Supplementary Information

SoluProt: Prediction of Soluble Protein Expression in *Escherichia coli*

Jiri Hon[1,2,3], Martin Marusiak[3], Tomas Martinek[3], Antonin Kunka[1,2], Jaroslav Zendulka[3], David Bednar[1,2], Jiri Damborsky[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

Table S1. TargetTrack experiment states signifying soluble expression. The list was compiled by the authors of PROSO II (Smialowski *et al.*, 2012).

Experiment states
soluble, purified, crystallized, hsqc, structure, in pdb, native diffraction-data, NMR assigned, phasing diffraction-data, diffraction, in bmrb, nmr structure, crystal structure, diffraction-quality crystals

Table S2. Specific keywords signifying expression in *E. coli*.

Specific keywords
BL21, DE3, rosetta, xl10, DH10B, CodonPlus, RIPL, RIL, DB3.1, DB3, arctic, origami

Table S3. Protocols identified by generic *E. coli* phrases and manually checked to signify expression in *E.coli*.

Protocol ids
NYSGXRC-SGX_MOLBIO_TOPO_TRANSFORM JCSG-E_Ecoli_GNF_1 CSGID-NU_SelMet_expression CSGID-NU_native_expression MPP-LP.4341 MCSG-NU_default_expression NYSGXRC-SGX_FERM_ECOLI_LB MPP-LP.4813 SSGCID-33 NYSGXRC-SGX_FERM_ECOLI_M9 CSGID-NU_default_expression SSGCID-2 SSGCID-31 SSGCID-1 CESG-MAXWELL 16 EXPRESSION TESTING (R D) v.1.0.0 MPP-LP.4814 SSGCID-128 EFI-SeMET expression in HY Media-PSI2 SGX-SGX_FERM_ECOLI_LB_CFTR SGX-SGX_MOLBIO_EXPR_SOL_CFTR

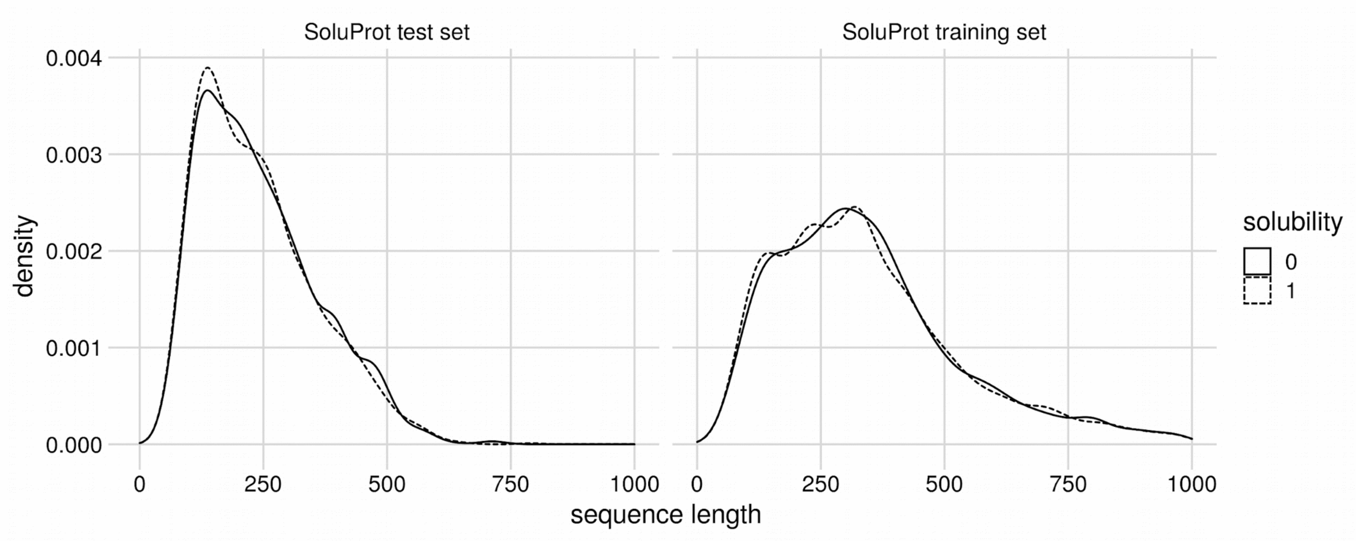

Figure S1. Sequence length distribution of soluble and insoluble proteins in the SoluProt datasets. The x-axis is limited to the range of 0–1000 amino acids to improve readability. The longest sequences in the test and training sets have 790 and 2842 amino acids, respectively.

Table S4. Sequence physicochemical features. Most of the features were extracted using the Biopython package (Cock *et al.*, 2009).

Name	**Description**
physico_chemical_fracnumcharge	Fraction of charged amino acids (R, K, D, E).
physico_chemical_kr_ratio	Ratio of K and R content.
physico_chemical_aa_helix	Fraction of helix amino acids (V, I, Y, F, W, L).
physico_chemical_aa_sheet	Fraction of sheet amino acids (E, M, A, L).
physico_chemical_aa_turn	Fraction of turn amino acids (N, P, G, S).
physico_chemical_molecular_weight	Molecular weight.
physico_chemical_avg_molecular_weight	Molecular weight normalized by the sequence length.
physico_chemical_aromaticity	Fraction of aromatic amino acids (Y, W, F)
physico_chemical_flexibility	Flexibility according to (Vihinen *et al.*, 1994)
physico_chemical_gravy	Grand average of hydropathy according to (Kyte and Doolittle, 1982)
physico_chemical_isoelectric_point	Isoelectric point using methods of Bjellqvist (Bjellqvist *et al.*, 1993, 1994)
physico_chemical_instability_index	Instability index according to (Guruprasad *et al.*, 1990)

Table S5. Sequence features and their importance in the final SoluProt model.

#	Feature	Importance
1	ecoli_usearch_identity_identity	14.11%
2	physico_chemical_isoelectric_point	6.20%
3	monomers_K	3.87%
4	tmhmm_first_60	3.43%
5	monomers_Q	3.31%
6	monomers_E	2.02%
7	monomers_M	1.94%
8	physico_chemical_aa_helix	1.87%
9	dimers_comb_DK	1.77%
10	physico_chemical_molecular_weight	1.56%
11	dimers_comb_EN	1.53%
12	dimers_comb_AA	1.49%
13	monomers_Y	1.39%
14	monomers_C	1.37%
15	dimers_comb_EK	1.25%
16	dimers_comb_AI	1.14%
17	dimers_comb_DT	1.11%
18	dimers_comb_DR	1.09%
19	dimers_comb_RR	1.09%
20	monomers_W	1.07%
21	dimers_comb_IS	1.05%
22	dimers_comb_PQ	1.02%
23	dimers_comb_GK	1.02%
24	dimers_comb_EI	1.01%
25	dimers_comb_DI	0.95%
26	dimers_comb_EE	0.95%
27	dimers_comb_LT	0.93%
28	dimers_comb_EM	0.90%
29	dimers_comb_LL	0.89%
30	dimers_comb_MV	0.89%
31	monomers_F	0.87%
32	dimers_comb_AQ	0.86%
33	dimers_comb_IL	0.85%
34	dimers_comb_LQ	0.85%
35	dimers_comb_GN	0.84%
36	dimers_comb_FP	0.82%
37	dimers_comb_KQ	0.82%
38	dimers_comb_QT	0.80%
39	dimers_comb_GL	0.79%
40	dimers_comb_FT	0.78%
41	dimers_comb_AM	0.78%
42	dimers_comb_TY	0.77%
43	dimers_comb_EV	0.76%
44	dimers_comb_EL	0.75%
45	dimers_comb_EP	0.75%
46	dimers_comb_VY	0.75%
47	dimers_comb_QV	0.72%
48	dimers_comb_LN	0.71%
26	dimers_comb_EE	0.95%
27	dimers_comb_LT	0.93%

#	Feature	Importance
49	dimers_comb_DE	0.71%
50	dimers_comb_SV	0.69%
51	dimers_comb_GG	0.68%
52	dimers_comb_DM	0.67%
53	monomers_H	0.67%
54	physico_chemical_fracnumcharge	0.66%
55	dimers_comb_IT	0.65%
56	dimers_comb_FI	0.65%
57	dimers_comb_AC	0.65%
58	dimers_comb_KV	0.63%
59	dimers_comb_AV	0.63%
60	dimers_comb_CP	0.63%
61	dimers_comb_MN	0.62%
62	dimers_comb_FL	0.62%
63	dimers_comb_RS	0.61%
64	dimers_comb_GH	0.57%
65	dimers_comb_EF	0.55%
66	dimers_comb_AK	0.55%
67	dimers_comb_MW	0.54%
68	dimers_comb_AG	0.54%
69	dimers_comb_NY	0.52%
70	dimers_comb_CI	0.52%
71	dimers_comb_HK	0.51%

#	Feature	Importance
74	dimers_comb_CG	0.49%
75	dimers_comb_KM	0.48%
76	dimers_comb_RW	0.48%
77	dimers_comb_AN	0.47%
78	dimers_comb_HT	0.47%
79	dimers_comb_EH	0.46%
80	dimers_comb_GM	0.46%
81	dimers_comb_CY	0.46%
82	dimers_comb_DW	0.44%
83	dimers_comb_HL	0.43%
84	dimers_comb_IY	0.42%
85	dimers_comb_PW	0.41%
86	dimers_comb_CS	0.39%
87	dimers_comb_KR	0.37%
88	dimers_comb_FM	0.37%
89	dimers_comb_FH	0.32%
90	dimers_comb_GT	0.30%
91	dimers_comb_MY	0.29%
92	dimers_comb_CC	0.27%
93	dimers_comb_HW	0.25%
94	dimers_comb_MM	0.24%
95	dimers_comb_WW	0.12%
96	tmhmm_pred_hel	0.01%

Table S6. Optimized hyperparameters of the Gradient Boosting classifier. In each stage, one or two parameters were optimized while the other parameters were left either at their final values from previous stages or at their default values if they had not been optimized previously. The parameters were first optimized using a large step size. Smaller steps were then used for refinement. The learning rate was lowered from the default value of 0.1 to 0.01 before optimizing the number of estimators. Parameters not mentioned here were left at their default values.

Stage	Parameter	Range	Step	Final value
1	n_estimators	20-100	10	-[a]
2	max_depth	3-17	2, 1	6
	min_samples_split	100-1400	100, 50	1250
3	min_samples_leaf	1-160	10, 5	6
4	max_features	5-96	5	40
5	subsample	0.5-1	1/40	0.525
6	learning_rate	-[b]	-[b]	0.01
7	n_estimators	200-1800	200, 50	1500

[a] The parameter was optimized again in the 7th stage, after which its final value was determined; [b] The learning rate was set to a fixed value; The final set of parameters was as follows: criterion='friedman_mse', init=None, learning_rate=0.01, loss='deviance', max_depth=6, max_features=40, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=6, min_samples_split=1250, min_weight_fraction_leaf=0.0, n_estimators=1500, n_iter_no_change=None, presort='auto', random_state=9, subsample=0.525, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False.

Table S7. Class disagreements between available training sets and the SoluProt test set when applying different binarization thresholds.

Dataset	FP1	FP2	FP3	FP4	FP5	FN1	FN2	FN3	FN4	FN5	E1	E2	E3	E4	E5
PROSO II initial	49	55	188	384	509	508	377	309	204	149	557	432	497	588	658
DeepSol/ SKADE	66	76	183	324	426	356	252	205	129	99	422	328	388	453	525
SWI	43	94	163	256	339	16	11	7	5	2	59	105	170	261	341
SOLpro	39	40	46	83	106	156	132	87	52	35	195	172	133	135	141
SoluProt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FP – false positives, FN – false negatives, E – total number of errors (FP + FN). The numerical suffix denotes the binarization threshold used for the SoluProt test set. For example, a binarization threshold of 2 means that all sequences with solubility scores of 2 or above are considered soluble, and all others are considered insoluble.

References

Bjellqvist,B. *et al.* (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**, 529–539.

Bjellqvist,B. *et al.* (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, **14**, 1023–1031.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Guruprasad,K. *et al.* (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 Eng Des Sel*, **4**, 155–161.

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

Smialowski,P. *et al.* (2012) PROSO II - a new method for protein solubility prediction. *FEBS J.*, **279**, 2192–2200.

Vihinen,M. *et al.* (1994) Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Bioinformatics*, **19**, 141–149.