SoluProt: Prediction of Protein Solubility

Jiří Hon[1,2,3], Martin Marušiak[3], Tomáš Martínek[3], Jaroslav Zendulka[3], David Bednář[1,2], and Jiří Damborský[1,2]

[1] Loschmidt Laboratories, Centre for Toxic Compounds in the Environment RECETOX and Department of Experimental Biology, Faculty of Science, Masaryk University, 625 00 Brno

[2] International Clinical Research Center, St. Annes's University Hospital Brno, 656 91 Brno

[3] IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 612 66 Brno

ihon@fit.vutbr.cz

Abstract. Protein solubility poses a major bottleneck in production of many therapeutic and industrially attractive proteins. Experimental solubilization attempts are plagued by relatively low success rates and often lead to the loss of biological activity. Therefore, any advance in computational prediction of protein solubility may reduce the cost of proteomics studies significantly. Here, we propose a novel software tool SoluProt for prediction of solubility from protein sequence. SoluProt achieved the best accuracy 58.2% and AUC 0.61 of all available tools at an independent balanced test set. While the absolute prediction performance is rather low, we show SoluProt can still help reduce costs of proteomic studies significantly by efficient prioritization of protein sequences. The main SoluProt contribution lies in improved preprocessing of noisy training data and sensible selection of sequence features included in the prediction model.

Keywords: protein · solubility · prediction · machine-learning

1 Introduction

The key biochemical process limited by protein solubility is a heterologous expression – a manipulation of a living organism, usually *E. coli*, to produce a target protein of interest. Unfortunately, many proteins heterogously expressed are not sufficiently soluble. Although various biochemical approaches aimed at improving protein solubility were developed, solubilization attempts are plagued by relatively low success rates and often lead to the loss of biological activity.

Recently, various software tools were developed for prediction of solubility based on protein sequence – from the first reasonable model proposed by Wilkinson and Harrison [4] (mWH) to a battery of tools using more advanced machine-learning techniques, namely SOLpro [7], PROSO II [9], ccSOL omics [1], Protein-Sol [5] and DeepSol [6]. The accuracy of available tools is, however, still very limited and the problem remains open. Moreover, the performance of the tools often turns up to be highly overestimated when later validated on larger test sets.

Here, we propose a novel software tool SoluProt that pushes the borders of solubility prediction from protein sequence. SoluProt was designed to reflect the current advance in protein research. The main effort was focused on dataset preprocessing, feature selection and unbiased comparison against existing methods.

2 SoluProt datasets

The main source of SoluProt training dataset is TargetTrack (TT) database [2]. As TT does not provide solubility information directly but only store reports from protein crystallization experiments, we inferred solubility computationally. All proteins that achieved soluble status or any subsequent stages requiring soluble expression, were considered soluble. When expression or purification failure was reported in the stop status, the protein was considered insoluble.

We did several additional steps to improve dataset quality. First of all, we performed expert-assisted keyword matching combined with manual checking of TT annotations to leave only proteins expressed in the most common expression organism *E. coli.* This significantly reduce the noise in the training dataset as it is known that protein soluble in one expression system might appear to be insoluble in another one. We also removed short proteins, sequences with undefined residues and transmembrane proteins predicted by *TMHMM* tool. Next, we removed insoluble sequences that now have protein structure solved in PDB database. This step reflects the current advance in molecular biology as some older TT proteins can now be produced and crystallized using current technology.

In the last step, we reduced sequence redundancy in the training set by clustering to 25% identity using *MMseqs2*. This was done separately for positive and negative samples to avoid simplification of the prediction problem. We also adjusted class and sequence length distribution of soluble and insoluble samples such that length alone was not decisive in the prediction and both classes were equally represented. Finally, the SoluProt training set consists of 10,912 protein sequences, having equal number of soluble and insoluble samples.

To build SoluProt test set, we used dataset of 9,703 proteins expressed in *E. coli* using unified high-throughput production pipeline of North East Structrual Consortium [8] (NESG dataset) that directly provides solubility value at integer scale from 0 to 5. The NESG dataset was processed with similar preprocessing and balancing pipeline as the training set, only skipping the computational solubility derivation and expression system filtration as these were pointless in this case. Finally, 3,788 sequences remained in the balanced SoluProt test set having no overlap with the training set.

3 SoluProt predictor

SoluProt predictor is implemented in Python using *scikit-learn*, *Biopython*, *pandas* and *NumPy* libraries. Random forests regression [3] (RF) model was used both for feature pruning and the final prediction.

Prediction features were selected from a set of 253 values in eight subgroups: i) single amino acid content, ii) amino acid dimer content, iii), average physico-chemical features, iv) average flexibility by *DynaMine*, v) secondary structure content by *FESS*, vi) average disorder by *ESPRITZ*, vii) amino acids in transmembrane helices by *TMHMM* and viii) maximal identity to sequences in PDB.

Finally, 36 features were used in the prediction model. The five most dominant features are: i) maximal identity to PDB, ii) lysine to arginine ratio, iii) amino acid content, iv) isoelectric point and v) amino acid dimer content.

Optimization of random forests hyper-parameters was performed using 5-fold cross-validation and grid search on the SoluProt training set. The overtraining was quantified by a difference between the accuracy on the training and validation part while lower difference was preferred. The final model had average cross-validation accuracy 69% on training part, 65.3% on validation part.

SoluProt achieved the best accuracy 58.2% (at threshold 0.53) and AUC 0.61 on the SoluProt test set from all available tools (see Table 1 and Figure 1). On the test set, SoluProt reports 969 true positives, 1,237 true negatives, 657 false positives and 925 false negatives. This indicates that proper prediction of positive samples is the most problematic part of the solubility prediction problem.

Table 1. Performance of sequence-based solubility prediction tools on SoluProt test set. The best possible accuracy and corresponding threshold is presented.

Tool	AUC	Accuracy	Threshold
SoluProt	0.607	0.582	0.530
PROSOII	0.596	0.572	0.610
ESPRESSO	0.571	0.555	0.561
DeepSol	0.552	0.541	0.443
mWH	0.544	0.539	0.561
Protein-Sol	0.545	0.536	0.617
SOLpro	0.521	0.523	0.376
ccSOL omics	0.512	0.517	0.750

Although SoluProt is the best performing tool, its absolute acurracy is rather low. However, SoluProt can still be very useful for protein sequence prioritization (see Figure 1). Using SoluProt prediction score on SoluProt test set to leave only 10% best proteins will increase the number of soluble proteins by nearly 50% in comparison to what would be expected when selecting the same amount of proteins from the test set just randomly.

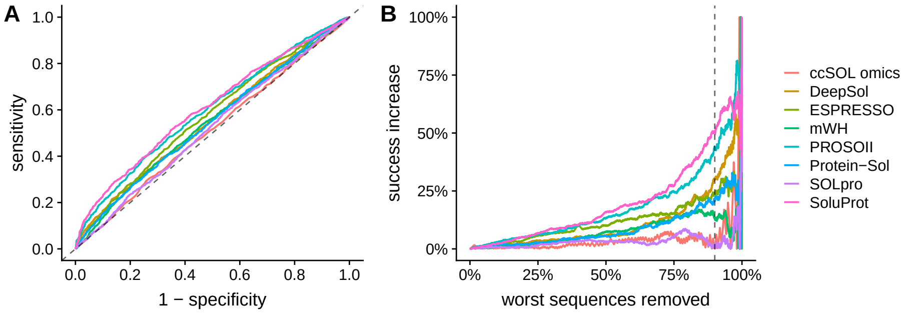

Fig. 1. Performance comparison. (A) ROC curves on SoluProt test set. (B) Increase of soluble sequences in prioritization task on the SoluProt test set in comparison to random subsampling. For SoluProt, removing 90% of worst scoring sequences (dashed line) will increase the number of soluble proteins by nearly 50%.

4 Conclusion

Our SoluProt tool moves the computational solubility prediction one step forward. The main SoluProt contribution lies in improved preprocessing of noisy TargetTrack training data and sensible selection of sequence features included in the prediction model. SoluProt achieved the best accuracy 58.2% and AUC 0.61 while being also the best tool in the task of protein prioritization.

However, there is still space for improvement. Currently, the protein sequence is often translated to numeric features that ignore amino acid order. Consequently a protein sequence and its randomized variant can have the same feature value while their solubility almost certainly differ. This could be solved by embedding protein sequence into a vector space that would preserve sequence similarity. Another option might be application of current advance in computational modelling of protein structures to build spatial model from the protein sequence and subsequently derive more descriptive features that reflect the protein structure.

References

1. Agostini, F., Vendruscolo, M., Tartaglia, G.G.: Sequence-based prediction of protein solubility. Journal of Molecular Biology **421**(2-3), 237–241 (2012)
2. Berman, H.M., Gabanyi, M.J., Protein Structure Initiative Network Of Investigators: Protein Structure Initiative – TargetTrack 2000–2017 – All Data Files (2017)
3. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
4. Davis, G.D., Elisee, C., Newham, D.M., Harrison, R.G.: New fusion protein systems designed to give soluble expression in Escherichia coli. Biotechnol. Bioeng. **65**(4), 382–388 (1999)
5. Hebditch, M., Carballo-Amador, M.A., Charonis, S., Curtis, R., Warwicker, J.: Protein–Sol: a web tool for predicting protein solubility from sequence. Bioinformatics **33**(19), 3098–3100 (2017)

6. Khurana, S., Rawi, R., Kunji, K., Chuang, G.Y., Bensmail, H., Mall, R.: DeepSol: a deep learning framework for sequence-based protein solubility prediction. Bioinformatics (2018)
7. Magnan, C.N., Randall, A., Baldi, P.: SOLpro: accurate sequence-based prediction of protein solubility. Bioinformatics **25**(17), 2200–2207 (2009)
8. Price, W.N., et al.: Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in *E. coli*. Microbial Informatics and Experimentation **1**(1), 6 (2011)
9. Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., Frishman, D.: PROSO II–a new method for protein solubility prediction. FEBS J. **279**(12), 2192–2200 (2012)

Acknowledgement: This work was supported by ELIXIR CZ research infrastructure project [MEYS Grant No: LM2015047] including access to computing and storage facilities, IT4Innovations excellence in science project of the Ministry of Education, Youth and Sports of the Czech Republic [IT4I XS – LQ1602] and ICT tools, methods and technologies for smart cities project of the Brno University of Technology [FIT-S-17-3964]. Computational resources were supplied by the CESNET [LM2015042] and the CERIT Scientific Cloud [LM2015085], provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures”.