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

Marcin J. Mizianty¹ and Lukasz Kurgan¹,*
¹Department 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_CRYYS 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.

<table>
<thead>
<tr>
<th>Step</th>
<th>sequencing</th>
<th>cloning</th>
<th>expression</th>
<th>purification</th>
<th>crystallization</th>
<th>diffraction</th>
<th>Crystallizable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>508</td>
<td>6 222</td>
<td>11 223</td>
<td>16 457</td>
<td>5 123</td>
<td>6 391</td>
<td>15 412</td>
</tr>
<tr>
<td>2</td>
<td>244</td>
<td>3 490</td>
<td>7 252</td>
<td>7 819</td>
<td>4 093</td>
<td>1 283</td>
<td>7 006</td>
</tr>
<tr>
<td>3</td>
<td>243</td>
<td>3 470</td>
<td>7 225</td>
<td>7 641</td>
<td>4 087</td>
<td>1 267</td>
<td>6 976</td>
</tr>
<tr>
<td>4</td>
<td>240</td>
<td>3 216</td>
<td>7 152</td>
<td>7 462</td>
<td>4 087</td>
<td>1 267</td>
<td>6 976</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>764</td>
<td>3 902</td>
<td>4 737</td>
<td>3 135</td>
<td>1 205</td>
<td>6 976</td>
</tr>
<tr>
<td>6</td>
<td>4 693</td>
<td>4 737</td>
<td>4 340</td>
<td>4 779</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2 486</td>
<td>1 431</td>
<td>849</td>
<td>2 408</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Datasets | Production of the protein material failed | Purification failed | Crystallization failed | Crystallizable |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DB_4CL</td>
<td>2 486</td>
<td>1 431</td>
<td>849</td>
<td>2 408</td>
</tr>
<tr>
<td>DB_MF</td>
<td>2 486</td>
<td>1 431</td>
<td>4 688</td>
<td></td>
</tr>
<tr>
<td>DB_PF</td>
<td>2 486</td>
<td>1 431</td>
<td>3 257</td>
<td></td>
</tr>
<tr>
<td>DB_CF</td>
<td>2 486</td>
<td>1 431</td>
<td>849</td>
<td></td>
</tr>
<tr>
<td>DB_CRYYS</td>
<td>4 766</td>
<td>1 431</td>
<td>849</td>
<td>2 408</td>
</tr>
</tbody>
</table>
Supplementary Table 2. List of considered 64 hydrophobicity- and energy-based indices. The names are based to the nomenclature in the AAIndex1 database.

<table>
<thead>
<tr>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
<th>Name</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARGP820101</td>
<td>BULH740101</td>
<td>CHAM820102</td>
<td>CIDH920105</td>
<td>EISD840101</td>
<td>EISD860101</td>
<td>EISD860102</td>
<td>EISD860103</td>
<td>FAUJ830101</td>
<td>GOLD730101</td>
<td>GUHY850101</td>
<td>HOPT810101</td>
<td>JANJ790102</td>
<td>JOND750101</td>
<td>KYTE820101</td>
<td>LAWE840101</td>
<td>LEVM760101</td>
<td>MANP780101</td>
<td>MIYS850101</td>
<td>NOZY710101</td>
<td>OOBM770101</td>
<td>OOBM770102</td>
<td>OOBM770103</td>
<td>OOBM770104</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Feature name</th>
<th>Biserial correlation</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB_MF</td>
<td>WILM950101_min_5</td>
<td>-0.375</td>
<td>Minimal average value of the hydrophobicity index (Wilce et al., 1995) in a window of 5 residues</td>
</tr>
<tr>
<td></td>
<td>AA_exp_E</td>
<td>0.107</td>
<td>Content of the predicted exposed Glu</td>
</tr>
<tr>
<td></td>
<td>DIS_RES_seg_15</td>
<td>-0.198</td>
<td>Content of the predicted disordered residues in segments of 15 or more residues</td>
</tr>
<tr>
<td></td>
<td>KIDA850101_min_5</td>
<td>0.099</td>
<td>Minimal average value of the hydrophobicity index (Kidera et al., 1985) in a window of 5 residues</td>
</tr>
<tr>
<td></td>
<td>WERD780104_min_5</td>
<td>0.088</td>
<td>Minimal average value of the energy index (Wertz and Scheraga, 1978) in a window of 5 residues</td>
</tr>
<tr>
<td></td>
<td>AA_C</td>
<td>-0.185</td>
<td>Composition of Cys</td>
</tr>
<tr>
<td></td>
<td>LAWE840101_max_20</td>
<td>-0.101</td>
<td>Maximal average value of the energy index (Lawson et al., 1984) in a window of 20 residues</td>
</tr>
<tr>
<td></td>
<td>YUTK870103_max_5</td>
<td>0.195</td>
<td>Maximal average value of the energy index (Yutani et al., 1987) in a window of 5 residues</td>
</tr>
<tr>
<td></td>
<td>RSA_REAL</td>
<td>-0.133</td>
<td>Average value of the predicted relative solvent accessibility</td>
</tr>
<tr>
<td></td>
<td>AA_bur_R</td>
<td>0.087</td>
<td>Content of the predicted buried Arg</td>
</tr>
<tr>
<td></td>
<td>OOBM770101_min_15</td>
<td>0.095</td>
<td>Minimal average value of the energy index (Oobatake and Ooi, 1977) in a window of 15 residues</td>
</tr>
<tr>
<td></td>
<td>AA_bur_S</td>
<td>-0.198</td>
<td>Content of the predicted buried Ser</td>
</tr>
<tr>
<td></td>
<td>GOLD730101_max_20</td>
<td>-0.129</td>
<td>Maximal average value of the hydrophobicity index (Goldsack and Chalifoux 1973) in a window of 20 residues</td>
</tr>
<tr>
<td></td>
<td>BULLH740101_max_10</td>
<td>-0.199</td>
<td>Maximal average value of the energy index (Bull and Breese, 1974) in a window of 10 residues</td>
</tr>
<tr>
<td></td>
<td>ROBB790101_exp</td>
<td>-0.098</td>
<td>Average value of the energy index (Robson and Osguthorpe, 1979) over the predicted exposed residues divided by the length of the sequence</td>
</tr>
<tr>
<td>DB_PF</td>
<td>MANP780101_min_5</td>
<td>0.149</td>
<td>Minimal average value of the hydrophobicity index (Manavalan and Ponnumswamy, 1978) in a window of 5 residues</td>
</tr>
<tr>
<td></td>
<td>AA_burr_C</td>
<td>-0.191</td>
<td>Content of the predicted buried Cys</td>
</tr>
<tr>
<td></td>
<td>pI</td>
<td>-0.181</td>
<td>Isoelectric point</td>
</tr>
<tr>
<td></td>
<td>ROBB790101_min_15</td>
<td>0.118</td>
<td>Minimal value of the energy index (Robson and Osguthorpe, 1979) in a window of 15 residues</td>
</tr>
<tr>
<td></td>
<td>AA_exp_N</td>
<td>-0.092</td>
<td>Content of the predicted exposed Asn</td>
</tr>
<tr>
<td></td>
<td>AA_exp_M</td>
<td>0.094</td>
<td>Content of the predicted exposed Met</td>
</tr>
<tr>
<td>DB_CF</td>
<td>GOLD730101_min_10</td>
<td>0.398</td>
<td>Minimal average value of the hydrophobicity index (Goldack and Chalifoux 1973) in a window of 10 residues</td>
</tr>
<tr>
<td></td>
<td>DIS_SEG</td>
<td>-0.288</td>
<td>Number of the predicted disorder segments</td>
</tr>
<tr>
<td></td>
<td>WILM950104_max_15</td>
<td>-0.201</td>
<td>Maximal average value of the hydrophobicity index (Wilce et al., 1995) in a window of 15 residues</td>
</tr>
<tr>
<td></td>
<td>EXP_RES_seg_5</td>
<td>-0.163</td>
<td>Content of the predicted exposed residues in segments of 5 or more residues</td>
</tr>
<tr>
<td></td>
<td>AA_exp_H</td>
<td>-0.207</td>
<td>Content of the predicted buried His</td>
</tr>
<tr>
<td></td>
<td>EISD860102_min_10</td>
<td>0.151</td>
<td>Minimal average value of the hydrophobicity index (Eisenberg and McLachlan, 1986) in a window of 10 residues</td>
</tr>
<tr>
<td></td>
<td>ROBB790101_min_15</td>
<td>0.241</td>
<td>Minimal value of the energy index (Robson and Osguthorpe, 1979) in a window of 15 residues</td>
</tr>
<tr>
<td></td>
<td>KIDA850101_min_5</td>
<td>0.149</td>
<td>Minimal average value of the hydrophobicity index (Kidera et al., 1985) in a window of 5 residues</td>
</tr>
<tr>
<td></td>
<td>WERD780103_min_5</td>
<td>0.278</td>
<td>Minimal average value of the energy index (Wertz and Scheraga, 1978) in a window of 5 residues</td>
</tr>
<tr>
<td></td>
<td>SWER830101_min_15</td>
<td>0.139</td>
<td>Minimal average value of the hydrophobicity index (Sweet and Eisenberg, 1983) in a window of 5 residues</td>
</tr>
<tr>
<td>DB_CRYS</td>
<td>SS_E_avg</td>
<td>0.192</td>
<td>Average length of the predicted strand segments</td>
</tr>
<tr>
<td></td>
<td>HOPT810101_min_10</td>
<td>0.151</td>
<td>Minimal average value of the hydrophobicity index (Hopp and Woods, 1981) in a window of 5 residues</td>
</tr>
<tr>
<td></td>
<td>AA_C</td>
<td>-0.206</td>
<td>Composition of Cys</td>
</tr>
<tr>
<td></td>
<td>DIS_SEG</td>
<td>-0.224</td>
<td>Number of the predicted disorder segments</td>
</tr>
<tr>
<td></td>
<td>GOLD730101_min_10</td>
<td>0.196</td>
<td>Minimal average value of the hydrophobicity index (Goldack and Chalifoux 1973) in a window of 10 residues</td>
</tr>
<tr>
<td></td>
<td>SIMZ760101_bur</td>
<td>0.123</td>
<td>Average value of the energy index (Simon, 1976) for the predicted buried residues divided by the length of the sequence</td>
</tr>
<tr>
<td></td>
<td>AA_bur_H</td>
<td>0.144</td>
<td>Content of the predicted buried His</td>
</tr>
<tr>
<td></td>
<td>YUTK870103_min_10</td>
<td>0.135</td>
<td>Minimal average value of the energy index (Yutani et al., 1987) in a window of 10 residues</td>
</tr>
<tr>
<td></td>
<td>AA_bur_S</td>
<td>-0.223</td>
<td>Content of the predicted buried Ser</td>
</tr>
<tr>
<td></td>
<td>JURD980101_min_10</td>
<td>0.212</td>
<td>Minimal average value of the hydrophobicity index (Juretic et al., 1998) in a window of 10 residues</td>
</tr>
<tr>
<td></td>
<td>BLAS910101_min_15</td>
<td>0.153</td>
<td>Minimal average value of the hydrophobicity index (Black and Mould, 1991) in a window of 15 residues</td>
</tr>
<tr>
<td></td>
<td>WILM950102_min_10</td>
<td>0.147</td>
<td>Minimal average value of the hydrophobicity index (Wilce et al., 1995) in a window of 10 residues</td>
</tr>
<tr>
<td></td>
<td>RADA880104_min_5</td>
<td>0.166</td>
<td>Minimal average value of the energy index (Radzicka and Wolfenden, 1988) in a window of 5 residues</td>
</tr>
<tr>
<td></td>
<td>RSA_AVG_VAL</td>
<td>-0.175</td>
<td>Average value of the predicted relative solvent accessibility</td>
</tr>
</tbody>
</table>

### REFERENCES


