Sequence analysis

Protein solubility: sequence based prediction and experimental verification

Pawel Smialowski, Antonio J. Martin-Galiano, Aleksandra Mikolajka¹, Tobias Girschick, Tad A. Holak² and Dmitrij Frishman*

Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany and ¹Max Planck Institute for Biochemistry, 82152 Martinsried, Germany

Received on September 25, 2006; revised on November 22, 2006; accepted on December 4, 2006

Advance Access publication December 6, 2006

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Obtaining soluble proteins in sufficient concentrations is a recurring limiting factor in various experimental studies. Solubility is an individual trait of proteins which, under a given set of experimental conditions, is determined by their amino acid sequence. Accurate theoretical prediction of solubility from sequence is instrumental for setting priorities on targets in large-scale proteomics projects.

Results: We present a machine-learning approach called PROSO to assess the chance of a protein to be soluble upon heterologous expression in Escherichia coli based on its amino acid composition. The classification algorithm is organized as a two-layered structure in which the output of primary support vector machine (SVM) classifiers serves as input for a secondary Naive Bayes classifier. Experimental progress information from the TargetDB database as well as previously published datasets were used as the source of training data. In comparison with previously published methods our classification algorithm possesses improved discriminatory capacity characterized by the Matthews Correlation Coefficient (MCC) of 0.434 between predicted and known solubility states and the overall prediction accuracy of 72% (75 and 68% for positive and negative class, respectively). We also provide experimental verification of our predictions using solubility measurements for 31 mutational variants of two different proteins.

Availability: A web server for protein solubility prediction is available at http://webclu.bio.wzw.tum.de:8080/proso

Contact: d.frishman@wzw.tum.de

Supplementary information: Supplementary data are available at Bioinformatics online

1 INTRODUCTION

Protein solubility is an important pre-requisite for structural and biophysical studies. Obtaining soluble proteins in sufficiently high-concentrations remains a major experimental challenge. Many heterologously expressed proteins are insoluble and often solubilization is a trial and error process with relatively low success rate (Chow et al., 2006; Singh and Panda, 2005). Although highly efficient overexpression into inclusion bodies is sometimes desirable as it results in relatively clean protein, refolding procedures pose a significant technical challenge (Armstrong et al., 1999; Chow et al., 2006) and frequently lead to the loss of biological activity (Singh and Panda, 2005).

Common strategies to improve protein solubility during heterologous expression include weak promoters, low-temperature (Makrides, 1996), modified growth media (Georgiou and Valax, 1996; Makrides, 1996), co-expression with molecular chaperons (Tresaugues et al., 2004), and fusion with solubility enhancing tags (Davis et al., 1999; Kapust and Waugh, 1999). If some knowledge of 3D structure is at hand, it is possible to enhance solubility by structure-guided mutagenesis (Dale et al., 1994). When a priori knowledge of structure is not available protein solubility can be increased by directed evolution method (Pedelaq et al., 2002; Waldo, 2003) that relies on high-throughput screening of large protein diversity libraries generated by random mutagenesis for more soluble variants.

An alternative way to increase the overall success rate of biophysical studies relying on protein solubility is to avoid potentially difficult targets altogether and focus experimental work on those proteins that offer better chances to be soluble. This approach is being frequently practiced by large-scale structural proteomics consortia that initially target more proteins than they can address and then select the most promising candidates (‘low-hanging fruits’) according to certain criteria (Edwards et al., 2000). Despite the fact that the experimental determination of solubility is relatively accessible, the ability to predict from protein sequence its potential for solubility when overexpressed in standard host cells, e.g. Escherichia coli, would be extremely beneficial for rational target selection in structural proteomics as well as for a variety of biophysical studies. Under a given set of experimental conditions, including the expression host, protein solubility is an individual trait ultimately determined by its primary structure.

A number of previous reports addressed the interconnection between protein solubility and various sequence-derived features. A simple method for calculating protein solubility from sequence was first proposed by Wilkinson and Harrison (1991) and then improved by Davis et al. (1999). Their solubility model is based on two parameters: average charge, determined by the relative numbers of Asp, Glu, Lys and Arg residues, and the content of turn-forming residues (Asn, G1y, Pro and Ser). Christendat et al. (2000) examined experimental success and failure data accumulated

¹To whom correspondence should be addressed.
in a high-throughput structural genomics project on 424 nonmembrane proteins from \textit{Methanobacterium thermoautotrophicum}. They demonstrated that insoluble proteins tended to have more hydrophobic stretches (longer than 20 amino acids), lower glutamine content (Q $<$4\%), fewer negatively charged residues (DE $<$17\%) and higher percentage of aromatic amino acids (FYW $>$7.5\%) than soluble ones. A set of simple rules was derived based on these observations allowing prediction of protein solubility with 65\% accuracy. A year later Bertone et al. (2001) reanalyzed 562 proteins from the same organism and confirmed that high-content of negative residues (DE $>$18\%) and absence of hydrophobic patches are associated with improved solubility. Additionally they found that low percentage of aspartic acid, glutamic acid, asparagines and glutamine residues (DENQ $<$16\%) increases the probability of a protein to be insoluble. Goh et al. (2004) applied random forest and decision tree analysis to various attributes of $>$27000 proteins from multiple organisms and found that protein solubility is influenced by (in decreasing order of importance): percentage of serine (S $<$6.4\%), fraction of negatively charged residues (DE $<$10.8\%), percentage of S, C, T and M amino acids, and length ($<$516 amino acids). However, in a study describing high-throughput overexpression in \textit{E.coli} of 10167 proteins of \textit{Caenorhabditis elegans}, no statistically significant correlation between protein’s pl, molecular weight, presence of rare codons, overall sequence hydrophobicity and protein solubility was observed (Luan et al., 2004). The authors indicate that proteins homologous to those with known structures have higher chances to be soluble (Luan et al., 2004). Finally, in a recent publication Idicula-Thomas et al. (2006) presented a solubility prediction method based on a support vector machine (SVM) trained on unbalanced datasets of 62 soluble and 130 insoluble proteins. According to this publication the method was able to predict correctly the increase/decrease in solubility upon mutations. This method is discussed and evaluated in detail below.

Here we present a novel prediction technique dubbed PROSO (PROtein SOLubility predictor) to assess the chance of a protein to be soluble upon heterologous expression in \textit{E.coli}. The method employs SVM and Naive Bayes classifiers trained on experimental progress data stored in the TargetDB (Chen et al., 2004) and PDB (Berman et al., 2000) databases as well as on an additional dataset extracted from literature. In comparison with previously published methods our classification algorithm possesses improved discriminatory capacity characterized by the Matthews Correlation Coefficient (MCC) of 0.434 between predicted and known solubility states and the overall prediction accuracy of 72\% (75 and 68\% for positive and negative class, respectively). We analyze the importance of diverse solubility determinants identified in our work as well as those reported previously. Finally, we provide experimental verification of our predictions using solubility measurements for 31 mutational variants of two different proteins.

2 METHODS

2.1 Datasets of soluble and insoluble proteins

The TargetDB database (Chen et al., 2004, http://targetdb.pdb.org/) stores amino acid sequences and experimental progress information of structural targets pursued by structural genomics consortia. For each protein, TargetDB lists its current experimental status, such as Selected, Cloned, Expressed, Purified, Soluble, Crystallized and so forth. Thus, all proteins that achieved the status Soluble or any subsequent status may be confidently considered soluble. Comparing the experimental status of sequences stored in TargetDB at two sufficiently distant time points – April 2005 and November 2005 – we divided all proteins into TargetDB-Soluble (annotated with the Soluble or any more advanced descriptor) and TargetDB-Insoluble (annotated as Expressed but not as Soluble in April 2005 and still remaining in that state seven months later). We did not consider work-stopped targets (those which were expressed but then either explicitly stated as ‘Work-stopped’ or those that disappeared from TargetDB) since the reason for aborting work on a given target can be unrelated to its experimental behavior. Furthermore, to remove targets dropped as a result of competitors having structure submitted to PDB all proteins with 100\% identity to any PDB protein were removed from the insoluble dataset.

Considering all proteins with known 3D structure soluble by definition, another dataset of soluble proteins (PDB-Soluble) was built from the PDB database (release 1.XII.2005). We selected PDB entries annotated with the descriptors ‘EXPRESSION_SYSTEM: \textit{ESCHERICHIA COLI}’ and ‘EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID’. We also used a further dataset of soluble (Lit-Soluble) and insoluble (Lit-Insoluble) proteins from \textit{E.coli} described by Idicula-Thomas and coworkers (Idicula-Thomas and Balaji, 2005; Idicula-Thomas et al., 2006). Each of the described datasets was refined by removing proteins with one or more transmembrane segments as predicted using TMHMM (Krogh et al., 2001), and those sequences containing more than one contiguous ‘X’ characters. The number of instances and other details on each dataset used in this work can be found in Supplementary Table 1. Finally we merged separately data for soluble and insoluble proteins and removed redundancy by homology clustering at the 50\% sequence identity level using the CD-HIT program (Li et al., 2001; Li et al., 2002).

2.2 Multiple/mono domain split and length distribution adjustment

Large multidomain proteins generally represent more challenging structural targets then smaller single-domain proteins. In particular, the relationship between amino acid sequence and solubility may be significantly different between single- and multidomain proteins since the latter involve sizeable sequence portions corresponding to inter-domain contacts. Multi-domain proteins are believed to fold via a hierarchical multi-step organization of individual co-translationally folded domains (Frydman, 2001; Seckler and Jaenicke, 1992). In contrast, it is common for monodomain proteins to fold in a fast two-state manner (Onuchic and Wolynes, 2004). In order to take into account these differences in the nature of folding/ misfolding we split our datasets into the subsets of long multiple domain and short monodomain proteins.

To find a reasonable length threshold for this split we analyzed size distribution of mono and multiple-domain proteins using domain sequences from the CATH database (Pearl et al., 2005) clustered at 50\% identity level (Supplementary Figure 1). Obviously, no perfect decision rule exists, but we found that threshold length in the range between 250 and 300 residues provides a good separation for dividing our data into two sections – one enriched in monodomain, and the other enriched in multidomain proteins; the threshold of 250 amino acid residues was finally adopted. Solubility classification was performed independently for each dataset.

Since sequence length distributions are somewhat different for insoluble and soluble datasets (data not shown), the composition of sequence datasets was adjusted to account for this effect using a simple approach. Proteins were divided into bins (five bins for mono and nine for multiple domain proteins) according to their length. The width of the bins was 50 and 200 for proteins with length below and above 250, respectively. For each bin, the number of sequences from the least populated dataset was used as a limit of sequence number for both datasets, thus assuring equal population of sequence length ranges and removing any length-related bias. Characterization of the final datasets used for classification can be found in Supplementary Table 2.
In summary, as a result of restrictive data selection from TargetDB and PDB databases we built a sufficiently large (>14,000 proteins) yet reasonably reliable input dataset. However, it is important to realize that the majority of sequences in our final dataset stem from TargetDB, which contains only brief descriptions of experimental results and does not impose standard requirements for annotation. Consequently, in contrast to the PDB-derived data, TargetDB entries lack detailed information about the expression system used for annotation. Consequently, in contrast to the PDB-derived data, sequences in our final dataset stem from TargetDB, which contains only brief reliable input dataset. However, it is important to realize that the majority of PDB databases we built a sufficiently large (2538

2.3 Protein sequence features

Amino acid sequences were represented by the frequencies of mono-, di- and tri-peptides. However, for di- and tri-peptides in the original 20-letter amino acid alphabet the total number of features would be very large – 400 and 8000, respectively. In order to reduce the dimensionality of the feature space and improve signal to noise ratio we clustered amino acids into groups with similar physico-chemical or structural properties as described in our previous work (Smialowski et al., 2006). For the original amino acid alphabet we calculated the frequencies of words of length one and two while for amino acid groups, words of length one, two and three were used.

To derive amino acid clustering pertinent to protein solubility a set of eight numeric scales related to protein solubility was obtained from the Amino Acid Index Database (Kawahine and Kanehsa, 2000): CHAM820102 (free energy of solution in water, kcal/mole) (Charton and Charton, 1982), PONP800101 (surrounding hydrophobicity in folded form) (Ponnuswamy et al., 1980), NOZY710101 (transfer energy and organic solvent/water) (Nozaki and Tanford, 1971), KRIJ790102 (fraction of site occupied by water) (Krigbaum and Komoriya, 1979a, b), ZHOH040103 (buriality) (Zhou and Zhou, 2004), BIOV880101 (information value for accessibility; average fraction 35%) (Bioy et al., 1988), ROSM880102 (side-chain hydrophathy, corrected for solvation) (Roseman, 1988), and JANJ780101 (average accessible surface area) (Janin and Wodak, 1978). Values from all scales were normalized and used together to cluster amino acids by Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Meila and Heckerman, 2001), into two alternative groupings, one with 14 and another one with 17 sets. These cluster numbers have been found optimal in our previous work on protein crystalizability (Smialowski et al., 2006). Amino acids clustering rules are listed in the Supplementary Table 3.

2.4 Classification and feature selection

We classified data using the two level procedure described in detail in our previous publication (Smialowski et al., 2006). Briefly, each set of input data for a fixed word size and a given amino acid grouping scheme was first classified using a primary classifier – a SVM with the Gaussian kernel (Keerthi et al., 2001; Platt, 1999) as described by Fan et al. (2005). The output of the primary classifier for each protein was obtained by 10-fold cross-validation and served as input for a secondary Naive Bayes classifier (Domingos and Pazzani, 1996). A 10-fold stratified cross-validation over input data were performed to obtain class assignment for each protein and to estimate the accuracy of the second level classifier.

For feature selection we used the wrapper method (Kohavi and John, 1997) with the Naive Bayes method (Domingos and Pazzani, 1996) as a classification procedure and the ‘Best first’ approach (Kohavi and John, 1997) as a search algorithm. The detailed procedure can be found in Smialowski et al. (2006). Additionally feature ranking was performed by measuring symmetrical uncertainty of attributes with respect to a given class (Hall and Holmes, 2003). While selecting features the grouping schema which performed best for a given word size was utilized.

2.5 Classifier evaluation

In order to quantify the performance of the classifiers the following measures were calculated:

- accuracy: the number of correctly classified instances divided by the total number of instances
- true positive rate (TP rate) (also called sensitivity or recall of positive class): the number of correctly classified instances from the positive class divided by the number of all instances from the positive class (TP/N + FN)
- true negative rate (TN rate) (also called sensitivity or recall of negative class): the number of correctly classified instances from the negative class divided by the number of all instances from the negative class (TN + FP)
- precision (selectivity): the ratio of the number of correctly classified positive (TP) or negative (TN) instances to the number of all instances classified as positive (TP + FP) or negative (TN + FN), for positive and negative class, respectively
- gain: proportion of given class precision (selectivity) to the ratio of the given class in full dataset
- MCC, calculated as:

\[
\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}
\]

where TP, FP and FN are the numbers of true positive, false positive and false negative instances, respectively.

In particular, gain is an important performance measure that quantifies how much better is the decision when guided by the classifier in comparison to random drawing of instances. MCC indicates correlation between the classifier assignments and the actual class in the two class case. It is a good measure of classifier performance even when classes are unbalanced. Additionally, the statistical relevance of the classification was compared separately for positive and negative classes using the chi square test against the null hypothesis, which states that the classifier has no discriminatory capacity (Motulsy, 1995).

2.6 Evaluation of additional global protein features and previously reported solubility prediction methods

We evaluated all previously published methods, which score protein solubility known to us. For every protein in our dataset we calculated solubility according to Wilkinson-Harrison (Davis et al., 1999; Harrison, 1999, http://www.biotech.ou.edu/). Wilkinson and Harrison, 1991) using their web server (http://www.biotech.ou.edu/) and according to Idicula-Thomas (Idicula-Thomas and Balaji, 2005) using our own implementation of their method. Unfortunately the newer method published in 2006 (Idicula-Thomas et al., 2006) is not available publicly and was not provided to us upon repeated Email requests to the authors. Discriminatory power of both methods was evaluated directly according to the rules specified by the authors. Additionally, the scores produced by these methods were used as input to train and evaluate a Naive Bayes classifier. We also calculated the following global sequence features: sequence length, isoelectric point (pI), grand average of hydropathicity index (GRAVY) (Gasteiger et al., 2003; Kyte and Doolittle, 1982), aliphatic index (AI) (Ikai, 1980) and fold index (FI) (Prilusky et al., 2005). Naive Bayes classifier was trained and evaluated with these features used as input to check whether any of them could result in reasonably good classification performance. Additionally, the combination of AI, FI, GRAVY and pI was also used.
2.7 Protein cloning, expression, purification and solubility measurements

For direct experimental verification of our predictions we utilized two proteins – FOP (FGFR1 oncogene partner, SWISSPROT accession no. O95684) and CAP350 (centrosome associated protein 350, SWISSPROT code Q5VT6) – as well as their multiple sequence variants (constructs) generated by directed mutagenesis. All studied protein constructs were cloned into the pET46 E/kLIC vector and two constructs named FC2SHisMut and FC2LisH_2 had shorter tags spanning 10 and 9 residues, respectively (Supplementary Table 4). To increase the efficiency of protein production, 10 FOP constructs were expressed as high-yielding 1–134 constructs with a factor Xa cleavage site at position 54. N-terminal 1–53 oligopeptides were subsequently removed by factor Xa protease cleavage. Additionally, to facilitate protein concentration measurements we added tryptophan at the position 135 to all FOP constructs spanning 1–134 and 54–134 amino acids. Sequences of all resulting constructs are listed in Supplementary Table 4.

All constructs were expressed in the E. coli strain BL21 STAR (DE3) (Invitrogen) grown in Luria–Bertani (LB) medium supplemented with 100 mg/ml of ampicillin by overnight induction with 0.5 mM isopropyl- β-D-thiogalactopyranoside (IPTG) at 25 °C for CAP350 constructs and 1 mM IPTG at 37 °C for FOP constructs. After expression cells were lysed in a lysis buffer (50 mM NaH2PO4 (pH 8), 300 mM NaCl, 10 mM βME and 10 mM imidazole), sonicated and centrifuged. Supernatants were passed through NiNTA column (Qiagen) in the first step of purification. The samples were loaded onto the column and washed with lysis buffer supplemented with 20 mM imidazole. Elution was performed with the lysis buffer supplemented with 250 mM imidazole. Xa protease cleavage was performed in 20 mM Tris–HCl at pH 7.5 supplemented with 100 mM NaCl and 2 mM CaCl2.

Identities of all the constructs were checked by DNA sequencing. All proteins were checked by western blot, mass spectrometry and/or N-terminal Edman-sequencing. Purity of all proteins was estimated to be 85–98% from SDS–PAGE and mass spectrometry. It was not possible to purify two tested constructs, namely FOP_FL (full-length) and FC3.

Protein solubility was first estimated by measuring relative amount of protein in the cytoplasmatic fraction and in inclusion bodies by SDS–PAGE and mass spectrometry. It was not possible to purify two tested constructs, namely FOP_FL (full-length) and FC3.

3 RESULTS AND DISCUSSION

3.1 Performance of the primary classifier

We classified independently data for mono and multiple domain proteins. Based on MCC values we chose the best classifiers for each amino acid word size and clustering schema. For each word length, the best classifier corresponding to clustering schemes was selected. In general, improvement of accuracy and MCC by increasing word size from one to two is limited. Performance of classifiers based on word size three is lower than the results for shorter word sizes. This effect can be associated with lower number of counts for word size three leading to less efficient problem generalization by SVM. A large difference was observed between classification results for mono and multidomain proteins. All primary classifiers for multidomain proteins had MCC higher than 0.4 and accuracy over 70% while monodomain protein classifiers had MCC not exceeding 0.35 and accuracies below 68%. Detailed results for mono and multiple domain proteins are listed in Supplementary Tables 5 and 6, respectively. All statistical descriptions of the classifiers and classification of all instances were obtained using 10-fold stratified cross-validation over the dataset. In other words, we always tested our classifier on the data which were never part of the training data. It also means that the test sequences did not share sequence similarity >49% with the training data.

3.2 Feature selection for primary classification

The most advantageous feature subsets were selected using the wrapper (Kohavi and John, 1997) method as described in Methods. Primary features selected for the best clustering schema for word size of one, two and three residues, are reported in Supplementary Table 7 separately for mono and multiple domain proteins. It is important to remember that wrapper is designed to establish subsets of features optimal for classification (Kohavi and John, 1997). In doing so, however, it does not provide directly the relations between feature values and solubility. It is interesting that out of eight frequencies of amino acids (R, D, C, E, G, L, M, S) selected by wrapper for both mono and multiple domain proteins four – G, R, D and E – are common with those used as input for the Wilkinson-Harrison method (Davis et al., 1999). Relative content of negatively charged residues (DE) seems to be the strongest determinant of protein solubility as it was selected as an important attribute by all researchers so far (Bertone et al., 2001; Christendat et al., 2000; Davis et al., 1999; Goh et al., 2004; Wilkinson and Harrison, 1991).

Frequencies of cysteine and methionine were also found to be important by Goh and coworkers (Goh et al., 2004). Amongst di-peptide frequencies five of them are common for mono and multiple domain proteins: RE, QA, EG, HM and KG. Notably, the first position in three out of five peptides is occupied by charged (in two cases basic, in one acidic) and in one by polar residues while in the second position in four cases there is a residue with non-polar side chain and only in one case with charged (acidic) side chain. Thus, frequencies of di-peptides with the first residue charged and the second non-polar stand out as important determinants of protein solubility. Interestingly, frequencies of all such peptides were shown to be neutral or even favorable for protein stability (Guruprasad et al., 1990).

Evaluation of symmetrical uncertainty confirmed the importance of these five peptides only partially. EG, HM and KG had scores at least five times larger then the average for each mono and multiple
The PROSO ROC curves are slightly asymmetrical. The shape of the curve reveals the detailed relation between classifier sensitivity (Sensitivity (P)) and indirectly specificity (1 − Specificity (P)) in the entire range of these variables. The diagonal line on the graph symbolizes random classifier. The PROSO ROC curves are slightly asymmetrical.

3.3 Performance of the meta-classifier

Naive Bayes was used as a second-level classifier to aggregate the information from primary SVM classifiers. Input data for the meta-classifier was formed by the classification results for each instance acquired upon 10-fold cross validation of the three best performing primary-classifiers (one for each of the three different word lengths). Combining inputs from different primary classifiers improved the overall separation power of the method and made classifiers less prone to overfitting.

In total our two-level method had an overall accuracy of 72%, with 75% for positive and 68.5% for negative class. The statistical relevance of the results for both classes is very high with $P$-value $<2.2E-16$. The MCC value achieved in this work is 0.434, the highest amongst all tested or previously reported classifiers. For mono and multiple domain proteins the accuracy is 68.0 and 74.6%, respectively. A more detailed characterization of the meta-classifier is given in the last column of Table 1, Supplementary Table 8 and by the receiver operating characteristic (ROC) curves shown in Supplementary Figure 3. The area under ROC curve is equal to 0.781. This last value is a good measurement of classifier performance and is usually interpreted as the probability that a classifier will assign a higher score to the positive example than to the negative, when both are randomly picked. The shape of the curve reveals the detailed relation between classifier sensitivity $\text{Sensitivity (P)} = \text{TPrate}$ and indirectly specificity $\text{1 − Specificity (P)}$ in the entire range of these variables. The diagonal line on the graph symbolizes random classifier. The PROSO ROC curves are slightly asymmetrical.

3.4 A comparative study of solubility predicting methods

We compared the performance of our predictor with other previously reported algorithms: the widely used Wilkinson-Harrison method (Wilkinson and Harrison, 1991), and the recently reported Idicula-Thomas method (Idicula-Thomas and Balaji, 2005). We assessed the results of the Wilkinson-Harrison and Idicula-Thomas approaches using classification thresholds as described by the authors (Harrison, 1999; Idicula-Thomas and Balaji, 2005). Additionally, using the output of these methods and a number of global protein features, such as pl, length, etc. we trained and evaluated (by 10-fold stratified cross-validation) a set of Naive Bayes classifiers as described in Methods. All results obtained on a dataset including 14 200 proteins are summarized in Table 1 and Supplementary Table 8. In this comparison, our method achieved the highest value of MCC 0.434 and the accuracy of 72%. The Wilkinson-Harrison approach resulted in much lower accuracy of 56.2% and the MCC of 0.127. Although the accuracy value reported by Idicula-Thomas (Idicula-Thomas et al., 2006) is 72% and is thus equal to ours, the MCC of this classifier is 18% lower (0.358). Furthermore, the latter classifier was originally tested only on a small set of 64 instances. A chi square test revealed that the results reported by Idicula-Thomas et al. (2006) on the small test dataset including 64 instances. WH: Wilkinson Harrison method. We also evaluated how strongly protein solubility is correlated with simple sequence features: aliphatic index (AI), fold index (FI), GRAVY index (GRAVY), and pI (pI). To optimize the performance a Naive Bayes classifier was trained with output values of a given method or a sequence feature. All results presented were obtained using stratified ten fold cross validation as described in the text. MCC: Matthews Correlation Coefficient, AUROC: Area under receiver operating characteristic curve. The letters P and N in parentheses refer to positive (soluble) and negative (insoluble) class, respectively. PROSO: our method presented in this work.

Table 1. Comparative analysis of solubility prediction methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances</td>
<td>64</td>
<td>64</td>
<td>14 200</td>
<td>14 200</td>
<td>14 200</td>
<td>14 200</td>
<td>14 200</td>
<td>14 200</td>
<td>14 200</td>
<td>14 200</td>
<td>14 200</td>
</tr>
<tr>
<td>Accuracy</td>
<td>72.2</td>
<td>67.2</td>
<td>53.1</td>
<td>54.5</td>
<td>57.6</td>
<td>50.5</td>
<td>58.4</td>
<td>56.0</td>
<td>59.5</td>
<td>60.7</td>
<td>71.7</td>
</tr>
<tr>
<td>MCC</td>
<td>0.358</td>
<td>-</td>
<td>0.078</td>
<td>0.091</td>
<td>0.153</td>
<td>0.011</td>
<td>0.171</td>
<td>0.134</td>
<td>0.201</td>
<td>0.222</td>
<td>0.434</td>
</tr>
<tr>
<td>TPrate</td>
<td>0.55</td>
<td>0</td>
<td>0.233</td>
<td>0.451</td>
<td>0.623</td>
<td>0.589</td>
<td>0.678</td>
<td>0.781</td>
<td>0.753</td>
<td>0.737</td>
<td>0.749</td>
</tr>
<tr>
<td>TNrate</td>
<td>0.803</td>
<td>1</td>
<td>0.829</td>
<td>0.639</td>
<td>0.530</td>
<td>0.422</td>
<td>0.490</td>
<td>0.339</td>
<td>0.438</td>
<td>0.477</td>
<td>0.685</td>
</tr>
<tr>
<td>Gain(P)</td>
<td>1.677</td>
<td>-</td>
<td>1.155</td>
<td>1.111</td>
<td>1.140</td>
<td>1.009</td>
<td>1.141</td>
<td>1.083</td>
<td>1.145</td>
<td>1.170</td>
<td>1.408</td>
</tr>
<tr>
<td>Gain(N)</td>
<td>1.195</td>
<td>1</td>
<td>1.039</td>
<td>1.076</td>
<td>1.169</td>
<td>1.013</td>
<td>1.207</td>
<td>1.215</td>
<td>1.278</td>
<td>1.290</td>
<td>1.463</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.573</td>
<td>0.601</td>
<td>0.565</td>
<td>0.598</td>
<td>0.499</td>
<td>0.611</td>
<td>0.576</td>
<td>0.628</td>
<td>0.650</td>
<td>0.781</td>
<td></td>
</tr>
</tbody>
</table>

*Performance of different methods for solubility prediction. Additionally in the first column we present results reported by Idicula-Thomas et al. (2006) on the small test dataset including 64 instances. WH: Wilkinson Harrison method. We also evaluated how strongly protein solubility is correlated with simple sequence features: aliphatic index (AI), fold index (FI), GRAVY index (GRAVY), and pI (pI). To optimize the performance a Naive Bayes classifier was trained with output values of a given method or a sequence feature. All results presented were obtained using stratified ten fold cross validation as described in the text. MCC: Matthews Correlation Coefficient, AUROC: Area under receiver operating characteristic curve. The letters P and N in parentheses refer to positive (soluble) and negative (insoluble) class, respectively. PROSO: our method presented in this work.
Table 8). The performance reported by Idicula-Thomas and coworkers (accuracy 72%) is due to unbalance in the class representation. Their method is strongly biased toward the negative class with TNrate = 0.8 and TRate = 0.55. On the same data a dummy classifier labeling all proteins as insoluble would have the accuracy of 67%, meaning that the reported classifier outperforms the dummy approach by only 5% points. PROSO surpasses the dummy classifier by over 20 percent points.

Amongst all classifiers based on single global features the one exploiting isoelectrical point (pl) was the most efficient: it reached MCC of 0.2 and accuracy of 59.5%. The combination of all global descriptors was slightly more successful, with MCC and accuracy values of 0.222 and 60.7%, respectively. Despite their relatively high-accuracy, both pl based and combined classifiers suffer from high-unbalance of classification (TPrate (highly soluble) ¼ 0.55. On the same data 10 out of 14 cases (the number of highly soluble proteins) achieving or equal 0.745 to the highly soluble category we were correct in 8 out of 14 cases (the number of highly soluble proteins) achieving TPrate (highly soluble) = 0.71. For medium soluble proteins we were able to classify correctly all instances (PROSO score higher or equal 0.5 and lower then 0.745). Detailed results are provided in Supplementary Tables 9 and 10.

3.6 Conclusions: advances, limitations and future directions

We report a novel sequence-based approach to classify proteins into ‘soluble’ and ‘insoluble’. It is able to categorize sequences with low or no sequence homology to training data. Based on the MCC value of 0.434 and accuracy values of 74.9 and 68.5% for positive and negative class, respectively, we believe that our classifier outperforms any previously reported solubility predictor and is thus expected to be helpful in selecting soluble proteins for biophysical studies as well as in detecting particularly hard cases. Predictive abilities of our method were furthermore confirmed by high-correlation of assigned scores with experimentally determined solubility. We also identified the subset of features, which have the strongest impact on protein solubility.

An obvious limitation of our method is that it is only applicable to non-membrane proteins. It is also unable to take into account factors unrelated to protein sequence, such as buffer composition etc. Instead of the currently used simplistic approach to dividing proteins into single- and multi-domain a more sophisticated prediction technique would be desirable, such as described, e.g. by Jones et al. (2005) or Liu and Rost (2004). Finally, a significant potential for further improvement of our method exists as the annotation of TargetDB proteins gets more accurate and detailed.

ACKNOWLEDGEMENTS

This work was funded by BMBF BIO/0312992A.

Conflict of Interest: none declared.

REFERENCES
