selection of the best candidate. We also anticipate that the CamSol method will be generally applicable to a wide range of proteins for which a structure or a homology model is available and will represent a cost-effective way to obtain soluble mutational variants of proteins of biotechnological and therapeutic interest.

Materials and Methods

Increasing the solubility of proteins with the CamSol method

Input

In order to be applied to increasing the solubility of a target protein, the CamSol method requires a knowledge of its native structure. This structure is needed to distinguish the poorly soluble residues required for fast and correct folding (e.g. the residues that form the hydrophobic core) from those that remain exposed to the solvent and might elicit the aggregation process. As additional input, one can provide a list of residues important for function or that cannot be otherwise mutated and the maximum number of mutations that the algorithm is allowed to perform so that the wild-type sequence is not changed too much.

Calculation of the intrinsic solubility profiles

In the first step, a score is assigned to each residue in the wild-type sequence to identify the residues that most affect the solubility.

Methods of calculating solubility scores have been proposed on the basis of machine-learning approaches trained on experimental databases of heterologous expression [38–41]. Although we plan to use similar approaches in the future, in this work we have exploited the connection between the aggregation propensity and the solubility of proteins. These two concepts are thermodynamically distinct (Fig. S1), but they are linked if one assumes that the free energy barriers are correlated with the stabilities. Indeed, methods of predicting aggregation propensity have been employed to predict protein solubility as well [41]. The use of databases of solubility-related measurements enables one to avoid making this approximation, but at the moment such databases are still of poor quality or insufficient size. We thus expect to be able to improve even further the CamSol predictions using machine-learning methods based on solubility measurements.

We thus use a strategy similar to that of the Zyggregator method of predicting protein aggregation propensity profiles [36,45]. As in the Zyggregator method, an initial score is assigned to each residue in the form of a linear combination of specific physicochemical properties

$$s_i = a_{\mathrm{H}} p_i^{\mathrm{H}} + a_{\mathrm{C}} p_i^{\mathrm{C}} + a_{\alpha} p_i^{\alpha} + a_{\beta} p_i^{\beta} \quad (1)$$

where p_i^{H}, p_i^{C}, p_i^{α} and p_i^{β} are the hydrophobicity, the charge (at neutral pH), the α-helix propensity, and the β-strand propensity of residue i, respectively, while a values are the parameters of the linear combination (Table S3). Then, in order to account for the effect of neighboring amino acids, the profile is smoothed over a seven-residue window and a correction is added to consider the possible presence of hydrophobic–hydrophilic patterns and the influence of charges of the same sign

$$S_i = \frac{1}{7}\left(\sum_{j=i-3}^{i+3} s_j\right) + a_{\mathrm{pat}} I_i^{\mathrm{pat}} + a_{\mathrm{gk}} I_i^{\mathrm{gk}} \quad (2)$$

where I_i^{pat} accounts for the presence of specific patterns of alternating hydrophobic and hydrophilic residues and it is described in Refs. [36] and [45], while I_i^{gk} takes into account the gatekeeping effect of individual charges. At variance with the Zyggregator method, the term I_i^{gk} has been refined here to encompass the relative distance of charged residues along the sequence

$$I_i^{\mathrm{gk}} = \sum_{j=-5}^{5} e^{-\frac{j^4}{200}} C_{i+j} \quad (3)$$

where C_{i+j} is the charge of the amino acid $i + j$.

Differently from Zyggregator, in Eq. (1), the CamSol method uses secondary structure propensities calculated from the Protein Data Bank (PDB) using representative structures at a 50% sequence identity and a hydrophobicity scale adapted using the Wimley–White scale [57] (Table S3). With this change, we removed the bias toward predicting amyloid-like aggregation and, in fact, the largest change in the behavior of individual amino acids is observed for proline and glycine residues because these two amino acids disfavor β-strand formation and thus act against amyloid formation, but their impact on solubility is weaker. Moreover, all signs where inverted so that larger S_i scores indicate a larger contribution of the ith residue to the predicted solubility.

Since the scores computed are dimensionless numbers, the solubility profiles are rescaled so that a random polypeptide yields a profile with mean 0 and standard deviation 1, using the procedure described in Ref. [36] ($S_i = [S_i - \mu_{\mathrm{random}}]/\sigma_{\mathrm{random}}$). Accordingly, amino acids with

a score smaller than −1 are regarded as poorly soluble and have a negative impact on the solubility of a protein, while scores larger than 1 denote highly soluble regions, yielding a positive contribution to the overall solubility.

Calculation of the structurally corrected solubility profiles

The structurally corrected solubility profile, or surface solubility propensity, is defined by projecting the intrinsic solubility profile onto the surface and smoothing over a surface patch of size A and dimension r_A. The structurally corrected solubility propensity score S_i^{surf} of residue i can be written as [47]

$$S_i^{\text{surf}} = w_i^{\text{E}}\left(\widetilde{S}_i^{\text{int}} + \sum_{j\notin[i-3,i+3]} w_j^{\text{D}} w_j^{\text{E}} \widetilde{S}_j^{\text{int}}\right) \quad (4)$$

where the sum is extended over all the residues of the protein within a distance r_A from residue i, excluding the residues that are contiguous along the sequence, as their proximity effect is already encompassed by the intrinsic solubility score. Respectively, w_i^{E} and w_i^{D} are the "exposure weight", which depends on the solvent exposure of residue j, and the "smoothing weight", defined as

$$w_j^{\text{D}} = \max\left(1-\frac{d_{ij}}{r_A}, 0\right) \quad (5)$$

where d_{ij} is the distance of residue j from residue i. This definition implies that neighboring residues contribute more to the local surface aggregation propensity than more distant ones. Furthermore, the smoothing weight does not bias toward a preselected surface patch size. In the present work, we set r_A to be 8 Å, as this value is consistent with the seven-amino-acid windows implemented in the prediction of the intrinsic solubility profile (in fact, a distance of 8 Å spans approximately three residues in a compact globular protein).

The exposure weight is defined as

$$w_j^{\text{E}} = \frac{\vartheta(x_j - 0.05)}{1 + e^{-a(x_j-b)}} \quad (6)$$

where x_j is the relative exposure of residue j, that is, the SASA (*s*olvent-*a*ccessible *s*urface *a*rea) of residue j in the given structure divided by the SASA of the same residue in a Gly-Xxx-Gly peptide in an extended conformation. The Heaviside step function, ϑ, is employed so that residues with less than 5% solvent exposure are not taken into account. Equation (6) thus describes a sigmoidal function, where a and b are parameters tuned so that the weight grows slowly up to a relative exposure $x \approx 20\%$ and then grows linearly reaching 1 at $x \approx 50\%$; this is accomplished by setting $a = -10$ and $b = 0.3$. When a residue is 50% solvent exposed, half of it faces inward in the structure while the other half, facing the solvent, already provides the largest surface for potential aggregation partners. With this correction, residues not exposed to the surface, such as those buried in the hydrophobic core and essential for the folding of a protein, are assigned a score close to 0 and, consequently, are not considered in the subsequent steps of the CamSol algorithm.

The quantity $\widetilde{S}_j^{\text{int}}$ in Eq. (4) is the intrinsic solubility of residue j computed using a modified version of Eq. (2), which reads

$$\widetilde{S}_i = \frac{1}{\sum_{j=i-3}^{i+3}\widetilde{x}_j}\left(\sum_{j=i-3}^{i+3}\widetilde{x}_j s_j\right) + a_{\text{pat}} I_i^{\text{pat}} + a_{\text{gk}}\widetilde{I}_i^{\text{gk}} \quad (7)$$

In essence, the average over the seven-residue window in Eq. (2) has been replaced here by a weighted average (over the same window) with weights $\widetilde{x}_j$, which are the relative exposures of residue j linearly rescaled in the range [0.25,1], so that division by 0 never occurs.

Similarly, $\tilde{I}_i^{\text{gk}}$ embodies the same idea as I_i^{gk} in Eq. (3), but the gatekeeping effect of charges of the same sign is now computed in the three-dimensional space

$$\widetilde{I}_i^{\text{gk}} = \sum_j w_j^{\text{D}}(d_{ij}, 2r_A) w_j^{\text{E}}\left(x_j^{\text{C}}\right) C_j \quad (8)$$

where C_j is the net charge of residue j at neutral pH, and the smoothing weight w_j^{D} is computed here using twice the patch radius r_A and the exposure weight w_j^{E} using the relative exposure x_j^{C} of the charged atom in residue j.

The calculation of the structurally corrected solubility profile requires the knowledge of the structure of the protein. However, there is no need for particularly high resolution. The predictions are accurate as long as the solvent exposure of the amino acids and their relative C^{α} distances are correctly represented. This fact makes the CamSol procedure applicable to a large number of cases where only the sequence is known, as a good-enough structure can be obtained by standard techniques, such as homology modeling (see below).

Calculation of the CamSol solubility score

Despite the approximate correspondence between solubility and aggregation [41], the overall solubility score of a protein does not represent an aggregation rate because, since the solubility is a thermodynamic parameter, different proteins could aggregate at different rates but have a similar solubility (Fig. S1). From the intrinsic solubility profile, we derive an overall solubility score for the whole protein

$$S_P = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} S_i & \text{if } S_i<-0.7 \text{ or } S_i>0.7 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where N is the length of the protein sequence. The choice of the ±0.7 thresholds is empirical and reflects the fact that residues with a score close to 0 have a negligible effect on the solubility of the protein, while residues with scores close to or less than −1 are poorly soluble. Similarly, residues with score close to or larger than 1 contribute positively to the solubility of the protein. The threshold ±0.7 was chosen over ±1 to yield a solubility score more sensitive to single mutations, as predicting the solubility change upon mutation is the aim of the CamSol method.

Selection of the sites for mutation

In this step, specific positions in the wild-type sequence are selected as candidates for residue substitutions or insertions. As mentioned above, a list

of residues important for function or that cannot be otherwise mutated can be provided as input. The knowledge of the residues required for protein activity does not, however, have to be very accurate. As long as false positives are favored over false negatives and hence functionality is preserved, solubility-enhancing mutations can be found in general.

As in the CamSol method, we aim to select positions at which amino acid substitutions or insertions maximally impact the solubility of the protein, we proceed as follows:

First, we select the sequence fragments on the structurally corrected profile containing at least a residue with a score smaller than −1. Each of these fragments is then assigned an overall score given by the sum of the scores of the residues contained in it (i.e., the integral under the profile), which is used to rank the fragments accounting for both their length (the size of the "dangerous" region) and their solubility scores (how insoluble its components are). We then sort these ranked fragments from less soluble to more soluble.

Second, we loop through the ensemble of fragments identified and flag as immutable those residues that are required for function (if given as input). Especially in the case of antibodies, after this filtering, some of the fragments can be completely immutable, as it is quite common for poorly soluble residues to be found within the CDR loops in the binding site. Nevertheless, we conserve the fragment positions and their scores also in these cases.

Third, we scan through the ensemble one fragment at a time, and if the fragment still contains some mutable residues, their positions in the sequence are flagged for residue substitutions; otherwise, we look at the positions at the side of the (immutable) fragment. This step exploits the fact that the presence of highly soluble residues (such as charged residues) has an effect on the solubility profile of the region that contains them. Hence, mutating soluble-neutral residues to highly soluble ones at the sides of a poorly soluble fragment can significantly increase the solubility profile of the whole region. For this reason, we look at the positions of these adjacent residues in the structure. If the amino acid under scrutiny is solvent exposed and it is not involved in interactions important for the overall stability of the protein (such as salt bridges or hydrogen bonds), its position is flagged for mutation. Otherwise, if the adjacent amino acids are not solvent exposed or form important interactions, the sides of the insoluble fragment are labeled as possible sites for insertions.

Screening of all possible mutants

The procedure described above provides a list of positions, mapped on both the sequence and the structure, suitable for mutations and/or insertions. Each position also has a score (the one given to the fragments) that reflects how large the effect on solubility of a substitution/insertion at that site is expected to be (small score, large effect). Mutations sites are therefore ranked using this score.

At this point, a choice needs to be made by the user. While it could be desirable to perform mutations at every position in order to maximize the solubility of the resulting protein, too many mutations could alter other properties of the protein in an unwanted manner. Large changes of this type are generally unsuitable for biotechnological and pharmaceutical applications. Moreover, a large number of mutations, even when they are solely on the surface, can affect the stability of the protein. In the CamSol method, we thus give as input the maximum number N of mutations that can be performed. The N positions with smaller scores are consequently chosen from the list, and candidate amino acid sequences corresponding to an extensive number of possible combination of mutations at those sites are generated. This step involves generally a very large number of candidate sequences, but such a number can be decreased by simple considerations, for example, by excluding from the list of candidates the strongly hydrophobic amino acids.

The intrinsic solubility profile predictor is then run on all the candidate sequences and the corresponding CamSol solubility scores are stored. Although the intrinsic predictor is designed to capture the solubility in the unfolded state, the sites selected for mutations are chosen on the surface of the protein using the structurally corrected profile. As a result, the intrinsic predictor accurately captures the change in solubility upon mutation at those sites.

These sequences and their CamSol solubility scores are the output of the program. The sequence with the highest score represents the most soluble variant.

Homology modeling

Three-dimensional models of the gammabody variants, including the wild type, were produced using Modeller (version 9.12) [55,56]. The crystal structure of an autonomous human VH domain (PDB code 3B9V) was used as a template for homology modeling [44]. The PDB file for this structure contains a crystal unit with the coordinates of four identical VH single-domain antibodies. The fact that the CDR3 of two of these antibodies did not crystallize completely—two residues are missing from chain D and three residues are missing from chain B—suggests that the loop is disordered. This possibility is further confirmed by the RMSD between the loop of chain A and the loop of chain C (residues 101–110), which is 6.53 Å.

The gammabody variants described in the main text have a sequence identity with this template that ranges from 91.6% (wild type) to 89.9% (QKLE) and most of the non-matching residues are contained in the CDR3 of the gammabodies. The Align2D command from Modeller was used to align query sequences to the template structure. Care was taken in tuning the gap costs of the alignment algorithm so that only one gap was opened in the region of the CDR3 loop. This aspect is particularly important, as we do not want homology-derived restraints to be imposed to that disordered region, which we rebuild using the loopmodel class.

Thirty different homology models were built for each gammabody variant. These models were ranked with a combination of two Modeller scores [55,56], the DOPE and the *molpdf* scores. The seven worst models were rejected, leaving 23 models per variant. We considered this high number of models to describe the conformations accessible to the disordered CDR3 loop, which is four amino acids longer in the wild-type gammabody than it is in the template structure.

All the selected models were energy minimized using the NAMD program [58] to remove possible steric clashes arising from the model building. After adding all hydrogen atoms, the minimization consisted of 600 steps of conjugate gradient using the default parameters.

All the models produced in this way were validated on the Swiss-Model QMEAN server [59], yielding an average QMEAN score of 0.71 with a standard deviation of 0.04.

This is particularly good given that the QMEAN score of the template crystal structure (chain A of 3B9V in the PDB) is 0.775.

Cloning and production of the different gammabody variants

Gammabody Aβ(33-42) variants were obtained by employing phosphorylated oligonucleotide PCR technique or QuikChange XLII kit (Qiagen, Venlo, Limburg, the Netherlands) on the wild-type pET17b/Aβ(33-42) cDNA, depending on the type of the mutation.

The different gammabodies were expressed in *Escherichia coli* BL21 (DE3)-pLysS strain (Stratagene, La Jolla, CA, USA) for 15 h at 30 °C using Overnight Express Instant TB Medium (Merck Millipore, Billerica, MA, USA) supplemented with ampicillin (100 μg/ml) and chloramphenicol (35 μg/ml). The cellular suspension was centrifuged twice at 6000 rcf and the supernatant was incubated with 2.5 ml/l of supernatant of Ni-NTA resin (Qiagen) at 18 °C overnight in mild agitation. The Ni-NTA beads were collected and the protein was eluted in phosphate-buffered saline (PBS) (pH 3), neutralized at pH 7 upon elution. The protein purity, as determined by NuPAGE (Life Technologies, Carlsbad, CA, USA), exceeded 95% (Fig. S2a). Protein concentrations and soluble protein yields were determined by absorbance measurements at 280 nm using theoretical extinction coefficients calculated with Expasy ProtParam.

Circular dichroism

Far-UV CD spectra for all protein variants were recorded using a Jasco J-810 spectropolarimeter equipped with a Peltier holder, using a 0.1-cm-pathlength cuvette. Typically, samples contained 10 μM protein in PBS. The far-UV CD spectra of all the variants were recorded from 200 to 250 nm at 25 °C, and the spectrum of the buffer was systematically subtracted from the spectra of all protein samples.

Dot blot assay and ELISA test

Dot blot assays were performed by applying four different amounts (0.9, 0.45, 0.28, and 0.14 μg) in 50 μl of volume of monomeric Aβ42 to a 0.1-μm nitrocellulose membrane (Merck Millipore) mounted on a vacuum manifold (Bio-Rad Laboratories, Inc., Hercules, CA, USA). Samples were vacuum-filtered and washed twice with 100 μl of PBS. Membranes were then blocked with 5% bovine serum albumin in PBS for 1 h at room temperature and then probed overnight at 4 °C with 7 μM gammabodies or anti-amyloid β, clone W0-2 monoclonal antibody (Merck Millipore) as a control in 5% bovine serum albumin in PBS. The next day, gammabodies and W0-2 probed membranes were washed three times for 15 min with 0.01% Tween in PBS and were subsequently incubated for 1 h at room temperature with Anti-His (C-term)-FITC Ab (Life Technologies) or Alexa Fluor® 488 Goat Anti-Mouse IgG (H + L) Antibody (Life Technologies), respectively. The excess of antibody was removed by washing the membranes three times for 15 min with 0.01% Tween in PBS. Immunofluorescence quantification was performed on a Typhoon Trio scanner (GE Healthcare Life Sciences, Little Chalfont, UK) and the images were analyzed with the program ImageQuant TL v2005 software (GE Healthcare Life Sciences).

ELISA tests were performed using Aβ42 Human ELISA Kit (Life Technologies). The wells of the ELISA plate coated with a monoclonal antibody to the N-terminal region of Aβ were incubated in the presence of increasing amounts of Aβ42 (from 0 to 40 μM) and 1 μM of the wild type or EEE gammabody in 100 μl volume, according to manufacturer's instructions. The amount of bound gammabody was detected using peroxidase-conjugated rat monoclonal anti-FLAG Ab (Sigma Aldrich, St. Louis, MO, USA) and provided reagents according to manufacturer's instructions. The increase in Abs_{450nm} was plotted as a function of Aβ42 and analyzed assuming single-site binding model. The fraction of bound ligand was plotted as a function of protein concentration in order to compare affinities between the different antibody variants.

Saturation concentration analysis

In order to determine the saturation concentration of the gammabody variants, we obtained protein samples at different concentrations by centrifugation steps using AmiconUltra-0.5, Ultracel-3 Membrane, 3 kDa (Merck Millipore), incubated for 30 min at room temperature and centrifuged at 90,000 rpm for 45 min at 4 °C. The concentration of the resulting supernatant was plotted as a function of the total protein concentration of the solution, which was measured before the centrifugation.

Dynamic light scattering

Representative size distributions of the gammabody variants at different protein concentrations in PBS were recorded at 25 °C on the Zetasizer Nano ZS instrument (Malvern Instruments Ltd., Malvern, UK) at 633 nm upon ultracentrifugation at 90,000 rpm. The acquired data were analyzed by Zetasizer Nano software (Malvern Instruments Ltd.).

Analytical SEC

Analytical SEC was performed at 4 °C using a Superdex 75 10/300 GL column (GE Healthcare Life Sciences), which had previously been calibrated using a mixture of standard proteins (albumin, 66,500 M_r, 35 Å; chymotrypsinogen A, 25,000 M_r, 20.9 Å; ovalbumin, 43,000 M_r, 30.5 Å; ribonuclease A, 13,700 M_r, 16.4 Å). Typically, samples of the various constructs contained 70 μM protein in PBS and were incubated for 3 h at room temperature before being loaded onto the column for analysis. The monomeric fractions as a function of total protein concentration for DED and EEE gammabody variants (Fig. S4) were estimated in a similar manner by the addition of a centrifugation step of the samples for 15 min at 16,000 rcf at 4 °C just prior the analysis in order to remove big aggregates that could have occluded the column (for protein samples up to 250 μM, no detectable protein pellet was observed). The relative quantities of the different protein species were estimated by analyzing

the chromatographic profile using a Gaussian function for each elution peak and measuring the relative area under each peak.

Native polyacrylamide gel electrophoresis

Native electrophoresis analysis was performed on a 10-μl protein fraction upon SEC by employing NativePAGE™ Bis-Tris 4–16% precast minigel system (Life Technologies), according to the manufacturer's instructions.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2014.09.026.

Acknowledgements

We are grateful to Dr. Peter Tessier for sending us the plasmid of the wild-type gammabody Aβ(33-42).

Received 29 May 2014;
Received in revised form 30 September 2014;
Accepted 30 September 2014
Available online 14 October 2014

Keywords:
protein solubility;
protein aggregation

†P.S. and F.A.A. contributed equally to this work.
‡The method is available as a Web server at http://www-mvsoftware.ch.cam.ac.uk/

Abbreviations used:
CDR, complementarity-determining region; SEC, size-exclusion chromatography; DOPE, discrete optimized protein energy; PDB, Protein Data Bank; PBS, phosphate-buffered saline.

References

[1] Pavlou AK, Reichert JM. Recombinant protein therapeutics—Success rates, market trends and values to 2010. Nat Biotechnol 2004;22:1513–9.
[2] Leader B, Baca QJ, Golan DE. Protein therapeutics: A summary and pharmacological classification. Nat Rev Drug Discov 2008;7:21–39.
[3] Goodman M. Market watch: Sales of biologics to show robust growth through to 2013. Nat Rev Drug Discov 2009;8:837.
[4] Carter PJ. Introduction to current and future protein therapeutics: A protein engineering perspective. Exp Cell Res 2011;317:1261–9.
[5] Winter G, Griffiths AD, Hawkins RE, Hoogenboom HR. Making antibodies by phage display technology. Annu Rev Immunol 1994;12:433–55.
[6] Sidhu SS. Phage display in pharmaceutical biotechnology. Curr Opin Biotechnol 2000;11:610–6.
[7] Hoogenboom HR. Selecting and screening recombinant antibody libraries. Nat Biotechnol 2005;23:1105–16.
[8] Bradbury AR, Sidhu SS, Dübel S, McCafferty J. Beyond natural antibodies: The power of *in vitro* display technologies. Nat Biotechnol 2011;29:245–54.
[9] Lee CC, Perchiacca JM, Tessier PM. Toward aggregation-resistant antibodies by design. Trends Biotechnol 2013;31:612–20.
[10] Elvin JG, Couston RG, van der Walle CF. Therapeutic antibodies: Market considerations, disease targets and bioprocessing. Int J Pharm 2013;440:83–98.
[11] Carter PJ. Potent antibody therapeutics by design. Nat Rev Immunol 2006;6:343–57.
[12] Jespers L, Schon O, Famm K, Winter G. Aggregation-resistant domain antibodies selected on phage by heat denaturation. Nat Biotechnol 2004;22:1161–5.
[13] Shire SJ. Formulation and manufacturability of biologics. Curr Opin Biotechnol 2009;20:708–14.
[14] Manning MC, Chou DK, Murphy BM, Payne RW, Katayama DS. Stability of protein pharmaceuticals: An update. Pharm Res 2010;27:544–75.
[15] Perchiacca JM, Tessier PM. Engineering aggregation-resistant antibodies. Annu Rev Chem Biomol Eng 2012;3:263–86.
[16] Balch WE, Morimoto RI, Dillin A, Kelly JW. Adapting proteostasis for disease intervention. Science 2008;319:916–9.
[17] Ciryam P, Tartaglia GG, Morimoto RI, Dobson CM, Vendruscolo M. Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins. Cell Rep 2013;5:781–90.
[18] Hartl FU, Bracher A, Hayer-Hartl M. Molecular chaperones in protein folding and proteostasis. Nature 2011;475:324–32.
[19] Dobson CM. Protein folding and misfolding. Nature 2003;426:884–90.
[20] Ventura S, Villaverde A. Protein quality in bacterial inclusion bodies. Trends Biotechnol 2006;24:179–85.
[21] Kopito RR. Aggresomes, inclusion bodies and protein aggregation. Trends Cell Biol 2000;10:524–30.
[22] Baneyx F, Mujacic M. Recombinant protein folding and misfolding in *Escherichia coli*. Nat Biotechnol 2004;22:1399–408.
[23] Davis GD, Elisee C, Newham DM, Harrison RG. New fusion protein systems designed to give soluble expression in *Escherichia coli*. Biotechnol Bioeng 1999;65:382–8.
[24] Arnau J, Lauritzen C, Petersen GE, Pedersen J. Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins. Protein Expression Purif 2006;48:1–13.
[25] Sørensen HP, Mortensen KK. Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. Microb Cell Factories 2005;4:1.
[26] Waldo GS. Genetic screens and directed evolution for protein solubility. Curr Opin Chem Biol 2003;7:33–8.
[27] Aharoni A, Gaidukov L, Yagur S, Toker L, Silman I, Tawfik DS. Directed evolution of mammalian paraoxonases PON1 and PON3 for bacterial expression and catalytic specialization. Proc Natl Acad Sci USA 2004;101:482–7.
[28] Frokjaer S, Otzen DE. Protein drug stability: A formulation challenge. Nat Rev Drug Discov 2005;4:298–306.
[29] Famm K, Hansen L, Christ D, Winter G. Thermodynamically stable aggregation-resistant antibody domains through directed evolution. J Mol Biol 2008;376:926–31.
[30] Jung S, Honegger A, Plückthun A. Selection for improved protein stability by phage display. J Mol Biol 1999;294:163–80.

[31] Ladiwala ARA, Bhattacharya M, Perchiacca JM, Cao P, Raleigh DP, Abedini A, et al. Rational design of potent domain antibody inhibitors of amyloid fibril assembly. Proc Natl Acad Sci USA 2012;109:19965–70.
[32] Perchiacca JM, Ladiwala ARA, Bhattacharya M, Tessier PM. Structure-based design of conformation- and sequence-specific antibodies against amyloid beta. Proc Natl Acad Sci USA 2012;109:84–9.
[33] Rouet R, Lowe D, Christ D. Stability engineering of the human antibody repertoire. FEBS Lett 2014;588:269–77.
[34] Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. Proc Natl Acad Sci USA 2009;106:11937–42.
[35] Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 2003;424:805–8.
[36] Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, Vendruscolo M. Prediction of aggregation-prone regions in structured proteins. J Mol Biol 2008;380:425–36.
[37] Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 2004;22:1302–6.
[38] Wilkinson DL, Harrison RG. Predicting the solubility of recombinant proteins in *Escherichia coli*. Nat Biotechnol 1991;9:443–8.
[39] Magnan CN, Randall A, Baldi P. Solpro: Accurate sequence-based prediction of protein solubility. Bioinformatics 2009;25: 2200–7.
[40] Smialowski P, Doose G, Torkler P, Kaufmann S, Frishman D. PROSO II—A new method for protein solubility prediction. FEBS J 2012;279:2192–200.
[41] Agostini F, Vendruscolo M, Tartaglia GG. Sequence-based prediction of protein solubility. J Mol Biol 2012;421:237–41.
[42] Wang X, Das TK, Singh SK, Kumar S. Potential aggregation prone regions in biotherapeutics: A survey of commercial monoclonal antibodies. MAbs 2009;1:254.
[43] Dudgeon K, Rouet R, Kokmeijer I, Schofield P, Stolp J, Langley D, et al. General strategy for the generation of human antibody variable domains with increased aggregation resistance. Proc Natl Acad Sci USA 2012;109: 10879–84.
[44] Barthelemy PA, Raab H, Appleton BA, Bond CJ, Wu P, Wiesmann C, et al. Comprehensive analysis of the factors contributing to the stability and solubility of autonomous human VH domains. J Biol Chem 2008;283:3639–54.
[45] Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. J Mol Biol 2005;350:379–92.
[46] Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. Proc Natl Acad Sci USA 2009;106:10159–64.
[47] Tsytlonok M, Sormanni P, Rowling PJ, Vendruscolo M, Itzhaki LS. Subdomain architecture and stability of a giant repeat protein. J Phys Chem B 2013;117:13029–37.
[48] Trevino SR, Scholtz JM, Pace CN. Measuring and increasing protein solubility. J Pharm Sci 2008;97:4155–66.
[49] Miklos AE, Kluwe C, Der BS, Pai S, Sircar A, Hughes RA, et al. Structure-based design of supercharged, highly thermoresistant antibodies. Chem Biol 2012;19:449–55.
[50] Tan PH, Chu V, Stray JE, Hamlin DK, Pettit D, Wilbur DS, et al. Engineering the isoelectric point of a renal cell carcinoma targeting antibody greatly enhances scFv solubility. Immunotechnology 1998;4:107–14.
[51] Pantaloni D, Carlier M, Coue M, Lal A, Brenner S, Korn E. The critical concentration of actin in the presence of ATP increases with the number concentration of filaments and approaches the critical concentration of actin.ADP. J Biol Chem 1984;259:6274–83.
[52] Aprile FA, Dhulesia A, Stengel F, Roodveldt C, Benesch JL, Tortora P, et al. Hsp70 oligomerization is mediated by an interaction between the interdomain linker and the substrate-binding domain. PLoS One 2013;8:e67961.
[53] Huertas ML, Carrasco B. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. Biophys J 2000;78:719–30.
[54] Arosio P, Barolo G, Müller-Späth T, Wu H, Morbidelli M. Aggregation stability of a monoclonal antibody during downstream processing. Pharm Res 2011;28:1884–94.
[55] Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234: 779–815.
[56] Fiser A, Sali A. Modeller: Generation and refinement of homology-based protein structure models. Methods Enzymol 2003;374:461–91.
[57] Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 1996;3:842–8.
[58] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem 2005;26:1781–802.
[59] Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res 2009;37: W510–4.