Briefings in Bioinformatics, 2023, **24(2)**, 1–7
https://doi.org/10.1093/bib/bbad004
Advance access publication date 31 January 2023
Problem Solving Protocol

Sequence-based prediction of pH-dependent protein solubility using CamSol

Marc Oeller, Ryan Kang, Rosie Bell, Hannes Ausserwöger, Pietro Sormanni and Michele Vendruscolo

Corresponding authors. Michele Vendruscolo, E-mail: mv245@cam.ac.uk; Pietro Sormanni, E-mail: ps589@cam.ac.uk

Abstract

Solubility is a property of central importance for the use of proteins in research in molecular and cell biology and in applications in biotechnology and medicine. Since experimental methods for measuring protein solubility are material intensive and time consuming, computational methods have recently emerged to enable the rapid and inexpensive screening of solubility for large libraries of proteins, as it is routinely required in development pipelines. Here, we describe the development of one such method to include in the predictions the effect of the pH on solubility. We illustrate the resulting pH-dependent predictions on a variety of antibodies and other proteins to demonstrate that these predictions achieve an accuracy comparable with that of experimental methods. We make this method publicly available at https://www-cohsoftware.ch.cam.ac.uk/index.php/camsolph, as the version 3.0 of CamSol.

Keywords: protein solubility, pH dependency, developability, drug formulation, solubility prediction

Introduction

Solubility is one of the key properties that underpins the developability potential of proteins in industrial pipelines [1–9]. Other such properties include expression yield, immunogenicity, chemical and conformational stability, viscosity and polyspecificity [1–9]. Although proteins have evolved to be soluble enough to be functional in the cellular environment [4, 10, 11], proteins for research, diagnostic and especially therapeutic purposes are commonly required to withstand the high concentrations necessary for storage and for certain administration routes, such as subcutaneous delivery. This consideration implies that in most cases protein solubility must be optimized beyond typical natural levels, and that specific formulation conditions, including the pH, must be identified to maximize solubility and stability of the product.

The solubility of proteins is defined thermodynamically in terms of the critical concentration, which is the level of concentration where the soluble and insoluble phases are in equilibrium [2, 9]. The solubility is therefore dependent on the formulation conditions. Therefore, formulation optimization is a key step in protein development pipelines, and it is important in particular to find the most suitable pH value to ensure that a protein is sufficiently stable.

Although several methods have been developed for the experimental measurement of protein solubility [2, 5], these methods are not readily amenable to high-throughput screening campaigns, which are required to assess the large number of candidates typically available the early stages of industrial pipelines. For this reason, many computational prediction methods have been developed in recent years. PON-sol [12], SOLpro [13] and PROSO II [14] use machine learning techniques to predict solubility in terms of soluble expression yield. Other methods derive the solubility from aggregation-prone regions [15] calculated using physicochemical descriptors of amino acid sequences, including TANGO [16], Aggrescan [17], Solubis [18] and the original CamSol method [19]. The use of molecular dynamics simulations to predict the exposure of hydrophobic regions and its link with aggregation propensity, such as in the case of the SAP method [20], has also been exploited. Despite many of these methods being highly reliable, there is still an unmet need for sequence-based predictors capable of accurately assessing the effects of formulation pH on the solubility of proteins.

In this work we generalize the CamSol method [19], which was introduced to predict the solubility of protein variants, to predict the effects of varying the pH on protein solubility. Our approach encompasses three main features: (i) the calculation of partial charges using the Henderson–Hasselbalch equation, (ii) the calculation of hydrophobicity values with pH-dependent logD values and (iii) the calculation of the context-dependent residue pK_a values, either from the 3D structure when available [21, 22] or through a sequence-based prediction ([23]; Figure 1). By employing CamSol 3.0, we show that we can accurately predict the solubility behavior at different pH values of proteins with varying sizes, including nanobodies, full-length antibodies and intrinsically disordered proteins.

Marc Oeller is a PhD student in Chemistry at the University of Cambridge. He is working on the physico-chemical principles of protein solubility.
Ryan Kang is a PhD student in Chemistry at the University of Cambridge. He did his master's project under Dr. Sormanni investigating protein solubility.
Rosie Bell is a postdoctoral research scientist at Birkbeck College, University of London. She did her PhD in Michele Vendruscolo's lab working on the biophysics of protein-protein interactions, and protein stability.
Hannes Ausserwöger is a PhD student in Chemistry at the University of Cambridge. His research focuses on studying the specificity of protein interactions.
Pietro Sormanni is a group leader supported by a University Research Fellowship from the Royal Society. His research focuses on the development of innovative technologies of protein engineering and antibody design at the interface between computation and experiments.
Michele Vendruscolo is Professor of Biophysics and Co-Director of the Centre for Misfolding Diseases at the University of Cambridge. His research concerns the principles that determine protein aggregation, solubility and homeostasis.
Received: August 31, 2022. **Revised:** December 10, 2022. **Accepted:** October 16, 2022

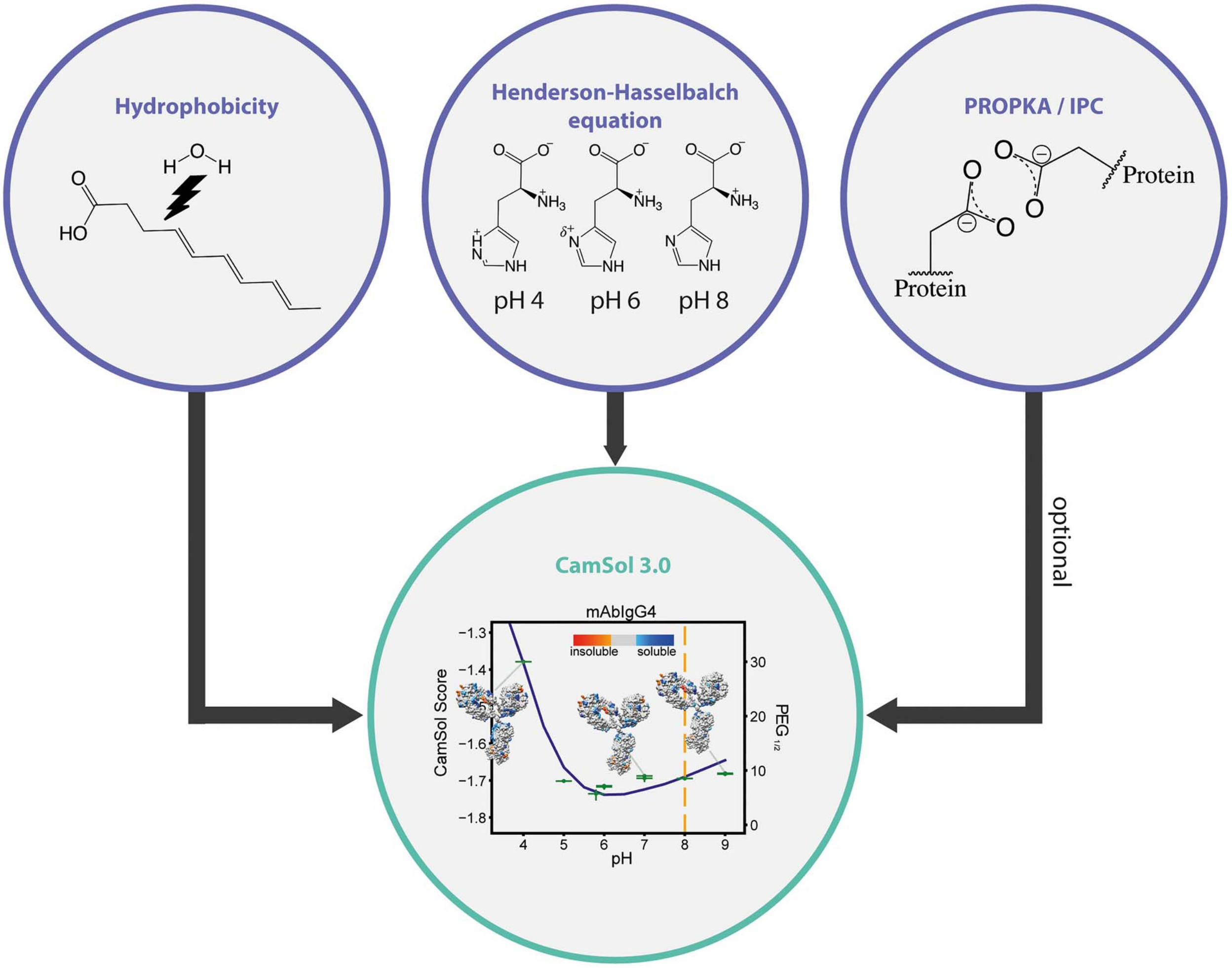

Figure 1. Schematic illustration of the sequence-based pH-dependent solubility predictions of CamSol. CamSol assesses partial charges using the Henderson–Hasselbalch equation. Hydrophobicity calculations are replaced by LogD calculations [26]. If a structure is supplied, amino acids pK_a values are calculated using PROPKA, otherwise the IPC method is used. Experimental data (green markers in lower circle) were generated using a recently developed PEG Assay.

Results

The CamSol method calculates the solubility of proteins based on the physicochemical properties of their amino acid sequences [19]. Changes in pH mainly affect ionisable residues, as the pH determines the protonation state and therefore the electrostatic charges of these residues. To accurately assess the charge of each residue, we implemented the Henderson–Hasselbalch equation [24] to determine the ratio between protonated and charged residues to estimate the partial charge of each amino acid.

An important component of these calculations is an accurate pK_a value, which determines the pH range where the residue is charged. In this work, we employed three alternative strategies to obtain accurate pK_a values. The user may choose which approach to use, depending on the information available on the protein under scrutiny.

In the first method, we revisited the tabulated pK_a values used in the original version of CamSol and updated them with more accurate values (Table S1). More specifically, we retrieved pK_a values from the computational biophysics and bioinformatics database (http://compbio.clemson.edu/pkad), which contains over 1500 experimentally measured residue pK_a values from a wide range of proteins. We then employed the median pK_a value observed in this database for each of the 20 amino acid types.

In the second method, sequence-based predictions of pK_a values using IPC 2.0 were applied to all proteins with a specific focus on disordered proteins such as α-synuclein and peptides with high structural heterogeneity such as insulin. IPC is especially useful for peptides and proteins that are intrinsically disordered or highly dynamic. Using pK_a values updated in this way helps refine the solubility predictions and improves accuracy while keeping the method completely sequence-based.

In the third method, we started from the observation that amino acid pK_a values are affected by the structural environment and neighboring residues. To take these effects into account we incorporated pK_a predictions using the PROPKA method [21, 22]. PROPKA is a widely used pK_a predictor for proteins [25], which requires the knowledge of the structure of the input protein or of an accurate structural model. For a comparison of how the pK_a values change upon usage of the different methods, Table S2 shows the pK_a values for each method for bovine serum albumin (BSA).

Furthermore, we also ensured that the effects of changes in pH are accurately reflected in the way hydrophobicity affects the solubility predictions. In the original CamSol method, hydrophobicity has been expressed as the partition coefficient logP. To take charged species into account, we replaced these logP values with their corresponding logD values. Whereas logP values only

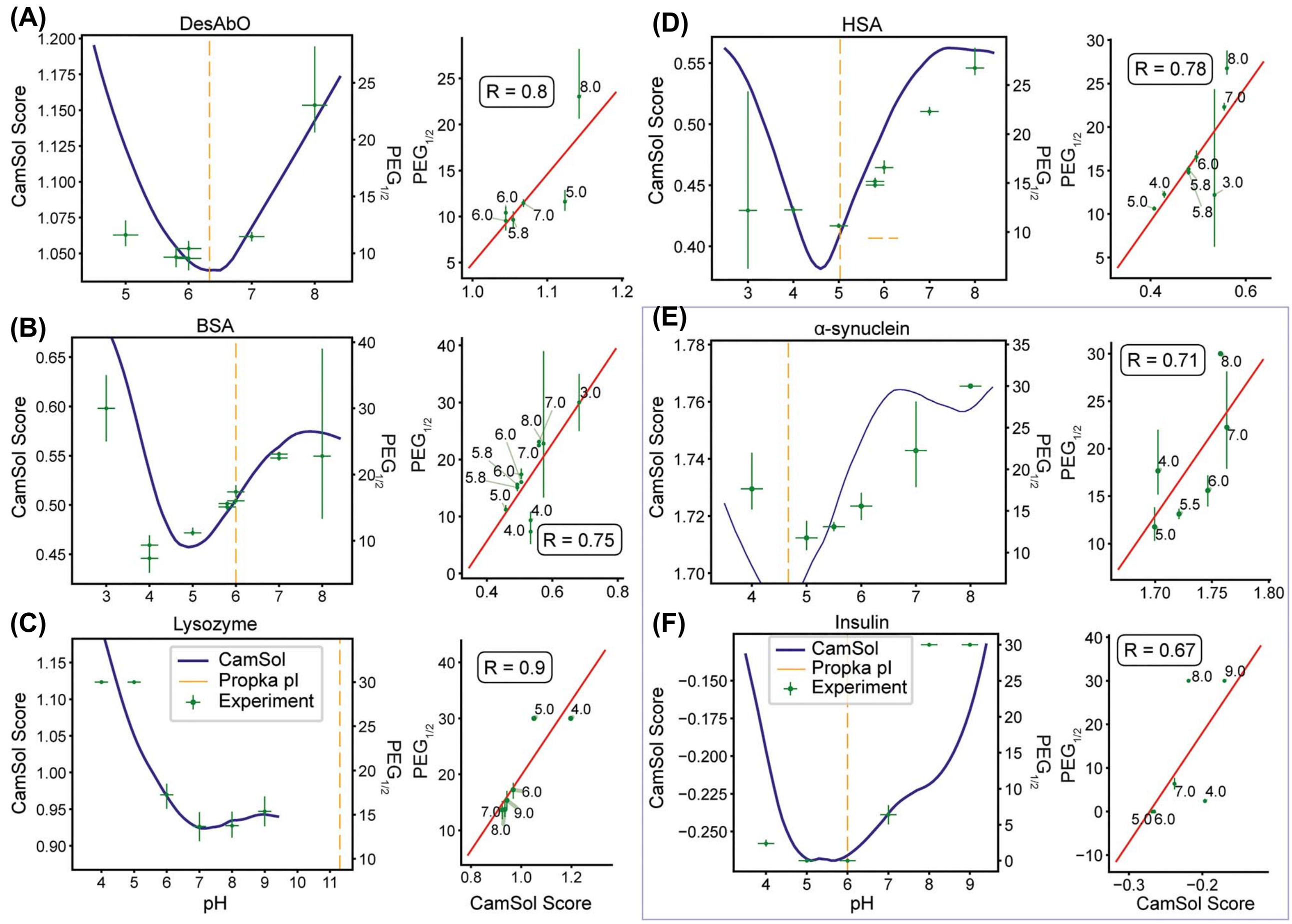

Figure 2. CamSol predicts solubility values that are highly correlated with experimental solubility values. Plots on the left-hand side in each column visualize how experimental and predicted values change over a range of pH values. The left axis and blue line report the predicted CamSol solubility score, the right axis and green markers the measured midpoints of PEG-precipitation, all as a function of pH (x-axis). The vertical yellow line is the theoretical isoelectric point. Plots on the right-hand side shows the correlation between the predicted and measured relative solubility values. CamSol calculations were carried out using pK_a values calculated by PROPKA for (A) DesAbO (nanobody), (B) bovine serum albumin (BSA), (C) hen egg white lysozyme and (D) human serum albumin (HSA), whereas for (E) α-synuclein and (F) insulin pK_a values were calculated with IPC (framed in blue box). *R* is the Pearson's coefficient of correlation.

consider the neutral state of a species, logD values take the neutral and charged states into account. We used data published by Zamora *et al.* [26] to predict LogD values for all ionisable residues. Since logD values are also pK_a-dependent, the calculated logD values can be updated with IPC or with PROPKA if a structure is available. We normalized and linearly fitted calculated logD values to the range of logP values previously used in CamSol to maintain consistency in the range of hydrophobicity values and to avoid the necessity to refit the scoring parameters of CamSol.

Since the availability of high-quality solubility data at varying pH values is limited, we set out to carry out experimental measurements of relative solubility. Specifically, we test our method on proteins that were either commercially available or already produced in our laboratory. A further criterion was that such proteins should not contain large co-factors (e.g. bound heme or large heteroatom groups) or metals, as these components can significantly alter their solubility and are not yet accounted for by the CamSol method. We measured the solubilities of α-synuclein, insulin, lysozyme, the single-domain antibody DesAbO [27], human serum albumin (HSA) and BSA at pH values ranging from 3 to 9 (Figure S1, see Supplementary Data available online), as well as that of the full-length IgG4 monoclonal antibody HzATNP and one of its mutational variants [3]. The range of pH values was limited by the experimental behavior of the proteins. At extreme pH values (3–4 and 8–9) the measurements become more inaccurate resulting in data with relatively large errors. However, these extreme pH values are not generally accessible for biocompatible formulation. To ensure that the solubility measurements are not affected by large conformational changes (e.g. unfolding) of the protein under scrutiny at certain pH values, we measured the circular dichroism (CD) spectra of every sample after the incubation time employed in the solubility assay (Figure S2, see Supplementary Data available online).

We calculated the correlation between measured relative solubility values and the predicted CamSol scores, testing also the different ways of calculating residue pK_a values implemented in the method (Figure 2). The results with IPC and without any pK_a corrections are shown in Figures S3 and S4 (see Supplementary Data available online), respectively. Our results show that CamSol not only captures the overall trend in solubility upon pH changes well, but the predicted values are also highly correlated with the experimental values with Pearson's coefficients of correlation always greater than 0.67 (Tables S3 and S4). Figure 2A–D depict predictions including PROPKA for folded, globular proteins, whereas Figure 2E and F (blue box) show the results for α-synuclein (an intrinsically disordered protein) and the peptide insulin, for which IPC was applied. With the proteins tested experimentally, we aimed at covering a broad range

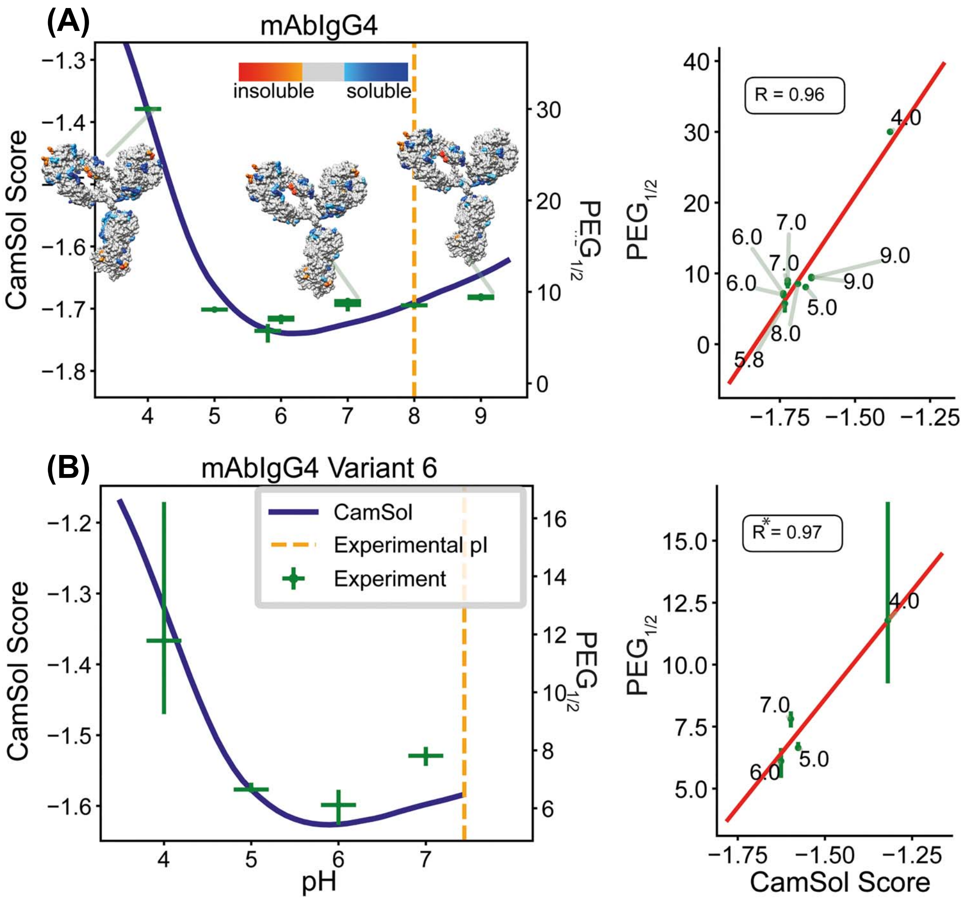

Figure 3. Change in solubility at varying pH for a monoclonal IgG4 antibody. As in Figure 2, plots on the left visualize how experimental and predicted values change over a range of pH values, and those on the right show the correlation between the predicted and measured relative solubility values. (A) The modelled structures of the IgG4 antibody are color-coded according to the CamSol structurally corrected profile for pH 4, 7 and 9 (blue: highly soluble region, red: highly insoluble region, see color-bar). (B) Same plot for a mutational variant of mAbIgG4 harboring two mutations that lower the pI (S70E V99D on heavy and light chain respectively). *R* is the Pearson's coefficient of correlation (*R*∗ indicates that *R* is estimated from the only four points that are available). Green crosses: experimental values; blue lines: predicted values; yellow dotted line: calculated pI and green dotted line: experimental pI.

(4.5–11) of theoretical pI values (Figure 2, yellow vertical line), as the pI is usually a good indicator for the pH value corresponding to the minimum in solubility.

BSA and HSA (Figure 2B and D) showed a good agreement between experimental values and predicted results with a Pearson's coefficient of correlation of 0.75 and 0.78, respectively. These correlations improved to 0.92 and 0.88 with the use of IPC, and are still relatively good even if the pK_a values are not corrected (0.63 and 0.78). The nanobody DesAbO shows an even better agreement with a coefficient of correlation of 0.8. However, the case of the nanobody indicates the limits of the IPC approach, as the coefficient of correlation drops to 0.47 (0.39 without any correction). Lysozyme shows almost a perfect correlation, with a correlation coefficient of 0.9. For α-synuclein and insulin only the IPC method was applied since intrinsically disordered Proteins (IDPs) and peptides do not form stable structures that can be used for PROPKA. The correlations are slightly lower, with a coefficient of correlation of 0.71 and 0.67, respectively (0.72 and −0.67 without any correction).

When we analyzed DesAbO, we realized that the charge effect of the C-terminal His-tag (i.e. the sequence of seven histidine residues used for purification) could be estimated more accurately. The His-tag was not part of the structure and the pK_a values were therefore not adjusted by using PROPKA. Nevertheless, it is clear from a physicochemical standpoint that the pK_a values must change if so many ionizable residues are near each other. By assuming the pK_a value shifts from 6.5 to a lower value [28], we carried out the calculations at pK_a 6.

We then tested the method on a IgG4 antibody (mAbIgG4). The coefficient of correlation is 0.96 with PROPKA pK_a values (Figure 3A), 0.88 with IPC (Figure S5, see Supplementary Data available online) and 0.85 with no correction (Figure S6, see Supplementary Data available online). We also visualized the change in solubility by color-coding the solubility profile onto the structure of the antibody (blue: highly soluble region and red: highly insoluble region). This approach highlights how some regions become less soluble at higher pH values and vice versa. We also tested a mutational variant of the antibody with a slightly lower pI to see whether CamSol was capable of capturing even small changes in pH-dependent solubility induced by minor mutations. CamSol can predict well the solubility of this variant with a coefficient of correlation of 0.97 (0.86 and 0.82, respectively for IPC and no pK_a correction; Figure 3B). To illustrate how the change in pH affects amino acids and sequence regions differently we plotted the solubility profile for BSA at three different pH values (Figure S7, see Supplementary Data available online). Looking at these solubility profiles helps pinpoint specific sequence regions whose contribution to solubility is most affected by changes in pH. Therefore, this analysis can aid the engineering of mutations to alter the pH-dependence of the solubility.

Discussion and conclusions

Predicting the solubility of proteins as a function of the pH is important in industrial pipelines as it can reduce the number and cost of labor-intensive *in vitro* assays to determine optimal formulations [1, 3–5]. With the new version of CamSol that we have reported here we took a step into this direction by including the effect of the pH in the predictions.

We have shown that CamSol 3.0 can reliably predict the solubility of globular proteins with a wide range of pI values (from 4.5 for α-synuclein up to 11 for lysozyme). CamSol 3.0 is also capable of capturing the changes induced by small mutations that shift the overall pI value of the protein. Using CamSol 3.0 we also visualized the changes of solubility on the surface of an IgG4 antibody and demonstrated how certain areas of a protein can move from highly soluble to neutral and neutral areas to highly insoluble over a short range of pH values.

The pI value is usually used as an approximation of the pH value at which a protein is least soluble as this is the point at which the overall charge on a protein is neutral. Although in many cases this assumption is rather accurate, in general the solubility of folded proteins does not only depend on the overall charge. The distribution of these charges is crucial, as patches of opposing charges can act as starting points for aggregation due to strong attractive intermolecular forces. Although the changes in hydrophobicity are related to the changes in charge upon pH variations, these are not perfectly correlated and hence can move the point of least solubility away from the pI value as well. CamSol 3.0 tries to capture all these effects to give a more accurate estimate of the point of lowest solubility. Moreover, we also predicted the behavior around the pI value that is more informative than just the point of least solubility.

We have provided the user with three different options to calculate amino acid pK_a values: (i) Using tabulated pK_a values, (ii) Using PROPKA if a structure is available and the protein of interest is stably folded and (iii) using IPC, if the protein under scrutiny is structurally heterogeneous or fully disordered. Option 2 is the most accurate for structured proteins, whereas option 3 works best for highly dynamic ones. Option 1 is provided as its calculations are extremely fast, and it can therefore be employed for very large-scale screenings of pH-dependence using little computational resources. The pH-dependence of the solubility of the top-ranking proteins from such screenings can then be calculated more accurately by employing either PROPKA or IPC calculated residue pK_a values.

By testing a wide variety of different proteins, from small, disordered peptides and proteins to large globular proteins including full-length antibodies, we illustrated the general applicability of our method. Nevertheless, we acknowledge that the new version of CamSol is still limited to proteins that do not contain large cofactors such as heme, as we expect that these can alter the pH-dependent solubility significantly and CamSol cannot currently account for these aspects. Moreover, other large modifications such as glycosylation, lipidation or the presence of co-factors are not yet taken into account in our method. In its current implementation, CamSol is aimed at performing relative comparisons of the solubility of a protein at varying conditions (including pH) or of similar proteins such as mutational variants.

In addition to aid the development of protein biologics, we expect our method to be useful in protein engineering and *de novo* protein design, adding to the increasingly powerful arsenal of computational methods emerging in these fields.

In conclusion, with version 3.0 of CamSol we have presented a sequence-based method that can accurately predict the solubility of proteins at varying pH values.

Methods

Theoretical methods

Here, we provide an overview of the CamSol method and explain the changes introduced to take into account the effect of the pH on protein solubility. For a detailed explanation of the CamSol approach, the reader is referred to the original CamSol paper [19]. CamSol is based on a phenomenological combination of physicochemical properties, and therefore its results are readily interpretable in terms of these properties. The software was based on the Zyggregator method [29], which predicts amyloid aggregation propensity of proteins. In CamSol, four physicochemical properties—charge, hydrophobicity, α-helical propensity and β-sheet propensity—are combined to assign a score to each amino acid, yielding a solubility profile which is then smoothed to account for the effect of neighboring residues, and then corrected for hydrophobic-hydrophilic patterns and gatekeeper effects (gatekeepers are charged residues flanking hydrophobic regions and modulate their effects on solubility). From this profile, an overall solubility score is calculated which was updated in version 2 of CamSol [30].

In the original CamSol method, it was already possible to provide the value of the pH as input. The consequence of changing the input pH was to adopt specific side-chain charges depending on tabulated pK_a values for the twenty standard amino acid. For example, all histidine residues would acquire a charge of +1 for an input pH below 6.5. Although rooted on general physicochemical principles, this description of the pH-dependence of residue pK_a is very coarse, and can be substantially improved. Therefore, in the current work, we updated the pK_a values tabulated in CamSol by compiling all experimentally determined pK_a values from http://compbio.clemson.edu/pkad (SI). We also introduced charges for the amide group at the N-terminus and the carboxylic acid at the C-terminus. The charge is calculated by using the Henderson–Hasselbalch equation, and partial charges are now allowed

$$\mathrm{pH} = \mathrm{p}Ka + \log\frac{\mathrm{Base}}{\mathrm{Acid}} \Longrightarrow \frac{\mathrm{Base}}{\mathrm{Acid}} = 10^{\mathbf{pH}-\mathbf{p}Ka}$$

Therefore, CamSol 3.0 not only relies on more accurate pK_a values (either from the updated table, or calculated with PROPKA or IPC), but employs partial charges when the pH is close to the pK_a of a charged amino acid.

Using the ratio of charged to neutral species calculated with the above equation, we also replaced the logP values representing hydrophobicity by pH-dependent hydrophobicity values (logD). LogD combines the partition coefficient logP of neutral and ionized species

$$\log D_{\mathrm{pH}} = \log\left(P_N + P_I * 10^{\delta}\right) - \log\left(1 + 10^{\delta}\right)$$

where δ is the difference between pK_a and pH (pK_a—pH for basic residues and pH—pK_a for acidic residues). We used the pH-dependent LogD calculations by Zamora and colleagues [26] for neutral and ionized LogP values for all standard amino acids.

In the original CamSol method cysteine residues were assumed to be reduced. We changed the default to assume that all cysteine residues are in salt bridges and therefore cannot be charged. Free cysteine residues can be assigned by replacing the letter ‘C’ with the letter ‘B’ in the input sequence.

To calculate accurate pK_a values, we employed PROPKA, an open-source available pK_a predictor [21, 22]. When given a structure, PROPKA calculates pK_a values for all ionisable residues based on their surrounding residues. PROPKA uses a combination of thermodynamic calculations and empirical shifts. It assesses the effect of surrounding residues on each ionisable residue by calculating desolvation contributions and charge–charge interactions. If a structure or a suitable 3D-model is not available, as it is the case for disordered proteins and peptides, a sequence-based prediction is carried out instead using IPC 2.0. This method uses a mixture of support vector regression and deep learning models trained on the PKAD database [31] to predict pK_a values based on the protein sequences [23]. Lastly, the pK_a for histidine residues that are part of a terminal His-tag is lowered to 6 from 6.5 since the proximity of these histidine residues to each other causes a lowering of their pK_a values to avoid electrostatic repulsion [28].

Taken together, these changes enable the accurate prediction of the sequence-based pH-dependence of the solubility for a broad range of proteins.

Experimental methods

Buffer

Ten-millimolar sodium phosphate dibasic heptahydrate (MP Biomedicals, 191 441) and 10-mM citric acid monohydrate (Fisher Scientific, 5949-29-1) were combined. For each experiment the pH was adjusted using NaOH or HCl.

Proteins

BSA (Sigma, A9418), HSA (Sigma, A3782) and chicken egg white lysozyme (Sigma, L6876) were resuspended in buffer and then further purified and buffer-exchanged by carrying out size exclusion chromatography (SEC; Cytiva, Superdex75 10/300 Increase for lysozyme and Superdex200 10/300 Increase for BSA and HSA). mAbIgG4 and its variant were provided by Novo Nordisk. DesAbO and α-synuclein were produced and purified in our group as described previously [27, 32]. Each protein was freshly buffer-exchanged into the correct buffer (Cytiva, HiTrap Desalting column) before each assay.

Solubility assay

The relative solubility of proteins was measured using a recently developed polyethylene glycol (PEG) solubility assay [33]. In brief, PEG 6000 is used as a crowding agent and titrated to final PEG concentrations of 0–33% and final protein concentration of 1 mg/ml, starting from PEG and protein stocks in the same buffer at the same pH, which typically needed to be re-adjusted after dissolving the PEG. After a 48-h incubation period, plates are centrifuged, the supernatant is transferred into a fresh plate and the soluble fraction is measured using a plate reader. A sigmoidal behavior is seen for the precipitation of proteins, and the inflection point of the curve is used as a solubility proxy.

Circular dichroism spectroscopy

Proteins were diluted to a concentration of 0.1 mg/ml. Three spectra were obtained using an Applied Photophysics Chirascan and a High Precision Cell made of quartz (Hellma Analytics, path length 1 mm) between 200 and 250 nm with a bandwidth of 1 nm, step size of 0.5 nm and scanning speed of 1 s/point.

Spectrophotometry

Absorbance was measured with a plate reader (BMG Clariostar) and the spectrum from 220 to 700 nm was recorded at 25°C.

Key Points

- We present a method of predicting the pH-dependence of the solubility of proteins.
- The method is incorporated in the CamSol software and, if the structure is not known, works only using the sequence.
- The quality of the prediction is validated through accurate solubility measurements.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

We thank Prof. Zamora for help and additional data provided to calculate LogD values. We also would like to thank Dr Kozlowski for help implementing IPC 2.0. We are grateful to collaborators at Novo Nordisk (Dr Nikolai Lorenzen, Dr Marie Pedersen and Dr Daniele Granata) for providing the purified antibody HzATNP and its variant. M.O. is a PhD student funded by AstraZeneca. P.S. is a Royal Society University Research Fellow (URF\R1\201461). The project was supported by the Wellcome Trust (203249/Z/16/Z).

Authors’ contributions

M.O. performed experiments and carried out data analysis. M.O., R.K. and R.B. purified alpha synuclein and M.O. and R.K. measured its solubility. H.A. provided material. M.O. and P.S. wrote the software. M.O., P.S. and M.V.led the drafting of the manuscript. M.O., P.S. and M.V. conceived and P.S. and M.V. supervised the project.

Supporting Information

Figures and Tables as referenced in the text including Table S4, which contains the complete list of measured solubility for all proteins tested.

Abbreviations

BSA, bovine serum albumin; CD, Circular dichroism spectroscopy; HSA, human serum albumin; PEG, polyethylene glycol; SEC, size exclusion chromatography

References

1. Norman RA, Ambrosetti F, Bonvin AMJJ, *et al.* Computational approaches to therapeutic antibody design: established methods and emerging trends. *Brief Bioinform* 2019;**00**:1–19.
2. Wolf Pérez AM, Lorenzen N, Vendruscolo M, *et al.* Assessment of therapeutic antibody developability by combinations of in vitro and in silico methods. *Methods Mol Biol* 2022;**2313**:57–113.
3. Wolf Pérez AM, Sormanni P, Andersen JS, *et al.* In vitro and in silico assessment of the developability of a designed monoclonal antibody library. *MAbs* 2019;**11**:388–400.
4. Sormanni P, Vendruscolo M. Protein solubility predictions using the CamSol method in the study of protein homeostasis. *Cold Spring Harb Perspect Biol* 2019;**11**:1–12. https://doi.org/10.1101/cshperspect.a033845.

5. Jain T, Sun T, Durand S, *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc Natl Acad Sci U S A* 2017;**114**:944–9.
6. Raybould MIJ, Marks C, Krawczyk K, *et al.* Five computational developability guidelines for therapeutic antibody profiling. *Proc Natl Acad Sci* 2019;**116**:4025–30.
7. Jarasch A, Koll H, Regula JT, *et al.* Developability assessment during the selection of novel therapeutic antibodies. *J Pharm Sci* 2015;**104**:1885–98.
8. Lauer TM, Agrawal NJ, Chennamsetty N, *et al.* Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. *J Pharm Sci* 2012;**101**:102–15.
9. Garidel P. Protein solubility from a biochemical , physicochemical and colloidal perspective. *Am Pharm Rev* 2013;1–12.
10. Vecchi G, Sormanni P, Mannini B, *et al.* Proteome-wide observation of the phenomenon of life on the edge of solubility. *Proc Natl Acad Sci* 2020;**117**:1015–20.
11. Tartaglia GG, Pechmann S, Dobson CM, *et al.* Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem Sci* 2007;**32**:204–6.
12. Yang Y, Niroula A, Shen B, *et al.* PON-Sol: prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics* 2016;**32**:2032–4.
13. Magnan CN, Randall A, Baldi P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* 2009;**25**: 2200–7.
14. Smialowski P, Doose G, Torkler P, *et al.* PROSO II-A new method for protein solubility prediction. *FEBS J* 2012;**279**:2192–200.
15. Trainor K, Broom A, Meiering EM. Exploring the relationships between protein sequence, structure and solubility. *Curr Opin Struct Biol* 2017;**42**:136–46.
16. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, *et al.* Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 2004;**22**: 1302–6.
17. Kuriata A, Iglesias V, Pujols J, *et al.* Aggrescan 3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res* 2019;**8220211**:1–8.
18. Ganesan A, Siekierska A, Beerten J, *et al.* Structural hot spots for the solubility of globular proteins. *Nat Commun* 2016;**7**:1–15.
19. Sormanni P, Aprile FA, Vendruscolo M. The CamSol method of rational design of protein mutants with enhanced solubility. *J Mol Biol* 2015;**427**:478–90.
20. Chennamsetty N, Voynov V, Kayser V, *et al.* Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci* 2009;**106**: 11937–42.
21. Søndergaard CR, Olsson MHM, Rostkowski M, *et al.* Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p Kavalues. *J Chem Theory Comput* 2011;**7**:2284–95.
22. Olsson MHM, Søndergaard CR, Rostkowski M, *et al.* PROPKA3: consistent treatment of internal and surface residues in empirical pka predictions. *J Chem Theory Comput* 2011;**7**:525–37.
23. Kozlowski LP. IPC 2.0: prediction of isoelectric point and pKa dissociation constants. *Nucleic Acids Res* 2021;**49**:W285–92.
24. Po HN, Snozan NM. The Henderson-Hasselbalch equation: its history and limitations. *J Chem Educ* 2001;**78**:1499–503.
25. Davies MN, Toseland CP, Moss DS, *et al.* Benchmarking pKa prediction. *BMC Biochem* 2006;**7**:1–12.
26. Zamora WJ, Campanera JM, Luque FJ. Development of a structure-based, pH-dependent Lipophilicity scale of amino acids from continuum solvation calculations. *J Phys Chem Lett* 2019;**10**:883–9.
27. Aprile FA, Sormanni P, Podpolny M, *et al.* Rational design of a conformation-specific antibody for the quantification of Aβ oligomers. *Proc Natl Acad Sci* 2020;**117**:13509–18. https://doi.org/10.1073/pnas.1919464117.
28. Qiagen. *The QIA Expressionist*. Germany: Qiagen GmbH, Düsseldorf, 2003.
29. Tartaglia GG, Pawar AP, Campioni S, *et al.* Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* 2008;**380**:425–36.
30. Sormanni P, Amery L, Ekizoglou S, *et al.* Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Sci Rep* 2017;**7**:8200.
31. Pahari S, Sun L, Alexov E. PKAD: a database of experimentally measured pKa values of ionizable groups in proteins. *Database* 2019;**2019**:1–7.
32. Staats R, Michaels T, Flagmeier P, *et al.* Screening of small molecules using the inhibition of oligomer formation in α-synuclein aggregation as a selection parameter. *Commun Chem* 2020;**3**:1–9.
33. Oeller M, Sormanni P, Vendruscolo M. An open-source automated PEG precipitation assay to measure the relative solubility of proteins with low material requirement. *Sci Rep* 2021;**11**: 1–10.