Protein solubility is controlled by global structural flexibility

Bikash K. Bhandari[1], Paul P. Gardner[1,2], Chun Shen Lim[1,*]

[1]Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand
[2]Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand

*Corresponding author. Email: chunshen.lim@otago.ac.nz

ABSTRACT

Summary

Recombinant protein production is a widely used technique in the biotechnology industry and biomedical research, yet only a quarter of target proteins are soluble and can be purified. Failures are largely due to low protein expression and solubility. We have discovered that global structural flexibility, which can be modeled by normalised B-factors, accurately predicts the solubility of 12,216 recombinant proteins expressed in *Escherichia coli*. We have optimised B-factors, and derived a new set of values for solubility scoring that further improves the prediction accuracy. We call this new predictor the ‘Solubility-Weighted Index’ (SWI). Importantly, SWI outperforms many existing protein solubility prediction tools. We have developed ‘SoDoPE’ (Soluble Domain for Protein Expression), a web interface that allows users to choose a protein region of interest for predicting and maximising both protein expression and solubility.

Availability

The SoDoPE web server and source code are freely available at https://tisigner.com/sodope and https://github.com/Gardner-BinfLab/TIsigner, respectively.
The code and data for reproducing our analysis can be found at https://github.com/Gardner-BinfLab/SoDoPE_paper_2019.

INTRODUCTION

High levels of protein expression and solubility are two major requirements of successful recombinant protein production (Esposito and Chatterjee 2006). However, recombinant protein production is a challenging process because almost half of the proteins fail to be expressed and half of the successfully expressed proteins are insoluble (http://targetdb.rcsb.org/metrics/). These failures hamper protein research, with particular implications for structural, functional and pharmaceutical studies, that require soluble and concentrated protein samples (Kramer *et al.* 2012, Hou *et al.* 2018). Therefore, predicting solubility, and engineering protein sequences for enhanced solubility is an active area of research. Notable protein engineering approaches include mutagenesis, truncation (i.e., expression of partial protein sequences), or fusion with a solubility-enhancing tag (Waldo

2003, Esposito and Chatterjee 2006, Trevino *et al.* 2007, Chan *et al.* 2010, Kramer *et al.* 2012, Costa *et al.* 2014).

Protein solubility depends on extrinsic factors such as ionic strength, temperature and pH, as well as intrinsic factors—the physicochemical properties of the protein sequence and structure—molecular weight, amino acid composition, hydrophobicity, aromaticity, isoelectric point, structural propensities and the polarity of surface residues (Wilkinson and Harrison 1991, Chiti *et al.* 2003, Tartaglia *et al.* 2004, Diaz *et al.* 2010). Many solubility prediction tools have been developed around these features, ranging from the use of simple statistical models (e.g., linear and logistic regressions) to sophisticated machine learning models (e.g., support vector machines and neural networks) (Hirose and Noguchi 2013, Habibi *et al.* 2014, Hebditch *et al.* 2017, Sormanni *et al.* 2017, Heckmann *et al.* 2018, Wu *et al.* 2019, Yang *et al.* 2019).

In this study, we investigated the experimental outcomes of 12,216 recombinant proteins expressed in *Escherichia coli* from the ‘Protein Structure Initiative:Biology’ (PSI:Biology) (Chen *et al.* 2004, Acton *et al.* 2005). We showed that protein structural flexibility is more accurate than other protein sequence properties in predicting solubility (Vihinen *et al.* 1994, Craveur *et al.* 2015). Flexibility is a standard feature that has previously been overlooked in solubility prediction. On this basis, we derived a set of 20 values for the standard amino acid residues and used them to predict solubility. We call this new predictor the ‘Solubility-Weighted Index’ (SWI). SWI is a powerful predictor of solubility, and a good proxy for global structural flexibility. In addition, SWI outperforms many protein solubility prediction tools.

RESULTS

Global structural flexibility performs well at predicting protein solubility

To determine which protein sequence properties can accurately predict protein solubility, we examined the experimental outcomes of 12,216 recombinant proteins expressed in *E. coli* (the PSI:Biology dataset; see Supplementary Table S1A) (Chen *et al.* 2004, Acton *et al.* 2005). These proteins were expressed either with a C-terminal or N-terminal 6xHis fusion tag (pET21_NESG and pET15_NESG expression vectors, N=8,780 and 3,436, respectively). They were previously curated and labeled as ‘Protein_Soluble’ or ‘Tested_Not_Soluble’ (Seiler *et al.* 2014), based on the soluble analysis of cell lysate using SDS-PAGE (Xiao *et al.* 2010). A total of 8,238 recombinant proteins were found to be soluble, in which 6,432 of them belong to the pET21_NESG dataset.

We first computed the standard protein sequence properties, namely molecular weight, isoelectric point, secondary structure composition (sheet, turn, and helix), aromaticity, Grand Average of Hydropathy (GRAVY), global structural flexibility and instability index using the ProtParam module of Biopython (Kyte and Doolittle 1982, Guruprasad *et al.* 1990, Bjellqvist *et al.* 1993, 1994, Lobry and Gautier 1994, Vihinen *et al.* 1994, Cock *et al.* 2009). We compared the prediction accuracy of these features using Receiver Operating Characteristic (ROC) analysis. To our surprise, flexibility outperformed other features in predicting protein solubility [Fig 1, Area Under the ROC Curve (AUC) = 0.67]. We then calculated 9,920 miscellaneous protein sequence properties using the 'protr' R package (Xiao *et al.* 2015), which include amino acid composition, pseudo-amino acid composition, autocorrelation, CTD (Composition, Transition, Distribution), conjoint triad descriptors, quasi-sequence-order descriptors and profile-based descriptors (Xiao *et al.* 2015). Among these features, the amphiphilic pseudo-amino acid composition for cysteine residues showed the highest AUC score, which is still lower than the AUC score for flexibility (Supplementary Fig S1 and Table S2, AUC = 0.65).

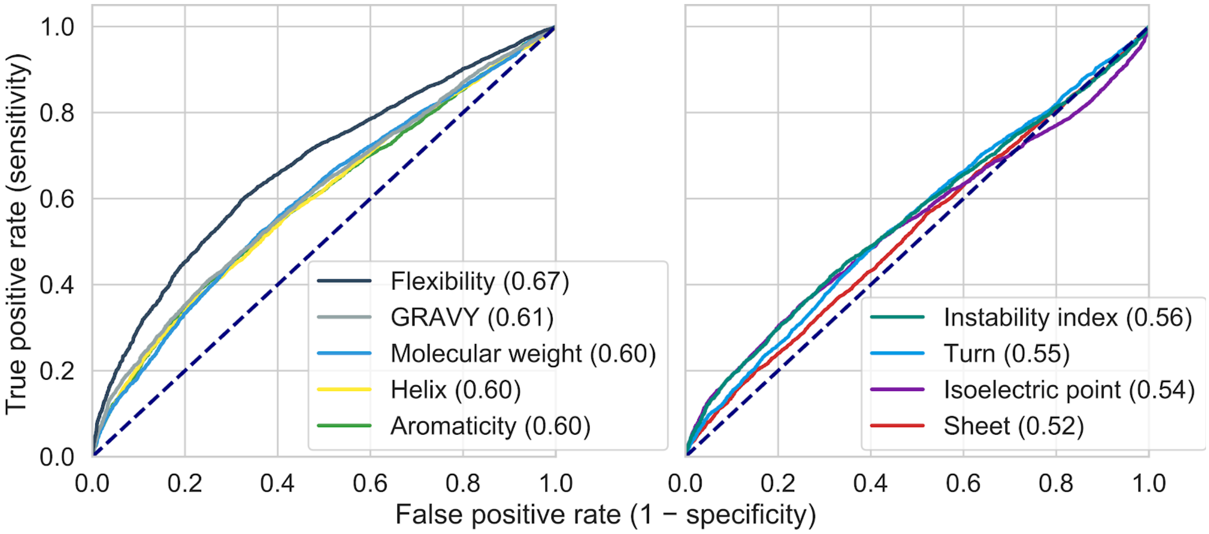

Fig 1. Global structural flexibility outperforms the other standard protein sequence properties in protein solubility prediction. ROC analysis of the standard protein sequence features for predicting the solubility of 12,216 recombinant proteins expressed in *E. coli* (the PSI:Biology dataset). AUC scores (perfect = 1.00, random = 0.50) are shown in parentheses. The ROC curves are shown in two separate panels for clarity. Dashed lines denote the performance of random classifiers. AUC, Area Under the ROC Curve; GRAVY, Grand Average of Hydropathy; PSI:Biology, Protein Structure Initiative:Biology; ROC, Receiver Operating Characteristic.

The Solubility-Weighted Index (SWI) is an improved approach to score solubility

Protein structural flexibility, in particular, the flexibility of local regions, is often associated with function (Craveur *et al.* 2015). The calculation of flexibility is usually performed by assigning a set of 20 normalised B-factors—a measure of vibration of C alpha atoms (see Discussion)—to a protein sequence and averaging the values by a sliding window approach (Karplus and Schulz 1985, Ragone *et al.* 1989, Vihinen *et al.* 1994, Smith *et al.* 2003). We reasoned that such sliding window can be approximated by a more straightforward arithmetic mean for calculating global structural flexibility, which is analogous to the computation of GRAVY. We applied this arithmetic mean approach to the PSI:Biology dataset and compared different sets of published, normalised B-factors (Bhaskaran and Ponnuswamy 1988, Ragone *et al.* 1989, Vihinen *et al.* 1994, Smith *et al.* 2003) as follows:

$$\frac{1}{L}\left(\sum_{i=1}^{L} B_i\right) \qquad (1)$$

where B_i is the normalised B-factor of the amino acid residue at the position i, and L is the sequence length. Among these sets of B-factors, solubility scoring using the most recently published set of normalised B-factors produced the highest AUC score (Supplementary Fig S2, AUC = 0.66).

To improve the prediction accuracy, we initialised an iterative refinement method with the most recently published set of normalised B-factors. This was done by maximising AUC scores with the Nelder-Mead optimisation algorithm (Nelder and Mead 1965). In order to account for phylogenetic relationships between proteins we clustered all 12,216 PSI:Biology protein sequences by 10% similarity using USEARCH (Fig 2A and Supplementary Fig S3). Cross-validations were conducted in a way that ensures training and testing is performed on unrelated sequences. We calculated the solubility scores for the optimised weights using Equation 1 and the AUC scores for each cross-validation step. Our training and test AUC scores were 0.72 ± 0.00 and 0.71 ± 0.03, respectively, showing an improvement over flexibility in solubility prediction (mean ± standard deviation; Fig 2B and Supplementary Table S3).

The final weights were derived from the arithmetic means of the weights for individual amino acid residues obtained from the cross-validation step (Supplementary Table S4). Interestingly, we observed over a 20% change on the weights for cysteine (C) and histidine (H) residues (Fig 2C and Supplementary Table S4). These results were in agreement with the contributions of cysteine and histidine residues as shown by the AUC scores of the amphiphilic pseudo-amino acid compositions for cysteine and histidine residues (Supplementary Fig S1B). To ensure that these results are not artifacts, in particular due to the presence of polyhistidine-tags in all the sequences, we repeated the iterative refinement method using the same cross-validation sets without His tag sequences. The final weights with and without His tags are nearly identical, suggesting that the approach is not confounded by tag use (Supplementary Table S4, Spearman's rho = 1).

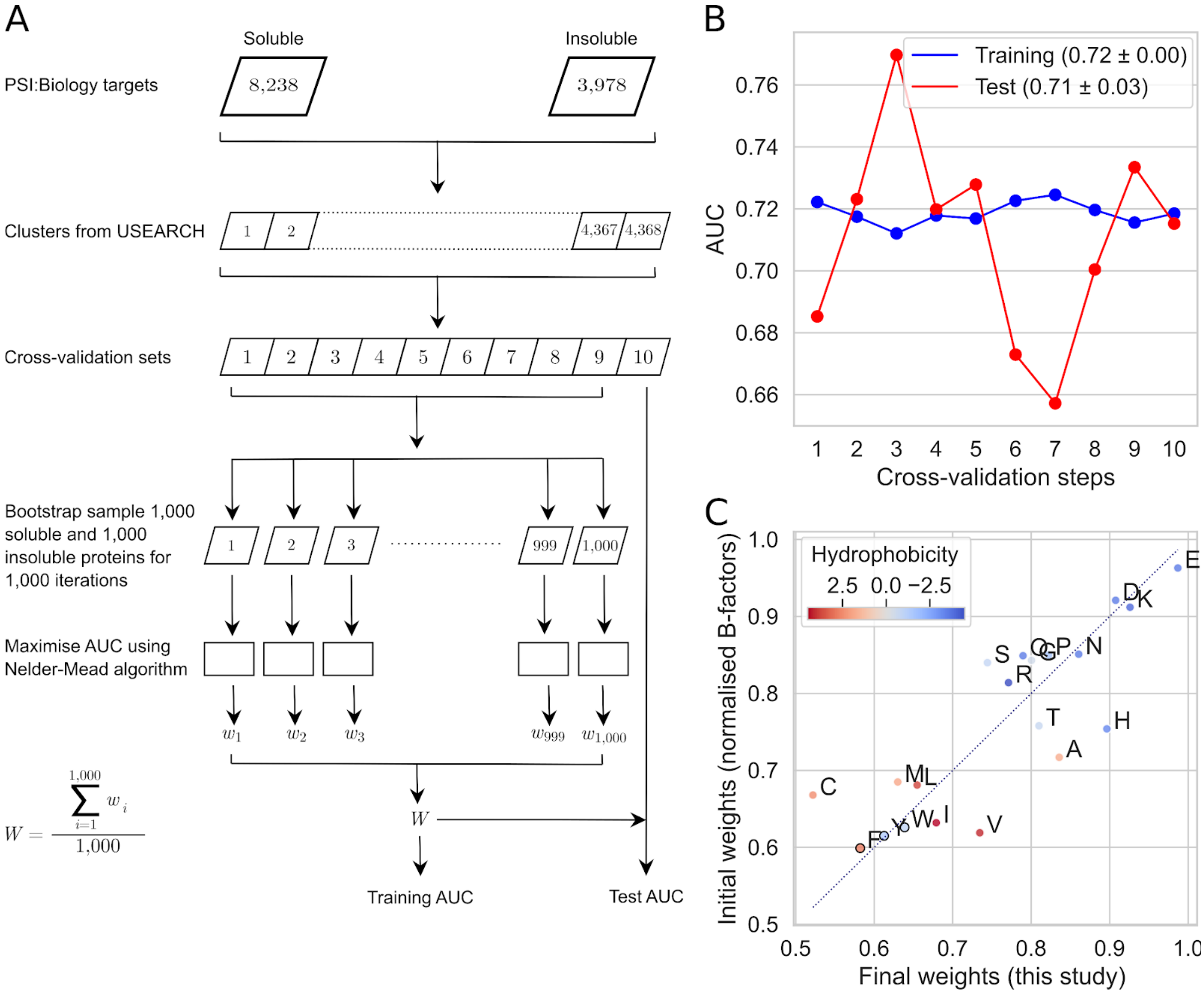

Fig 2. Derivation of the Solubility-Weighted Index (SWI). (A) Flow chart shows an iterative refinement of the most recently published set of normalised B-factors for solubility prediction (Smith *et al.* 2003). The solubility score of a protein sequence was calculated based on an arithmetic mean of the optimised weights as Equation 1 (using W instead of B). These scores were used to compute the AUC scores for training and test datasets. **(B)** Training and test performance of solubility prediction using the optimised weights for 20 amino acid residues in a 10-fold cross-validation (mean AUC ± standard deviation). Related data and figures are available as Supplementary Table S3 and Supplementary Fig S3. **(C)** Comparison between the 20 initial and final weights for amino acid residues. The final weights are derived from the arithmetic mean of the optimised weights from the cross-validation step. These weights are used to calculate SWI, the solubility score of a protein sequence, in the subsequent analyses. Filled circles, which represent amino acid residues, are colored by hydrophobicity (Kyte and Doolittle 1982). Solid black circles denote aromatic amino acid residues phenylalanine (F), tyrosine (Y), tryptophan (W). Dotted diagonal line represents no change in weight. Related data is available as Supplementary Table S4. AUC, Area Under the ROC Curve; ROC, Receiver Operating Characteristic; W , arithmetic mean of the weights of an amino acid residue optimised from 1,000 bootstrap samples in a cross-validation step.

To validate the cross-validation results, we used an independent dataset known as eSOL (Niwa *et al.* 2009). This dataset consists of the solubility percentages of *E. coli* proteins determined using an *E. coli* cell-free system (N = 3,198). Solubility scoring using the final weights showed a significant improvement in correlation with *E. coli* protein solubility over the initial weights (normalised B-factors) [Spearman's rho of 0.50 (P = 9.46 x 10^{-206}) vs 0.40 (P = 4.57 × 10^{-120})]. We call the solubility score of a protein sequence calculated using the final weights as the Solubility-Weighted Index (SWI).

We performed Spearman's correlation analysis for both the PSI:Biology and eSOL datasets. SWI shows the strongest correlation with solubility compared to the standard and 9,920 protein sequence properties (Fig 3 and Supplementary Fig S1). SWI also strongly correlates with flexibility, suggesting that SWI is still a good proxy for global structural flexibility.

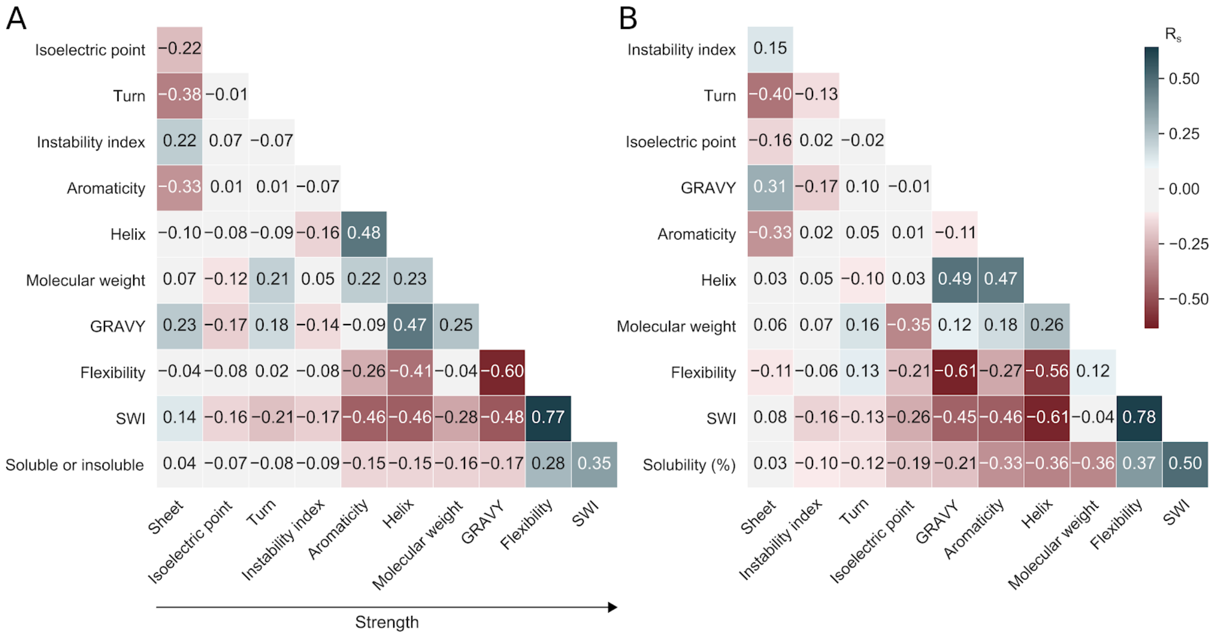

Fig 3. SWI strongly correlates with solubility. (A) Correlation matrix plot of the solubility of recombinant proteins expressed in *E. coli* and their standard protein sequence properties and SWI. These recombinant proteins are the PSI:Biology targets (N = 12,216) with a solubility status of 'Protein_Soluble' or 'Tested_Not_Soluble'**.** Related data is available as Supplementary Table S5. **(B)** Correlation matrix plot of the solubility percentages of *E. coli* proteins and their standard protein sequence properties and SWI. The solubility percentages were previously determined using an *E. coli* cell-free system (eSOL, N=3,198). Related data is available as Supplementary Table S6. GRAVY, Grand Average of Hydropathy; PSI:Biology, Protein Structure Initiative:Biology; R_s, Spearman's rho; SWI, Solubility-Weighted Index.

Next, we asked whether protein solubility can be predicted by surface amino acid residues. To address this question, we examined a previously published dataset for the protein surface 'stickiness' of 397 *E. coli* proteins (Levy *et al.* 2012). This dataset has the annotation for surface residues based on the protein crystal structures. Interestingly, we observed no correlation between the protein surface ‘stickiness’ and the solubility data from eSOL (Spearman’s rho = 0.05, P = 0.34). Optimising weights for surface residues as above led to no further improvements (i.e., the approach used to derive SWI; Spearman’s rho = 0.05, P = 0.31). In contrast, the SWI for these sequences has a significant correlation with solubility (Spearman’s rho = 0.45, P = 3.88 × 10^{-19}). These results suggest that full-length sequence should be taken into account when predicting protein solubility.

To understand the properties of soluble and insoluble proteins, we determined the enrichment of amino acid residues in the PSI:Biology targets relative to the eSOL sequences (see Methods). We observed that the PSI:Biology targets are enriched in charged residues lysine (K), glutamate (E) and aspartate (D), and depleted in aromatic residues tryptophan (W), albeit to a lesser extend for insoluble proteins (Fig 4A). As expected, cysteine residues (C) are enriched in the PSI:Biology insoluble proteins, supporting previous findings that cysteine residues contribute to poor solubility in the *E. coli* expression system (Wilkinson and Harrison 1991, Diaz *et al.* 2010).

In addition, we compared the SWI for random sequences with the PSI:Biology and eSOL sequences. In general, soluble proteins have higher SWI than insoluble proteins (Fig 4B). Interestingly, true biological sequences tend to have higher SWI than random sequences, highlighting a clear evolutionary selection for solubility.

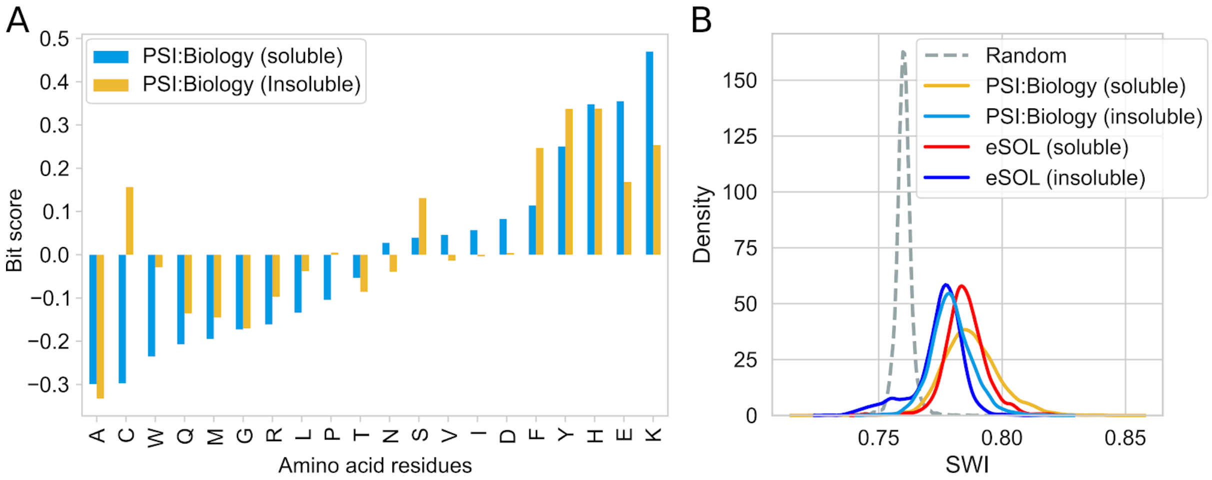

Fig 4. Properties of soluble and insoluble proteins. (A) Enrichment of amino acid residues in the PSI:Biology targets relative to eSOL sequences (N = 12,216 and 3,198, respectively). **(B)** Distribution of the SWI for soluble and insoluble proteins, and random sequences. eSOL sequences were grouped into soluble and insoluble proteins, i.e, <30% and >70% solubilities, respectively (Niwa *et al.* 2009) (Supplementary Table S1B). Random sequences were generated from a length of 50 to 6,000 amino acid residues, with an increment of 50 residues. A total of 12,000 random sequences were generated, 100 sequences for each length. PSI:Biology, Protein Structure Initiative:Biology; SWI, Solubility-Weighted Index.

SWI outperforms many protein solubility prediction tools

To confirm the usefulness of SWI in solubility prediction, we compared it with the existing tools including Protein-Sol (Hebditch *et al.* 2017), CamSol v2.1 (Sormanni *et al.* 2015, 2017), PaRSnIP (Rawi *et al.* 2018), DeepSol v0.3 (Khurana *et al.* 2018), the Wilkinson-Harrison model (Wilkinson and Harrison 1991, Davis *et al.* 1999, Harrison 2000), and ccSOL omics (Agostini *et al.* 2014). SWI outperforms other tools except for Protein-Sol in predicting *E. coli* protein solubility (Table 1). SWI is also the fastest solubility prediction algorithm (Table 1, Fig 5 and Supplementary Table S7).

Table 1. Comparison of protein solubility prediction methods and software.

	Approaches	Features	Runtime[a] (s per sequence)	PSI:Biology[b] (AUC)	eSOL[b] [R_s (P-value)]
SWI	● Arithmetic mean (this study). ● A set of 20 values for amino acid residues derived from normalised B-factors (Smith *et al.* 2003) by the Nelder-Mead simplex algorithm. ● Trained and tested using the PSI:Biology dataset curated by DNASU (Seiler *et al.* 2014). ● Available at https://tisigner.com/sodope	1	**0.00 ± 0.00**	**0.71 ± 0.03**[c]	0.50 (9.46 x 10^{-206})
Protein-Sol	● Linear model (Hebditch *et al.* 2017). ● Trained and tested using eSOL dataset (Niwa *et al.* 2009). ● Available at https://protein-sol.manchester.ac.uk/	10	1.16 ± 0.75	0.68	**0.54 (2.37 x 10^{-240})**
Flexibility	● A sliding window of 9 amino acid residues (Vihinen *et al.* 1994). ● Normalised B-factors derived from PDB. ● Available at https://github.com/biopython/biopython	1	0.38 ± 0.04	0.67	0.37 (7.73 x 10^{-106})
DeepSol S2	● Neural network models (Khurana *et al.* 2018). ● Trained and tested using a PSI:Biology dataset curated by ccSOL omics.	57 (11 types)	2069.77 ± 1613.63	0.67[d]	0.23 (5.82 x 10^{-41})[d]
DeepSol			2075.93 ±	0.66[d]	0.35

S3	● Available at https://github.com/sameerkhurana10/DSOL_rv0.2		1613.80		(7.48 x 10^{-91})[d]
DeepSol S1			2081.93 ± 1612.71	0.64[d]	0.39 (9.52 x 10^{-116})[d]
CamSol intrinsic web server	● Linear and logistic regression models (Sormanni *et al.* 2015, 2017). ● Trained and tested using previously published datasets (Família *et al.* 2015). ● Available at http://www-vendruscolo.ch.cam.ac.uk/camsolmethod.html	4	NA	0.66	0.43 (4.53 x 10^{-148})
PaRSnIP	● Gradient boosting machine model (Rawi *et al.* 2018). ● Trained and tested using a PSI:Biology dataset curated by ccSOL omics. ● Available at https://github.com/RedaRawi/PaRSnIP	8,477 (14 types)	2055.50 ± 1621.11	0.61	0.29 (3.57 x 10^{-65})
Wilkinson-Harrison model	● Linear model using charge average and turn-forming residue fraction (Wilkinson and Harrison 1991, Davis *et al.* 1999, Harrison 2000). ● Available at https://github.com/brunoV/bio-tools-solubility-wilkinson	2	0.09 ± 0.00	0.55	-0.06 (1.16x 10^{-4})
ccSOL omics web server	● Support vector machine model (Agostini *et al.* 2014). ● Trained and tested using a PSI:Biology dataset curated in-house. ● Available at http://s.tartaglialab.com/new_submission/ccsol_omics_file	5	NA	0.51	-0.02 (0.18)

Boldface values are the best results.

[a] The runtime was reported at the level of machine precision (mean seconds ± standard deviation). A total of 10 sequences were chosen from the PSI:Biology and eSOL datasets, related to Fig 5 (see Methods).

[b] The sample sizes for PSI:Biology and eSOL datasets are 12,216 and 3,198, respectively.

[c] Mean AUC ± standard deviation calculated from a 10-fold cross-validation (see Methods).

[d] DeepSol reports solubility prediction as probability and binary classes. The probability of solubility was used to calculate AUC and Spearman's correlation due to better results.

AUC, Area Under the ROC Curve; NA, not applicable; PDB, Protein Data Bank; PSI:Biology, Protein Structure Initiative:Biology; ROC, Receiver Operating Characteristic; R_s, Spearman's rho; SWI, Solubility-Weighted Index; s, seconds.

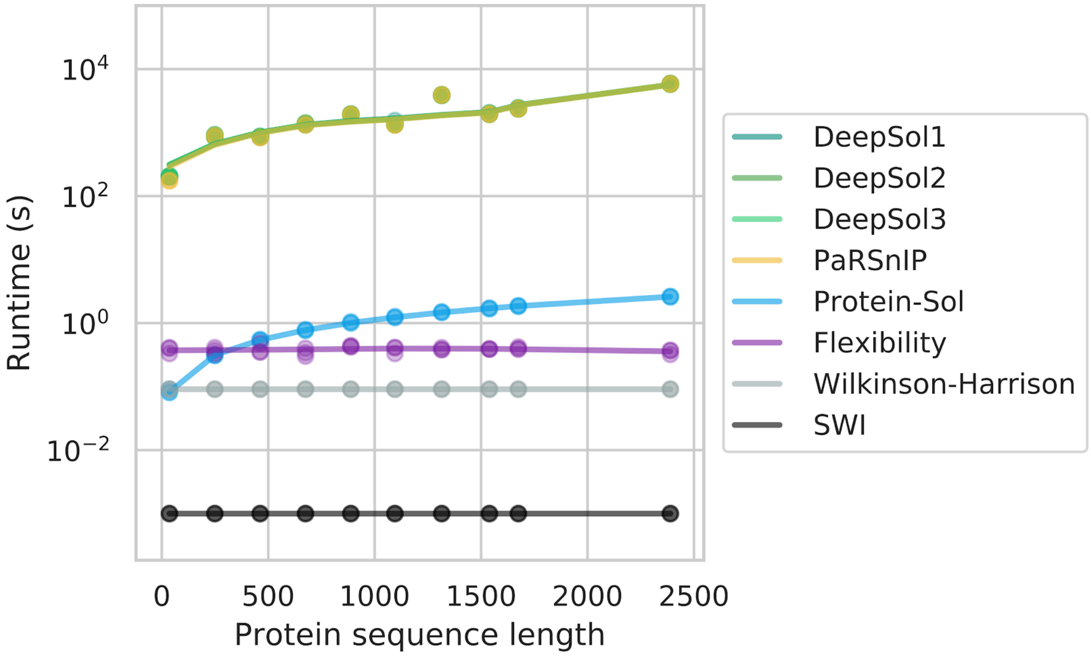

Fig 5. Runtime of protein solubility prediction tools per sequence. All the tools were run three times using 10 sequences selected from the PSI:Biology and eSOL datasets. A pseudocount of 0.001 s was used because the runtime of our SWI C program is 0.00 s per sequence, which is determined by machine precision. Related data is available as Supplementary Table S7. SWI, Solubility-Weighted Index; s, seconds.

To demonstrate a use case for SWI, we developed the Soluble Domain for Protein Expression (SoDoPE) web server (see Methods and https://tisigner.com/sodope). Upon sequence submission, the SoDoPE web server enables users to navigate the protein sequence and its domains for predicting and maximising protein expression and solubility.

DISCUSSION

The B-factor or temperature factor of the atoms in a crystalline structure is the measure of vibration around their mean position (u) that reflects the uncertainty in X-ray scattering structure determination (Schlessinger and Rost 2005, Bramer and Wei 2018, Carugo 2018).

$$B = 8\pi^2 u \qquad (2)$$

The profile of normalised B-factors along a protein sequence can be used to infer the flexibility and dynamics of the protein structure (Karplus and Schulz 1985, Vihinen *et al.*

1994). Protein structural flexibility has been associated with conformal variations, functions, thermal stability, ligand binding and disordered regions (Vihinen 1987, Teague 2003, Radivojac 2004, Ma 2005, Schlessinger and Rost 2005, Yuan *et al.* 2005, Yin *et al.* 2011). However, the use of flexibility in solubility prediction has been overlooked although their relationship has previously been proposed (Tsumoto *et al.* 2003). In this study, we have shown that flexibility strongly correlates with solubility (Fig 3). Based on the normalised B-factors used to compute flexibility, we have derived a new position and length independent weights to score the solubility of a given protein sequence. We call this protein solubility score as SWI.

Upon further inspection, we observe some interesting properties in SWI. SWI anti-correlates with helix propensity, GRAVY, aromaticity and isoelectric point (Fig 2C and 3). Amino acid residues with a lower aromaticity or hydrophilic are known to improve protein solubility (Han *et al.* n.d., Wilkinson and Harrison 1991, Trevino *et al.* 2007, Niwa *et al.* 2009, Kramer *et al.* 2012, Warwicker *et al.* 2014). Consistent with previous studies, the charged residues aspartate (D), glutamate (E) and lysine (K) are associated with high solubility, whereas the aromatic residues phenylalanine (F), tryptophan (W) and tyrosine (Y) are associated with low solubility (Fig 2C and 4A). Interestingly, histidine residue (H) appears as one of the heavily weighted residues in scoring solubility, which might be due to its positive charge. In contrast, cysteine residue (C) has been strongly downweighted, probably because disulfide bonds couldn't be properly formed in the *E. coli* expression hosts (Stewart *et al.* 1998, Aslund and Beckwith 1999, Rosano and Ceccarelli 2014, Jia and Jeon 2016). The weights are likely different if the solubility analysis was done using the reductase-deficient, *E. coli* Origami host strains, or eukaryotic hosts.

Higher helix propensity has been reported to increase solubility (Idicula-Thomas and Balaji 2005, Huang *et al.* 2012). However, our analysis has shown that helical and turn propensities anti-correlate with solubility, whereas sheet propensity lacks correlation with solubility, suggesting that disordered regions may tend to be more soluble (Fig 3). In accordance with these, SWI has stronger negative correlations with helix and turn propensities. These findings also suggest that protein solubility can be largely explained by overall amino acid composition, not just the surface amino acid residues. This idea aligns with our understanding that protein solubility and folding are closely linked, and folding occurs cotranscriptionally, a complex process that is driven various intrinsic and extrinsic factors (Wilkinson and Harrison 1991, Chiti *et al.* 2003, Tartaglia *et al.* 2004, Diaz *et al.* 2010). However, it is unclear why sheet propensity has little contribution to solubility because β-sheets have been shown to link closely with protein aggregation (Idicula-Thomas and Balaji 2005).

We conclude that SWI is a well-balanced index that is relatively simple and easy to use. To demonstrate the usefulness of SWI, we developed the SoDoPE web server for predicting solubility and designing protein sequences (see Methods and https://tisigner.com/sodope). In addition, SoDoPE is integrated with TIsigner, our gene optimisation web server for protein expression. This pipeline provides a holistic approach to improve the outcome of recombinant protein expression.

METHODS

Protein sequence properties

The standard protein sequence properties were calculated using the Bio.SeqUtils.ProtParam module of Biopython v1.73 (Cock *et al.* 2009). All miscellaneous protein sequence properties were computed using the R package protr v1.6-2 (Xiao *et al.* 2015).

Protein solubility prediction

We used the standard and miscellaneous protein sequence properties to predict the solubility of the PSI:Biology and eSOL targets (N=12,216 and 3,198, respectively) (Niwa *et al.* 2009, Seiler *et al.* 2014). For method comparison, we chose the protein solubility prediction tools that are scalable (Table 1). Default configurations were used for running the command line tools.

To benchmark the runtime of these solubility prediction tools, we selected 10 sequences with a large range of lengths from the PSI:Biology and eSOL datasets (from 36 to 2389 residues). All the tools were run and timed using a single process without using GPU on a high performance computer [/usr/bin/time <command>; CentOS Linux 7 (Core) operating system, 72 cores in 2× Broadwell nodes (E5-2695v4, 2.1 GHz, dual socket 18 cores per socket), 528 GiB memory]. Single sequence fasta files were used as input files.

SWI

To improve protein solubility prediction, we optimised the most recently published set of normalised B-factors using the PSI:Biology dataset (Smith *et al.* 2003) (Fig 2). To avoid bias due to protein sequence homology, we first clustered the PSI:Biology targets using USEARCH v11.0.667, 32-bit (Edgar 2010). His tag sequences were removed from all sequences before clustering to minimise bias. We obtained 4,368 clusters using the parameters: -cluster_fast <input_file> -id 0.1 -msaout <output_file> -threads 4. These clusters were divided into 10 groups with approximately 1,200 sequences per group. The subsequent steps were done with or without His tag sequences. We used the normalised B-factors as the initial weights to maximise AUC using these 10 groups with a 10-fold cross-validation. Since AUC is non-differentiable, we used the Nelder-Mead optimisation method (implemented in SciPy v1.2.0), which is a derivative-free, heuristic, simplex-based optimisation (Nelder and Mead 1965, Oliphant 2007, Millman and Aivazis 2011). For each step in cross-validation, we did bootstrap resampling for 1,000 times with each sample containing 1,000 soluble and 1,000 insoluble proteins. Optimisation was done for each sample, giving 1,000 sets of weights. The arithmetic mean of these weights was used to determine the training and test AUC for the cross-validation step (Fig 2A).

Bit score

To compute the bit scores for each amino acid residue in the PSI:Biology soluble and insoluble groups (Fig 4A), we normalised the count of each residue (x) in each group by the total number of residues in that group. We used the normalised count of amino acid residues using the eSOL sequences as the background. The bit score of residue (x) for soluble or insoluble group is then given by the following equation:

$$bit\ score\ (x)_i = log_2\left(\frac{f_i(x)}{f_{eSOL}(x)}\right),\ i = [soluble,\ insoluble] \qquad (3)$$

where $f_i(x)$ is the normalised count of residue (x) in the PSI:Biology soluble or insoluble group and $f_{eSOL}(x)$ is the normalised count in the eSOL sequences.

For control, random protein sequences were generated by incrementing the length of sequence, starting from a length of 50 residues to 6,000 residues with a step size of 50 residues. A hundred of random sequences were generated for each length, giving a total of 12,000 unique random sequences.

The SoDoPE web server

To estimate the probability of solubility using SWI, we fitted the following logistic regression to the PSI:Biology dataset:

$$probability\ of\ solubility = 1/(1 + exp(-(ax + b))) \qquad (4)$$

where, x is the SWI of a given protein sequence, $a = 81.1496$ and $b = -62.8379$. The P-value of log-likelihood ratio test was less than machine precision. Equation 4 can be used to predict the solubility of a protein sequence given that the protein is successfully expressed in *E. coli*.

On this basis, we developed a solubility prediction webservice called the Soluble Domain for Protein Expression (SoDoPE). Our web server accepts either a nucleotide or amino acid sequence. Upon sequence submission, a query is sent to the HMMER web server to annotate protein domains (https://www.ebi.ac.uk/Tools/hmmer/) (Potter *et al.* 2018). Once the protein domains are identified, users can choose a domain or any custom region (including full-length sequence) to examine the probability of solubility, flexibility and GRAVY. This functionality enables protein biochemists to plan their experiments and opt for the domains or regions with high probability of solubility. Furthermore, we implemented a simulated annealing algorithm that maximised the probability of solubility for a given region by generating a list of regions with extended boundaries. Users can also predict the improvement in solubility by selecting a commonly used solubility tag or a custom tag.

We linked SoDoPE with TIsigner, which is our existing web server for maximising the accessibility of translation initiation site (Bhandari *et al.* 2019). This pipeline allows users to predict and optimise both protein expression and solubility for a gene of interest. The SoDoPE web server is freely available at https://tisigner.com/sodope.

Statistical analysis

Data analysis was done using Pandas v0.25.3 (McKinney 2010), scikit-learn v0.20.2 (Pedregosa *et al.* 2011), numpy v1.16.2 (van der Walt *et al.* 2011) and statsmodel v0.10.1(Seabold and Perktold 2010). Plots were generated using Matplotlib v3.0.2 (Caswell *et al.* 2018) and Seaborn v0.9.0 (Waskom *et al.* 2014).

Code and data availability

Jupyter notebook of our analysis can be found at https://github.com/Gardner-BinfLab/SoDoPE_paper_2019. The source code for our solubility prediction server (SoDoPE) can be found at https://github.com/Gardner-BinfLab/TIsigner.

ACKNOWLEDGEMENTS

We thank New Zealand eScience Infrastructure for providing a high performance computing platform. We are grateful to Harry Biggs for proofreading our manuscript and providing feedback for the web server. This work was supported by the Ministry of Business, Innovation and Employment, New Zealand (MBIE grant: UOOX1709).

AUTHOR CONTRIBUTIONS

C.S.L. conceived the work; B.K.B. and C.S.L. analysed the data and C.S.L. contributed flexibility analysis; B.K.B. and P.P.G formulated SWI; B.K.B. developed the SoDoPE web server; B.K.B., P.P.G. and C.S.L. wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

Acton, T.B., Gunsalus, K.C., Xiao, R., Ma, L.C., Aramini, J., Baran, M.C., Chiang, Y.-W., Climent, T., Cooper, B., Denissova, N.G., Douglas, S.M., Everett, J.K., Ho, C.K., Macapagal, D., Rajan, P.K., Shastry, R., Shih, L.-Y., Swapna, G.V.T., Wilson, M., Wu, M., Gerstein, M., Inouye, M., Hunt, J.F., and Montelione, G.T., 2005. Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. *Methods in enzymology*, 394, 210–243.

Agostini, F., Cirillo, D., Livi, C.M., Delli Ponti, R., and Tartaglia, G.G., 2014. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. *Bioinformatics* , 30 (20), 2975–2977.

Aslund, F. and Beckwith, J., 1999. The thioredoxin superfamily: redundancy, specificity, and gray-area genomics. *Journal of bacteriology*, 181 (5), 1375–1379.

Bhandari, B.K., Lim, C.S., and Gardner, P.P., 2019. Highly accessible translation initiation sites are predictive of successful heterologous protein expression. *bioRxiv*.

Bhaskaran, R. and Ponnuswamy, P.K., 1988. Positional flexibilities of amino acid residues in globular proteins. *International Journal of Peptide and Protein Research*.

Bjellqvist, B., Basse, B., Olsen, E., and Celis, J.E., 1994. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis*, 15 (3-4), 529–539.

Bjellqvist, B., Hughes, G.J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.C., Frutiger, S., and Hochstrasser, D., 1993. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, 14 (10), 1023–1031.

Bramer, D. and Wei, G.-W., 2018. Blind prediction of protein B-factor and flexibility. *The Journal of chemical physics*, 149 (13), 134107.

Carugo, O., 2018. How large B-factors can be in protein crystal structures. *BMC bioinformatics*, 19 (1), 61.

Caswell, T.A., Droettboom, M., Hunter, J., Firing, E., Lee, A., Stansby, D., de Andrade, E.S., Nielsen, J.H., Klymak, J., Varoquaux, N., Root, B., Elson, P., Dale, D., May, R., Lee, J.-J., Seppänen, J.K., Hoffmann, T., McDougall, D., Straw, A., Hobson, P., cgohlke, Yu, T.S., Ma, E., Vincent, A.F., Silvester, S., Moad, C., Katins, J., Kniazev, N., Ariza, F., and Würtz, P., 2018. *matplotlib/matplotlib v3.0.2*.

Chan, W.-C., Liang, P.-H., Shih, Y.-P., Yang, U.-C., Lin, W.-C., and Hsu, C.-N., 2010. Learning to predict expression efficacy of vectors in recombinant protein production. *BMC bioinformatics*, 11 Suppl 1, S21.

Chen, L., Oughtred, R., Berman, H.M., and Westbrook, J., 2004. TargetDB: a target registration database for structural genomics projects. *Bioinformatics* , 20 (16), 2860–2862.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C.M., 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, 424 (6950), 805–808.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* , 25 (11), 1422–1423.

Costa, S., Almeida, A., Castro, A., and Domingues, L., 2014. Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: the novel Fh8 system. *Frontiers in microbiology*, 5, 63.

Craveur, P., Joseph, A.P., Esque, J., Narwani, T.J., Noël, F., Shinada, N., Goguet, M., Leonard, S., Poulain, P., Bertrand, O., Faure, G., Rebehmed, J., Ghozlane, A., Swapna, L.S., Bhaskara, R.M., Barnoud, J., Téletchéa, S., Jallu, V., Cerny, J., Schneider, B., Etchebest, C., Srinivasan, N., Gelly, J.-C., and de Brevern, A.G., 2015. Protein flexibility in the light of structural alphabets. *Frontiers in molecular biosciences*, 2, 20.

Davis, G.D., Elisee, C., Newham, D.M., and Harrison, R.G., 1999. New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnology and bioengineering*, 65 (4), 382–388.

Diaz, A.A., Tomba, E., Lennarson, R., Richard, R., Bagajewicz, M.J., and Harrison, R.G., 2010. Prediction of protein solubility in Escherichia coli using logistic regression. *Biotechnology and bioengineering*, 105 (2), 374–383.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* , 26 (19), 2460–2461.

Esposito, D. and Chatterjee, D.K., 2006. Enhancement of soluble protein expression through the use of fusion tags. *Current opinion in biotechnology*, 17 (4), 353–358.

Família, C., Dennison, S.R., Quintas, A., and Phoenix, D.A., 2015. Prediction of Peptide and

Protein Propensity for Amyloid Formation. *PloS one*, 10 (8), e0134679.

Guruprasad, K., Reddy, B.V., and Pandit, M.W., 1990. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein engineering*, 4 (2), 155–161.

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

Han, X., Ning, W., Ma, X., Wang, X., and Zhou, K., n.d. Improve Protein Solubility and Activity based on Machine Learning Models.

Harrison, R.G., 2000. Expression of soluble heterologous proteins via fusion with NusA protein. *Innovations*, 11, 4–7.

Hebditch, M., Carballo-Amador, M.A., Charonis, S., Curtis, R., and Warwicker, J., 2017. Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* , 33 (19), 3098–3100.

Heckmann, D., Lloyd, C.J., Mih, N., Ha, Y., Zielinski, D.C., Haiman, Z.B., Desouki, A.A., Lercher, M.J., and Palsson, B.O., 2018. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature communications*, 9 (1), 5252.

Hirose, S. and Noguchi, T., 2013. ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics*, 13 (9), 1444–1456.

Hou, Q., Bourgeas, R., Pucci, F., and Rooman, M., 2018. Computational analysis of the amino acid interactions that promote or decrease protein solubility. *Scientific Reports*.

Huang, H.-L., Charoenkwan, P., Kao, T.-F., Lee, H.-C., Chang, F.-L., Huang, W.-L., Ho, S.-J., Shu, L.-S., Chen, W.-L., and Ho, S.-Y., 2012. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC bioinformatics*, 13 Suppl 17, S3.

Idicula-Thomas, S. and Balaji, P.V., 2005. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein science: a publication of the Protein Society*, 14 (3), 582–592.

Jia, B. and Jeon, C.O., 2016. High-throughput recombinant protein expression in Escherichia coli: current status and future perspectives. *Open biology*, 6 (8).

Karplus, P.A. and Schulz, G.E., 1985. Prediction of chain flexibility in proteins. *Die Naturwissenschaften*, 72 (4), 212–213.

Khurana, S., Rawi, R., Kunji, K., Chuang, G.-Y., Bensmail, H., and Mall, R., 2018. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* , 34 (15), 2605–2613.

Kramer, R.M., Shende, V.R., Motl, N., Nick Pace, C., and Martin Scholtz, J., 2012. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophysical Journal*.

Kyte, J. and Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157 (1), 105–132.

Levy, E.D., De, S., and Teichmann, S.A., 2012. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proceedings of the National Academy of Sciences*.

Lobry, J.R. and Gautier, C., 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic acids research*, 22 (15), 3174–3180.

Ma, J., 2005. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* , 13 (3), 373–380.

McKinney, W., 2010. Data Structures for Statistical Computing in Python. *In*: *Proceedings of the 9th Python in Science Conference*. 51–56.

Millman, K.J. and Aivazis, M., 2011. Python for Scientists and Engineers. *Computing in Science Engineering*, 13 (2), 9–12.

Nelder, J.A. and Mead, R., 1965. A Simplex Method for Function Minimization. *Computer Journal*, 7 (4), 308–313.

Niwa, T., Ying, B.-W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H., 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (11), 4201–4206.

Oliphant, T.E., 2007. Python for Scientific Computing. *Computing in Science Engineering*, 9 (3), 10–20.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. *Journal of machine learning research: JMLR*, 12 (Oct), 2825–2830.

Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., and Finn, R.D., 2018. HMMER web server: 2018 update. *Nucleic acids research*, 46 (W1), W200–W204.

Radivojac, P., 2004. Protein flexibility and intrinsic disorder. *Protein Science*.

Ragone, R., Facchiano, F., Facchiano, A., Facchiano, A.M., and Colonna, G., 1989. Flexibility plot of proteins. *'Protein Engineering, Design and Selection'*.

Rawi, R., Mall, R., Kunji, K., Shen, C.-H., Kwong, P.D., and Chuang, G.-Y., 2018. PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics*.

Rosano, G.L. and Ceccarelli, E.A., 2014. Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in microbiology*, 5, 172.

Schlessinger, A. and Rost, B., 2005. Protein flexibility and rigidity predicted from sequence. *Proteins*, 61 (1), 115–126.

Seabold, S. and Perktold, J., 2010. Statsmodels: Econometric and statistical modeling with python. *In*: *Proceedings of the 9th Python in Science Conference*.

Seiler, C.Y., Park, J.G., Sharma, A., Hunter, P., Surapaneni, P., Sedillo, C., Field, J., Algar, R., Price, A., Steel, J., Throop, A., Fiacco, M., and LaBaer, J., 2014. DNASU plasmid and PSI:Biology-Materials repositories: resources to accelerate biological research. *Nucleic acids research*, 42 (Database issue), D1253–60.

Sharp, P.M. and Li, W.H., 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 15 (3), 1281–1295.

Smith, D.K., Radivojac, P., Obradovic, Z., Dunker, A.K., and Zhu, G., 2003. Improved amino acid flexibility parameters. *Protein science: a publication of the Protein Society*, 12 (5), 1060–1072.

Sormanni, P., Amery, L., Ekizoglou, S., Vendruscolo, M., and Popovic, B., 2017. Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Scientific reports*, 7 (1), 8200.

Sormanni, P., Aprile, F.A., and Vendruscolo, M., 2015. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology*.

Stewart, E.J., Aslund, F., and Beckwith, J., 1998. Disulfide bond formation in the Escherichia coli cytoplasm: an in vivo role reversal for the thioredoxins. *The EMBO journal*, 17 (19), 5543–5550.

Tartaglia, G.G., Cavalli, A., Pellarin, R., and Caflisch, A., 2004. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein science: a publication of the Protein Society*, 13 (7), 1939.

Teague, S.J., 2003. Implications of protein flexibility for drug discovery. *Nature reviews. Drug discovery*, 2 (7), 527–541.

Trevino, S.R., Martin Scholtz, J., and Nick Pace, C., 2007. Amino Acid Contribution to Protein Solubility: Asp, Glu, and Ser Contribute more Favorably than the other Hydrophilic Amino Acids in RNase Sa. *Journal of Molecular Biology*.

Tsumoto, K., Ejima, D., Kumagai, I., and Arakawa, T., 2003. Practical considerations in refolding proteins from inclusion bodies. *Protein expression and purification*, 28 (1), 1–8.

Vihinen, M., 1987. Relationship of protein flexibility to thermostability. *'Protein Engineering, Design and Selection'*.

Vihinen, M., Torkkila, E., and Riikonen, P., 1994. Accuracy of protein flexibility predictions. *Proteins*, 19 (2), 141–149.

Waldo, G.S., 2003. Genetic screens and directed evolution for protein solubility. *Current opinion in chemical biology*, 7 (1), 33–38.

van der Walt, S., Colbert, S.C., and Varoquaux, G., 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in science & engineering*, 13 (2), 22–30.

Warwicker, J., Charonis, S., and Curtis, R.A., 2014. Lysine and arginine content of proteins: computational analysis suggests a new tool for solubility design. *Molecular pharmaceutics*, 11 (1), 294–303.

Waskom, M., Botvinnik, O., Hobson, P., Cole, J.B., Halchenko, Y., Hoyer, S., Miles, A., Augspurger, T., Yarkoni, T., Megies, T., Coelho, L.P., Wehner, D., cynddl, Ziegler, E., diego, Zaytsev, Y.V., Hoppe, T., Seabold, S., Cloud, P., Koskinen, M., Meyer, K., Qalieh, A., and Allan, D., 2014. seaborn: v0.5.0 (November 2014).

Wilkinson, D.L. and Harrison, R.G., 1991. Predicting the solubility of recombinant proteins in Escherichia coli. *Bio/technology* , 9 (5), 443–448.

Wu, Z., Kan, S.B.J., Lewis, R.D., Wittmann, B.J., and Arnold, F.H., 2019. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences of the United States of America*, 116 (18), 8852–8858.

Xiao, N., Cao, D.-S., Zhu, M.-F., and Xu, Q.-S., 2015. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* , 31 (11), 1857–1859.

Xiao, R., Anderson, S., Aramini, J., Belote, R., Buchwald, W.A., Ciccosanti, C., Conover, K., Everett, J.K., Hamilton, K., Huang, Y.J., Janjua, H., Jiang, M., Kornhaber, G.J., Lee, D.Y., Locke, J.Y., Ma, L.-C., Maglaqui, M., Mao, L., Mitra, S., Patel, D., Rossi, P., Sahdev, S., Sharma, S., Shastry, R., Swapna, G.V.T., Tong, S.N., Wang, D., Wang, H., Zhao, L., Montelione, G.T., and Acton, T.B., 2010. The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *Journal of structural biology*, 172 (1), 21–33.

Yang, K.K., Wu, Z., and Arnold, F.H., 2019. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16 (8), 687–694.

Yin, H., Li, Y.-Z., and Li, M.-L., 2011. On the relation between residue flexibility and residue interactions in proteins. *Protein and peptide letters*, 18 (5), 450–456.

Yuan, Z., Bailey, T.L., and Teasdale, R.D., 2005. Prediction of protein B-factor profiles. *Proteins: Structure, Function, and Bioinformatics*.