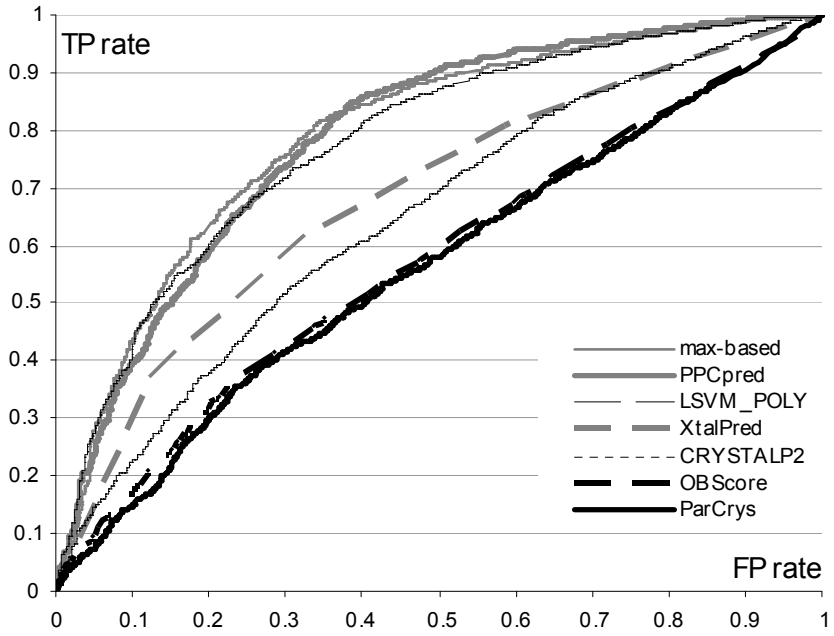
Supplement for the article entitled “In-silico, sequence-based prediction of protein crystallization, purification, and production propensity”

Marcin J. Mizianty[1] and Lukasz Kurgan[1,*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, CANADA



Supplementary Fig 1. The ROC curves for the considered crystallization propensity predictors computed for the DB_CRYS test dataset.

Supplementary Table 1. Number of samples in each step used to create datasets (top of the table) and the sizes of the final datasets (bottom of the table where shading denotes the data aggregated for a given class label). The steps include: 1. Selecting proteins with the completed stop status; 2. Filtering out trials with the same sequence; 3. Filtering out the non-crystallizable proteins against PDB and CDB; 4. Filtering out the non-crystallizable proteins against trials in PepcDB based on their current status field; 5. Selecting trials between 2006 and 2009; 6 Assigning class labels; 7. Removing sequence identity within each class.

Step	Non-crystallizable, with the following failed stop status						Crystallizable
	sequencing	cloning	expression	purification	crystallization	diffraction	
1	508	6 222	11 223	16 457	5 123	6 391	15 412
2	244	3 490	7 252	7 819	4 093	1 283	7 006
3	243	3 470	7 225	7 641	4 087	1 267	6 976
4	240	3 216	7 152	7 462	4 087	1 267	6 976
5	27	764	3 902	4 737	3 135	1 205	6 976
6		4 693		4 737	4 340		4 779
7		2 486		1 431	849		2 408
Datasets	Production of the protein material failed			Purification failed	Crystallization failed		Crystallizable
DB_4CL	2 486			1 431	849		2 408
DB_MF	2 486				4 688		
DB_PF				1 431	3 257		
DB_CF					849		2 408
DB_CRYS	4 766						2 408

Supplementary Table 2. List of considered 64 hydrophobicity- and energy-based indices. The names are based to the nomenclature in the AAIndex1 database.

ARGP820101	BULH740101	CHAM820102	CIDH920105	EISD840101
EISD860101	EISD860102	EISD860103	FAUJ830101	GOLD730101
GUYH850101	HOPT810101	JANJ790102	JOND750101	KYTJ820101
LAWE840101	LEVM760101	MANP780101	MIYS850101	NOZY710101
OOBM770101	OOBM770102	OOBM770103	OOBM770104	OOBM770105
OOBM850103	OOBM850104	PONP800101	PONP800102	PONP800103
PRAM900101	RADA880101	RADA880102	RADA880103	RADA880104
RADA880105	ROBB790101	ROSM880101	ROSM880102	SIMZ760101
SWER830101	VHEG790101	WERD780102	WERD780103	WERD780104
YUTK870101	YUTK870102	YUTK870103	YUTK870104	ZIMJ680101
PONP930101	WILM950101	WILM950102	WILM950103	WILM950104
KUHL950101	JURD980101	WOLR790101	KIDA850101	COWR900101
BLAS910101	CASG920101	ENGD860101	FASG890101	

Supplementary Table 3. Summary of results for the considered SVM models (based on 3 kernel types) and the two methods to aggregate the SVM predictions into the 4-class prediction, *max-based* and *order-based*, that use the SVMs with the highest MCC scores. The results are based on the five-fold cross validation on the training datasets and the methods are sorted in ascending order based on their MCC scores. The highest values for each quality index and dataset are shown in bold.

Training dataset	Method	MCC	ACC	SPEC	SENS	AUC
DB_CRYS	SVM_SIG	0.337	69.3	73.2	61.5	0.730
	SVM_RBF	0.349	73.2	91.0	37.8	0.732
	SVM_POLY	0.373	73.7	**88.2**	45.0	0.766
	order-based	0.455	**76.1**	**83.6**	61.0	0.790
	max-based	**0.456**	75.4	80.3	65.8	**0.795**
DB_MF	SVM_POLY	0.316	69.4	53.4	77.8	0.730
	SVM_SIG	0.339	69.6	59.3	75.1	0.736
	max-based	0.343	71.7	46.5	85.1	0.613
	SVM_RBF	0.425	**74.9**	54.9	**85.5**	**0.779**
	order-based	**0.449**	74.5	**67.6**	78.2	0.742
DB_PF	SVM_SIG	0.197	66.4	42.5	76.9	0.664
	SVM_RBF	0.216	70.4	28.5	88.9	0.677
	max-based	0.217	70.2	30.3	87.7	0.581
	SVM_POLY	0.238	**72.0**	22.9	**93.6**	**0.73**
	order-based	**0.292**	71.4	**43.7**	83.6	0.656
DB_CF	SVM_SIG	0.346	77.3	38.1	91.1	0.781
	SVM_RBF	0.402	78.9	43.3	91.5	0.827
	SVM_POLY	0.440	**80.3**	44.9	**92.8**	**0.843**
	max-based	0.492	78.1	73.2	79.9	0.830
	order-based	**0.499**	78.3	**74.4**	79.7	0.834

Supplementary Table 4. List of the selected features, sorted in the order in which they were added in the feature selection, along with average (over 5 training folds) biserial correlation with the corresponding class labels.

Dataset	Feature name	Biserial correlation	Brief description
DB_MF	WILM950101_min_5	-0.375	Minimal average value of the hydrophobicity index (Wilce et al., 1995) in a window of 5 residues
	AA_exp_E	0.107	Content of the predicted exposed Glu
	DIS_RES_seg_15	-0.198	Content of the predicted disordered residues in segments of 15 or more residues
	KIDA850101_min_5	0.099	Minimal average value of the hydrophobicity index (Kidera et al., 1985) in a window of 5 residues
	WERD780104_min_5	0.088	Minimal average value of the energy index (Wertz and Scheraga, 1978) in a window of 5 residues
	AA_C	-0.185	Composition of Cys
	LAWE840101_max_20	-0.101	Maximal average value of the energy index (Lawson et al., 1984) in a window of 20 residues
	YUTK870103_max_5	0.195	Maximal average value of the energy index (Yutani et al., 1987) in a window of 5 residues
	RSA_REAL	-0.133	Average value of the predicted relative solvent accessibility
	AA_bur_R	0.087	Content of the predicted buried Arg
	OOBM770101_min_15	0.095	Minimal average value of the energy index (Oobatake and Ooi, 1977) in a window of 15 residues
DB_PF	AA_bur_S	-0.198	Content of the predicted buried Ser
	GOLD730101_max_20	-0.129	Maximal average value of the hydrophobicity index (Goldsack and Chalifoux 1973) in a window of 20 residues
	BULH740101_max_10	-0.199	Maximal average value of the energy index (Bull and Breese, 1974) in a window of 10 residues
	ROBB790101_exp	-0.098	Average value of the energy index (Robson and Osguthorpe, 1979) over the predicted exposed residues divided by the length of the sequence
	MANP780101_min_5	0.149	Minimal average value of the hydrophobicity index (Manavalan and Ponnuswamy, 1978) in a window of 5 residues
	AA_burr_C	-0.191	Content of the predicted buried Cys
	pI	-0.181	Isoelectric point
	ROBB790101_min_15	0.118	Minimal value of the energy index (Robson and Osguthorpe, 1979) in a window of 15 residues
	AA_exp_N	-0.092	Content of the predicted exposed Asn
	AA_exp_M	0.094	Content of the predicted exposed Met
DB_CF	GOLD730101_min_10	0.398	Minimal average value of the hydrophobicity index (Goldsack and Chalifoux 1973) in a window of 10 residues
	DIS_SEG	-0.288	Number of the predicted disorder segments
	WILM950104_max_15	-0.201	Maximal average value of the hydrophobicity index (Wilce et al., 1995) in a window of 15 residues
	EXP_RES_seg_5	-0.163	Content of the predicted exposed residues in segments of 5 or more residues
	AA_exp_H	-0.207	Content of the predicted buried His
	EISD860102_min_10	0.151	Minimal average value of the hydrophobicity index (Eisenberg and McLachlan, 1986) in a window of 10 residues
	ROBB790101_min_15	0.241	Minimal value of the energy index (Robson and Osguthorpe, 1979) in a window of 15 residues
	KIDA850101_min_5	0.149	Minimal average value of the hydrophobicity index (Kidera et al., 1985) in a window of 5 residues
	WERD780103_min_5	0.278	Minimal average value of the energy index (Wertz and Scheraga, 1978) in a window of 5 residues
	SWER830101_min_5	0.139	Minimal average value of the hydrophobicity index (Sweet and Eisenberg, 1983) in a window of 5 residues
DB_CRYS	SS_E_avg	0.192	Average length of the predicted strand segments
	HOPT810101_min_10	0.151	Minimal average value of the hydrophobicity index (Hopp and Woods, 1981) in a window of 5 residues
	AA_C	-0.206	Composition of Cys
	DIS_SEG	-0.224	Number of the predicted disorder segments
	GOLD730101_min_10	0.196	Minimal average value of the hydrophobicity index (Goldsack and Chalifoux 1973) in a window of 10 residues
	SIMZ760101_bur	0.123	Average value of the energy index (Simon, 1976) for the predicted buried residues divided by the length of the sequence
	AA_bur_H	0.144	Content of the predicted buried His
	YUTK870103_min_10	0.135	Minimal average value of the energy index (Yutani et al., 1987) in a window of 10 residues
	AA_bur_S	-0.223	Content of the predicted buried Ser
	JURD980101_min_10	0.212	Minimal average value of the hydrophobicity index (Juretic et al., 1998) in a window of 10 residues
	BLAS910101_min_15	0.153	Minimal average value of the hydrophobicity index (Black and Mould, 1991) in a window of 15 residues
	WILM950102_min_10	0.147	Minimal average value of the hydrophobicity index (Wilce et al., 1995) in a window of 10 residues
	RADA880104_min_5	0.166	Minimal average value of the energy index (Radzicka and Wolfenden, 1988) in a window of 5 residues
	RSA_AVG_VAL	-0.175	Average value of the predicted relative solvent accessibility

REFERENCES

Black, S.D. and Mould D.R. (1991) Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *J Analytical Biochem*. 193, 72-82

Bull, H.B. and Breese, K. (1974) Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys*. 161, 665-670

Eisenberg, D. and McLachlan, A.D. (1986) Solvation energy in protein folding and binding. *Nature* 319, 199-203

Goldsack, D.E. and Chalifoux, R.C. (1973) Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure *J. Theor. Biol*. 39, 645-651

Hopp, T.P. and Woods, K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* 78, 3824-3828

Juretic, D., Lucic, B., Zucic, D. and Trinajstic, N. (1998) Protein transmembrane structure: recognition and prediction by using hydrophobicity scales through preference functions. *J Theor. Comp. Chem*. 5, 405-445

Kidera, A., Konishi, Y., Oka, M., Ooi, T. and Scheraga, A. (1985) Statistical Analysis of the Physical Properties of the 20 Naturally Occuring Amino Acids. *J. Prot. Chem*. 4, 23-55

Lawson, E.Q., Sadler, A.J., Harmatz, D., Brandau, D.T., Micanovic, R. MacElroy, R.D. and Middaught, C.R. (1984) A simple experimental model for hydrophobic interactions in proteins. *J. Biol. Chem*. 259, 2910-2912 (1984)

Manavalan, P. and Ponnuswamy, P.K. (1978) Hydrophobic character of amino acid residues in globular proteins. *Nature* 275, 673-674

Oobatake, M. and Ooi, T. (1977) An analysis of non-bonded energy of proteins. *J. Theor. Biol*. 67, 567-584

Radzicka, A. and Wolfenden, R. (1988) Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *J Biochemistry* 27, 1664-1670

Robson, B. and Osguthorpe, D.J. (1979) Refined models for computer simulation of protein folding: Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor. *J. Mol. Biol*. 132, 19-51

Simon, Z. (1976) Quantum Biochemistry and Specific Interactions, Abacus Press, Tunbridge Wells, Kent, England

Sweet, R.M. and Eisenberg, D. (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol*. 171, 479-488

Wertz, D.H. and Scheraga, H.A. (1978) Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules* 11, 9-15

Wilce, M.C., Aguilar, M.I. and Hearn, M.T. (1995) Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides. *Anal Chem*. 67, 1210-1219

Yutani, K., Ogasahara, K., Tsujita, T. and Sugino, Y. (1987) Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit. *Proc. Natl. Acad. Sci. USA* 84, 4441-4444