

Application Note

ccSOL omics: a webserver for large-scale prediction of endogenous and heterologous solubility in *E. coli*Federico Agostini^{1,2}, Davide Cirillo^{1,2}, Carmen Maria Livi^{1,2}, Riccardo Delli Ponti^{1,2} and Gian Gaetano Tartaglia^{1,2,*}¹Gene Function and Evolution, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.² Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.

* Corresponding author: Gian Gaetano Tartaglia. Telephone +34 933160116. Email: gian.tartaglia@crg.es.

Associate Editor: Prof. Alfonso Valencia

ABSTRACT**Summary:** Here we introduce *ccSOL omics*, a webserver for large-scale calculations of protein solubility. Our method allows (i) proteome-wide predictions; (ii) identification of soluble fragments within each sequences; (iii) exhaustive single-point mutation analysis.**Results:** Using coil/disorder, hydrophobicity, hydrophilicity, β -sheet and α -helix propensities, we built a predictor of protein solubility. Our approach shows an accuracy of 79% on the training set (36990 Target Track entries). Validation on three independent sets indicates that *ccSOL omics* discriminates soluble and insoluble proteins with an accuracy of 74% on 31760 proteins sharing less than 30% sequence similarity.**Availability:** *ccSOL omics* can be freely accessed on the web at http://s.tartaglialab.com/page/ccsol_group.**Supplementary information:** Documentation and tutorial are available at http://s.tartaglialab.com/static_files/shared/tutorial_ccsol_omics.html.**1 INTRODUCTION**

Algorithms for prediction of protein solubility (Wilkinson and Harrison, 1991) and aggregation (Fernandez-Escamilla *et al.*, 2004) provide a solid basis to investigate physico-chemical determinants of amyloid formation and related diseases (Conchillo-Solé *et al.*, 2007; Tartaglia *et al.*, 2004). In the last years, an *in vitro* reconstituted translation system allowed the large-scale investigation of *E. coli* proteins solubility (Niwa *et al.*, 2009), thus providing the development of predictive methods such as *ccSOL* (Agostini *et al.*, 2012). In *ccSOL*, coil/disorder, hydrophobicity, hydrophilicity, β -sheet and α -helical propensities are combined together into a solubility propensity score that is useful to investigate protein expression (Baig *et al.*, 2014) as well as bacterial evolution (Warnecke, 2012). Other methods have been developed to predict protein solubility based on amino acid characteristics. For instance, PROSO II (Smialowski *et al.*, 2012) exploits occurrence of mono-peptides and dipeptides to estimate heterologous expression in *E. coli*. PROSO II was trained on the pepcDB database [now Target Track (Berman *et al.*, 2009)] that stores target and protocol information provided by Protein Structure Initiative centers. Both *ccSOL* and PROSO II perform accurate predictions when used to respectively predict endogenous or heterologous soluble expressions [*ccSOL*: 76% accuracy; PROSO II: 75% accuracy (Smialowski *et al.*, 2012)]. We found that the experimental status of several Target Track entries (<http://sbkb.org/tt/>) has been recently updated and new data are available to train predictive methods (see

Supplementary information). Here, we introduce a novel implementation of the *ccSOL* method, called *ccSOL omics*, to perform large-scale predictions of endogenous and heterologous expression in *E. coli*. Our algorithm has been trained on non-redundant Target Track entries to identify soluble and insoluble regions within protein sequences. We envisage that *ccSOL omics* will be useful for protein engineering studies, as it allows the investigation of sequence variants in large datasets.

2 WORKFLOW AND IMPLEMENTATION

The *ccSOL omics* server allows the investigation of large protein datasets (see Supplementary information). Once the user provides sequences in FASTA format, the algorithm calculates:

- **Solubility profiles.** To identify soluble fragments within each polypeptide chain, protein sequences are divided into elements and calculate individual solubility propensities. Starting from the N-terminus of a protein, we use a sliding window of 21 amino acids that is moved one residue at a time until the C-terminus is reached. The solubility propensity profile of each fragment is calculated as defined in our original publication (Agostini *et al.*, 2012).
- **Sequence susceptibility.** For each sequence analyzed, the algorithm computes the effect of single amino acid mutations at different positions. This approach is particularly useful to identify regions susceptible to solubility change upon mutation. All variants are reported along with their scores, which provides a basis to engineer protein sequences and test hypotheses such as the occurrence of specific mutations in pathology.
- **Solubility score.** The solubility profile represents a unique *signature* containing information on all fragments arranged in sequential order. In our approach, the profile is used to estimate the solubility upon expression in the *E. coli* system. As sequences have different lengths, we exploit a method based on Fourier's transform (Bellucci *et al.*, 2011; Tartaglia *et al.*, 2007) that allows comparison of polypeptide chains with different sizes. Using 100 Fourier's coefficients, we trained an algorithm that has the same architecture developed for the analysis of protein expression levels in *E. coli* [i.e., neural network approach (Tartaglia *et al.*, 2009)].
- **Reliability score.** The webserver provides a confidence score based on statistical analysis of both training and testing sets (i.e., sequence range employed to validate the method; see Supplementary information).

All the aforementioned analyses are performed for each submitted protein set if the number of entries is below 500. Due to intense CPU usage, sequence susceptibility scores are not computed for datasets larger than 500 entries.

© The Author(s) 2014. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

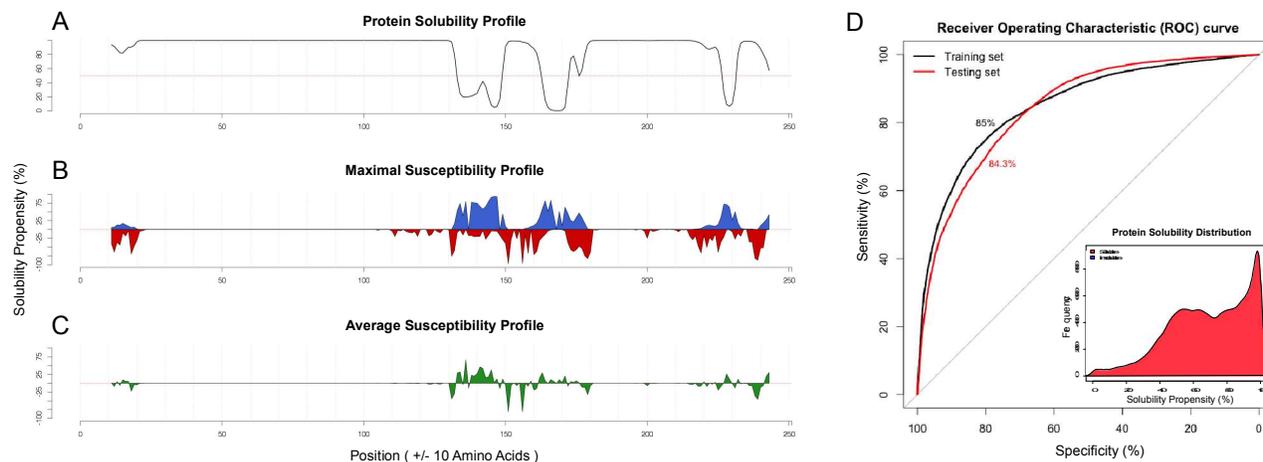


Figure 1. Human Prion Solubility and ccSOL Performances. A) Starting from the N-terminus, ccSOL computes the solubility profile using a sliding window moved towards the C-terminus. ccSOL identifies the fragment 130-170 as the most insoluble within the C-terminus of human PrP. B-C) Maximal and average susceptibility upon single point mutation. D) We trained on the Target Track set (AUROC = 85.5%) and tested on *E. coli* [AUROC = 93.3%; (Niwa *et al.*, 2009)], SOLpro [AUROC = 85.7%; (Magnan *et al.*, 2009)] and PROSO II [AUROC = 82.9%; (Smialowski *et al.*, 2012)] proteins. Inset: overall score distribution for soluble (red) and insoluble (blue) proteins.

3 PERFORMANCES

Expression of human prion (PrP) in *E. coli* is particularly difficult, as the protein accumulates in inactive aggregates (Baneix and Mujacic, 2004). ccSOL *omics* correctly predicts that PrP is insoluble and identifies the fragment 130-170 as the least soluble (Figure 1A-C) together with region 231-253 (not present in the mature form). This finding is very well in agreement with what has been previously reported in literature (Tartaglia *et al.*, 2005, 2008). Moreover, the analysis of susceptible fragments identifies a number of experimentally validated mutations (e.g., G131V, S132I, R148H, V176I, D178N) associated with lower solubility and located in the region promoting PrP aggregation [see Supplementary information (Corsaro *et al.*, 2012)]. As for the large scale performances of ccSOL *omics*, we employed a 10-fold cross validation on Target Track [total of 36990 entries with 30% redundancy (Fu *et al.*, 2012)] and observed 79% accuracy in discriminating between soluble and insoluble proteins. Furthermore, we tested the algorithm on three independent datasets containing protein expression data [total of 31760 entries taken from: *E. coli* (Niwa *et al.*, 2009), SOLpro (Magnan *et al.*, 2009) and PROSO II (Smialowski *et al.*, 2012) and found 74% accuracy (Figure 1D; see also Supplementary information)].

4 CONCLUSIONS

The ccSOL *omics* algorithm shows excellent performances in predicting solubility of endogenous and heterologous genes in *E. coli*. We hope that the webserver will be useful for biotechnological purposes, as it could be for instance employed to design fusion tags for soluble expression. Although accurate, our calculations are based on sequence features and integration with structural characteristics will dramatically increase the predictive power. We plan to combine ccSOL *omics* with information on chaperone (Tartaglia *et al.*, 2010) and RNA (Choi *et al.*, 2009; Bellucci *et al.*, 2011) interactions as these molecules greatly contribute to the solubility of protein products.

ACKNOWLEDGMENTS

We thank A. Zanzoni and G. Bussotti for stimulating discussions.

Funding: The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), through the European Research Council, under grant agreement RIBOMYLOME_309545, and from the Spanish Ministry of Economy and Competitiveness (SAF2011-26211). We also acknowledge support from the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017' (SEV-2012-0208).

REFERENCES

- Agostini, F. *et al.* (2012) Sequence-based prediction of protein solubility. *J. Mol. Biol.*, **421**, 237–241.
- Baig, F. *et al.* (2014) Dynamic transcriptional response of *Escherichia coli* to inclusion body formation. *Biotechnol. Bioeng.*, **111**, 980–999.
- Baneix, F. and Mujacic, M. (2004) Recombinant protein folding and misfolding in *Escherichia coli*. *Nat. Biotechnol.*, **22**, 1399–1408.
- Bellucci, M. *et al.* (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
- Berman, H.M. *et al.* (2009) The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Res.*, **37**, D365–368.
- Choi, S.I. *et al.* (2009) RNA-mediated chaperone type for de novo protein folding. *RNA Biol.*, **6**, 21–24.
- Conchillo-Solé, O. *et al.* (2007) AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65.
- Corsaro, A. *et al.* (2012) Role of prion protein aggregation in neurotoxicity. *Int J Mol Sci.*, **13**, 8648–8669.
- Fernandez-Escamilla, A.-M. *et al.* (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol.*, **22**, 1302–6.
- Fu, L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
- Magnan, C.N. *et al.* (2009) SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*, **25**, 2200–2207.
- Niwa, T. *et al.* (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 4201–4206.
- Smialowski, P. *et al.* (2012) PROSO II – a new method for protein solubility prediction. *FEBS Journal*, **279**, 2192–2200.
- Tartaglia, G.G. *et al.* (2009) A relationship between mRNA expression levels and protein solubility in *E. coli*. *J. Mol. Biol.*, **388**, 381–389.
- Tartaglia, G.G. *et al.* (2010) Physicochemical determinants of chaperone requirements. *J. Mol. Biol.*, **400**, 579–588.
- Tartaglia, G.G. *et al.* (2005) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.*, **14**, 2723–34.
- Tartaglia, G.G. *et al.* (2008) Prediction of aggregation-prone regions in structured proteins. *J Mol Biol*, **380**, 425–36.
- Tartaglia, G.G. *et al.* (2007) Prediction of local structural stabilities of proteins from their amino acid sequences. *Structure*, **15**, 139–43.
- Tartaglia, G.G. *et al.* (2004) The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.*, **13**, 1939–41.
- Warnecke, T. (2012) Loss of the DnaK-DnaJ-GrpE Chaperone System among the Aquificales. *Mol Biol Evol.*, **29**, 3485–3495.
- Wilkinson, D.L. and Harrison, R.G. (1991) Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology (N.Y.)*, **9**, 443–448.