Bioinformatics, 32(13), 2016, 2032–2034
doi: 10.1093/bioinformatics/btw066
Advance Access Publication Date: 19 February 2016
Applications Note

OXFORD

Sequence analysis

PON-Sol: prediction of effects of amino acid substitutions on protein solubility

Yang Yang[1,2,3], Abhishek Niroula[3], Bairong Shen[1] and Mauno Vihinen[3,*]

[1]Center for Systems Biology, [2]School of Computer Science and Technology, Soochow University, Suzhou 215006, China and [3]Department of Experimental Medical Science, Lund University, Lund SE 221 84, Sweden

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on August 25, 2015; revised on January 12, 2016; accepted on January 30, 2016

Abstract

Motivation: Solubility is one of the fundamental protein properties. It is of great interest because of its relevance to protein expression. Reduced solubility and protein aggregation are also associated with many diseases.

Results: We collected from literature the largest experimentally verified solubility affecting amino acid substitution (AAS) dataset and used it to train a predictor called PON-Sol. The predictor can distinguish both solubility decreasing and increasing variants from those not affecting solubility. PON-Sol has normalized correct prediction ratio of 0.491 on cross-validation and 0.432 for independent test set. The performance of the method was compared both to solubility and aggregation predictors and found to be superior. PON-Sol can be used for the prediction of effects of disease-related substitutions, effects on heterologous recombinant protein expression and enhanced crystallizability. One application is to investigate effects of all possible AASs in a protein to aid protein engineering.

Availability and implementation: PON-Sol is freely available at http://structure.bmc.lu.se/PON-Sol. The training and test data are available at http://structure.bmc.lu.se/VariBench/ponsol.php

Contact: mauno.vihinen@med.lu.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Numerous experimental methods are used to investigate variations affecting protein function, structure and properties. In addition, computational approaches are increasingly utilized for predicting effects and mechanisms of variations. Solubility has been of great interest because of its relevance to protein (over)expression. Reduced solubility and protein aggregation are associated with many human diseases. Solubility is very important also for protein structural studies.

Solubility and aggregation are related but different physical concepts. Solubility means concentration in saturated solution where the soluble protein is in equilibrium with solid phase (Arakawa and Timasheff, 1985). Protein in solid phase may be made soluble by dilution. In case of aggregation, protein is in solid phase but this process is typically accompanied by changes to protein structure, integrity and/or modification (Arakawa and Timasheff, 1985). Aggregation is an irreversible process occurring for example in Alzheimer's disease and systemic amyloidoses.

During the last decade computational methods have been developed to predict protein solubility especially for heterologous protein overexpression. These methods utilize different approaches, typically in the field of machine learning.

Much less effort has been put on predicting the effects of variations to protein solubility. It is well known that amino acid substitutions (AASs) can have profound effects on solubility and lead to diseases including severe complex V deficiency (Meulemans *et al.*, 2010) and cataract (Andley and Reilly, 2010). Until now two tools have been released for such predictions OptSolMut (Tian *et al.*, 2010) and CamSol (Sormanni *et al.*, 2014).

2 Methods

Our data includes three kinds of variants: those increasing or decreasing solubility and those that do not affect solubility. As there was no database available, we performed an extensive literature search in PubMed and collected 443 AASs from 71 proteins (Supplementary materials). Leucine, arginine and lysine are the most frequent original residues (Supplementary Table S1). Alanine is the most common substituted residue which is probably because alanine scanning approach is frequently used in solubility experiments. Substitutions to serine are also frequent. Otherwise the variants are widely spread.

The dataset was split into two parts: a training set containing 397 variations and a blind test set of 46 variations. We further split the training dataset into 5 partitions. The partitions were done so that the variations from the same protein and closely related proteins were always kept in the same partition (Supplementary Table S2).

In total, 1080 features based on amino acid sequence were collected to train a machine learning method for solubility prediction, including 617 amino acid features from AA index database (Kawashima and Kanehisa, 2000), 2 conservation features based on SIFT prediction (Ng and Henikoff, 2003), 436 variation type features and 25 neighborhood features (Supplementary materials).

Random forests (RF) algorithm implemented in R package was used for classification. To eliminate the redundant and non-relevant features we used a two step combined greedy feature-selection algorithm as previously described (Niroula *et al.*, 2015) (see Supplementary Fig. S1).

For comprehensive reporting of the binary classifiers, we used a number of measures as previously suggested (Vihinen, 2012, 2013). For 3-class predictors, the correct prediction ratio (CPR) and squared correlation (GC^2) were calculated instead of accuracy and MCC, respectively (Baldi *et al.*, 2000).

3 Results

We designed a 2-layer RF predictor called PON-Sol that classifies the variants into three classes. In the first step, it predicts whether the variants lead to increased solubility or not. Then the variants that do not increase solubility are classified into those that decrease or have no effect on solubility (Fig. 1). Five features were informative for increase/no-increase classifier and eight for decrease/no effect classifier (Supplementary Table S3).

The normalized CPR of PON-Sol is 0.491 and 0.432 in 5-fold cross validation (CV) and blind test, respectively (Supplementary Tables S4 and S6). Note that the result for a random predictor would be 0.333 as there are three classes.

We tested the performance of two previously published methods: CamSol correctly predicted 100 variations out of 392 with normalized CPR 0.353. OptSolMut requires 3-D protein structure. 83 variations in our CV dataset which were not used to train OptSolMut could be mapped to protein 3-D structures. 22 variations were correctly predicted by OptSolMut with normalized CPR 0.280 (Supplementary Table S5).

We tested also some general solubility and aggregation predictors whether they could predict the protein solubility effects due to the substitutions. We compared the predicted solubility scores for the normal and variant sequences (Supplementary Table S5). These results are similar to the dedicated tools OptSolMut and CamSol, and clearly worse than for PON-Sol.

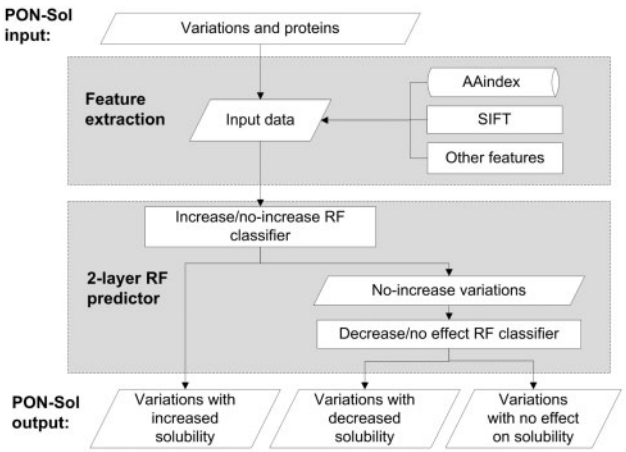

Fig. 1. Overview of PON-Sol's 2-layer random forest predictor. The method combines two two-layer predictors to obtain classification of variants to three groups

In the blind test, PON-Sol also has the best performance among the tested methods (Supplementary Table S6).

4 Web service and application

PON-Sol is freely available as a web service at http://structure.bmc.lu.se/PON-Sol. The user can provide either a list of variations and protein gi numbers or a protein gi for all possible AASs.

PON-Sol could be used for designing variations in protein engineering. As an example we applied PON-Sol to human interleukin-1β (gi: 157831412). We predicted the solubility change for all possible 2907 (153*19) AASs and studied the distribution of solubility increasing and decreasing variants in the 3-dimensional protein structure (Supplementary Fig. S2). The variants that have the highest predicted probability to affect solubility are listed in Supplementary Table S7.

Funding

This work has been supported by the National Nature Science Foundation of China [91230117, 31470821, 31170795], Jiangsu Government Scholarship for oversea studies (JS-2014-185); Swedish Research Council and Barncancerfonden.

Conflict of Interest: none declared.

References

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

Arakawa,T. and Timasheff,S.N. (1985) Theory of protein solubility. *Methods Enzymol.*, **114**, 49–77.

Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.

Meulemans,A. *et al.* (2010) Defining the pathogenesis of the human Atp12p W94R mutation using a *Saccharomyces cerevisiae* yeast model. *J. Biol. Chem.*, **285**, 4099–4109.

Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.

Niroula,A. *et al*. (2015) PON-P2: prediction method for fast and reliable identification of harmful variants. *PLoS One*, **10**, e0117380.

Sormanni,P. *et al*. (2014) The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol*., **427**, 478–490.

Tian,Y. *et al*. (2010) Scoring function to predict solubility mutagenesis. *Algorithms Mol. Biol*., **5**, 33.

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

Vihinen,M. (2013) Guidelines for reporting and using prediction tools for genetic variation analysis. *Hum. Mutat*., **34**, 275–282.