Bioinformatics Advance Access published December 6, 2006

Protein solubility: sequence based prediction and experimental verification

Pawel Smialowski[1], Antonio J. Martin-Galiano[1], Aleksandra Mikolajka[2], Tobias Girschick[1], Tad A. Holak[2] and Dmitrij Frishman[1,*]

[1] Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany

[2] Max Planck Institute for Biochemistry, 82152 Martinsried, Germany

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 *E. 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 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 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 three-dimensional 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 (Pedelacq 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 et al. (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, Gly, Pro, and Ser). Christendat and coworkers (2000) examined experimental success and failure data accumulated in a high-throughput structural genomics project on 424 nonmembrane proteins from *Methanobacterium thermoautotrophicum.* They demonstrated that insoluble proteins tended to have more hydro-

*To whom correspondence should be addressed.

phobic stretches (longer then 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 and coworkers (2004) applied random forest and decision tree analysis to various attributes of more then 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 aa). However, in a study describing high-throughput overexpression in *E. coli* of 10167 proteins of *Caenorhabditis elegans*, no statistically significant correlation between protein's pI, 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 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 *E. coli*. The method employs support vector machine 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 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 previously reported. 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 three-dimensional 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: ESCHERICHIA COLI" and "EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID". We also used a further dataset of soluble (Lit-Soluble) and insoluble (Lit-Insoluble) proteins from *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 Suppl. 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). By contrast, it is common for monodomain proteins to fold in a fast two-state process (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 (Suppl. 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 (5 bins for mono and 9 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 Suppl. Table 2

In summary, as a result of restrictive data selection from TargetDB and PDB databases we built a sufficiently large (more then 14000 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 con-

tains 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 and the setup of expression experiment. Nevertheless *E.coli* is predominantly being used as expression host by structural genomics consortia: out of 3171 PDB entries which originate from TargetDB (as of September 2006) 2381, or 75.1%, had annotation "EXPRESSION_SYSTEM: ESCHERICHIA COLI".

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 (Kawashima and Kanehisa, 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, organic solvent/water) (Nozaki and Tanford, 1971), KRIW790102 (fraction of site occupied by water) (Krigbaum and Komoriya, 1979a; Krigbaum and Komoriya, 1979b), ZHOH040103 (buriability) (Zhou and Zhou, 2004), BIOV880101 (information value for accessibility; average fraction 35%) (Biou et al., 1988), ROSM880102 (side chain hydropathy, 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 wih 14 and another one with 17 sets. These cluster numbers have been found optimal in our previous work on protein crystallizability (Smialowski et al., 2005). Amino acids clustering rules are listed in the Suppl. 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 support vector machine (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). Ten-fold stratified cross-validation over input data was 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+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
- Matthews Correlation Coefficient (MCC), calculated as:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

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 (Motulsky, 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; 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 e-mail 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 number O95684) and CAP350 (centrosome associated protein 350, SWISSPROT code Q5VT06) – as well as their multiple sequence variants (constructs)

generated by directed mutagenesis. All studied protein constructs were cloned into the pET46 Ek/LIC vector (Novagen) according to the manufacturer's manual. Mutations were introduced by the Quick-Change Site Directed Mutagenesis Kit (Stratagene).

All used proteins excluding ten FOP constructs spanning residues 54-134, had the N-terminal His-tag. 19 of them had a 15 amino acid His-tag from the pET46 Ek/Lic vector and two constructs named FC2SHisMut and FC2LisH_2 had shorter tags spanning 10 and 9 residues respectively (Suppl. Table 4). To increase the efficiency of protein production, ten 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 amino acids 1-134 and 54-134. Sequences of all resulting constructs are listed in Suppl. Table 4.

All constructs were expressed in the *E. coli* strain BL21 STAR (DE3) (Invitrogen) grown in LB medium supplemented with 100 mg/ml of ampicillin by overnight induction with 0.5 mM IPTG at 25 ºC for CAP350 constructs and 1 mM IPTG at 37ºC for FOP constructs. After expression cells were lysed in a lysis buffer (50 mM NaH_2PO_4, pH 8, 300 mM NaCl, 10 mM βME, 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 $CaCl_2$.

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, dot-blot and/or Western-blot with appropriate antibodies. Concentration of purified proteins was calculated by measuring absorbance at 280 nm (Layne, 1957; Stoscheck, 1990). Extinction coefficients for all constructs were calculated with the ExPASy ProtParam tool (Gasteiger et al., 2005; Gill and Hippel, 1989). For some of them maximal concentration achievable in the given buffer without causing precipitation was measured after purification as described by Golovanov and coworkers (2004). All concentrations were performed using 10 ml stirring concentrator (Amicon Inc). Oligomerization state of all proteins was determined using the ÄKTA explorer 10 chromatography system (Amersham-Pharmacia) with the following gel filtration chromatography columns: Superdex 75 HR 10/30 or HiLoad 26/60 Superdex 75 prep grade (both Amersham-Pharmacia).

A protein was considered insoluble when it predominantly accumulated in inclusion bodies and resisted standard attempts to refold from denaturizing solution (Singh and Panda, 2005; Tsumoto et al., 2004) using the dilution-refolding method. Stability in solution was assumed when the protein did not precipitate nor degrade (as checked by SDS-PAGE) over a one week period at 4°C.

For all proteins we calculated solubility scores using our classifier. When the protein was classified as a positive (soluble) the class probability was used as a score, in the case of negative it was 1-class probability. This resulted in score values in the range from 0 to 1, with insoluble proteins having values less then 0.5 and soluble greater or equal 0.5.

3 RESULTS AND DISCUSSION

3.1 Performance of the primary classifier

We classified independently data for mono and multiple domain proteins. Based on Matthews Correlation Coefficient (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 Suppl. 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 greater than 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 Materials and Methods. Primary features selected for the best clustering schema for word size of one, two and three residues, are reported in Suppl. 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 2 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 domain dataset, while RE and QA were not found to correlate significantly with protein solubility by this approach (data not shown). Overall, the frequencies of these five di-peptides when considered alone for Naive Bayes classification yielded 56.5% and 62% accuracy for mono and multiple domain datasets, respec-

tively, 10% less then the classification based on frequencies of all 400 peptides.

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, Suppl. Table 8 and by the receiver operating characteristic (ROC) curves shown in Suppl. 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 the 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 the classifier sensitivity (Sensitivity(P)=TPrate) and indirectly specificity (FPrate=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. For the insoluble class the curve goes up steeper in the first half of the chart. This type of ROC is typical for classifiers that can discriminate relatively modest fractions of examples, but do so with high degree of certainty. On the other hand the curve for soluble class grows slower and reaches higher values on the right side of the chart, which is characteristic for classifiers able to retrieve high percentage of instances of the given class. As revealed by the ROC curve, PROSO classified ~70% of positive or negative instances correctly while having only ~25% instances misclassified into the opposite class.

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 pI, length, etc. we trained and evaluated (by 10-fold stratified cross-validation) a set of Naive Bayes classifiers as described in Materials and Methods. All results obtained on a dataset including 14200 proteins are summarized in Table 1 and Suppl. 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) are only slightly statistically significant for the negative class (p-value 0.0065) and not significant for the positive class (p-value 0.988) (Suppl. 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 TPrate=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 only by 5 percentage points. PROSO surpasses the dummy classifier by over 20 percent points.

Amongst all classifiers based on single global features the one exploiting isoelectrical point (pI) 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 pI based and combined classifiers suffer from high unbalance of classification (TPrate ~0.7 and TNrate ~0.4) and relatively limited statistical significance of results for the negative class (p-value > 0.001). It is also worth noting that Naive Bayes based on the combination of all global descriptors was the only method except PROSO to reach both positive and negative gain values of more then 1.15. Amongst dedicated solubility prediction methods used in combination with Naive Bayes, Wilkinson-Harrison clearly outperformed Idicula-Thomas approach on our dataset reaching the accuracy of 57.6% (MCC=0.153) as contrasted with 54.5% (MCC=0.091) of the later method. Our solubility prediction method reaches gain values of 1.41 for soluble and 1.46 for insoluble class, which means that it improves the likelihood of correctly distinguishing between soluble and insoluble proteins by more then 40% relative to random selection (gain=1). It is important to realize that the use of any classifier which shows the gain value 1 or lower is pointless as the equivalent can be achieved by randomly drawing instances from input data. From this comparison we also conclude that the performance of our classifier can not be explained by merely detecting features indirectly linked to pI (accuracy = 60.7%) nor the presence of unfolded regions (accuracy = 55.5%).

3.5 Experimental verification of solubility predictions

We tested our method against experimental data on solubility measured for 31 different constructs of two proteins: FOP (FGFR1 oncogene partner) and CAP350 (centrosome associated protein 350). For two FOP constructs FOP_FL and FC3 we were unable to obtain proteins pure enough for reliable estimation of concentration. For these two constructs protein solubility was quantified as the percentage of soluble fraction in relation to inclusion bodies after expression in *E. coli* and was measured by SDS-PAGE and/or Western-blot (Suppl. Table 9). The rest of the proteins were divided into insoluble, medium soluble and highly soluble based on their estimated maximum concentration, oligomerization state, and stability in solution as described in Materials and Methods. Insoluble proteins were those that could not stay in solution without denaturizing agents like urea or gauanidinium hydrochloride. Highly soluble proteins were defined as those that stayed in solution in

Table 1 Comparative analysis of solubility prediction methods.[a]

Method	Idicula-Thomas, (2006)	Dummy classifier (all insoluble)	Idicula-Thomas (2005)	WH	Naive Bayes							
					Idicula-Thomas (2005)	WH	Length	AI	GRAVY	pI	AI+FI+GRAVY+pI	PROSO
Instances	64	64	14200	14200	14200	14200	14200	14200	14200	14200	14200	14200
Accuracy	72	67.2	53.1	56.2	54.5	57.6	50.5	58.4	56.0	59.5	60.7	71.7
MCC	0.358	-	0.078	0.127	0.091	0.153	0.011	0.171	0.134	0.201	0.222	0.434
TPrate	0.55	0	0.233	0.447	0.451	0.623	0.589	0.678	0.781	0.753	0.737	0.749
TNrate	0.803	1	0.829	0.676	0.639	0.530	0.422	0.490	0.339	0.438	0.477	0.685
Gain(P)	1.677	-	1.155	1.160	1.111	1.140	1.009	1.141	1.083	1.145	1.170	1.408
Gain(N)	1.195	1	1.039	1.100	1.076	1.169	1.013	1.207	1.215	1.278	1.290	1.463
AUROC	-	-	0.573	0.601	0.565	0.598	0.499	0.611	0.576	0.628	0.650	0.781

[a]Performance of different methods for solubility prediction. Additionally in the first column we present results reported by Idicula-Thomas and coworkers (2006) on the small test dataset inclusing 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.

high concentrations, did not oligomerize strongly (percentage of oligomers lower then 50%) and were stable (did not precipitate or aggregate). As an example, Suppl. Figure 2 demonstrates experimental evaluation of solubility for one of the constructs used in this work (FC2SHisMut). Proteins were labeled medium soluble when more than half of the protein in solution was in high oligomeric stage. In all cases our method categorized instances correctly to soluble/insoluble class. Additionally there was a correlation between PROSO score (see Methods) and membership in medium or highly soluble class. By attributing the proteins with PROSO score higher or equal 0.745 to the highly soluble category we were correct in 10 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 Suppl. Table 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, for example, by Jones et al. (2005) or Liu & 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

REFERENCES

Armstrong, N., de Lencastre, A. and Gouaux, E. (1999) A new protein folding screen: application to the ligand binding domains of a glutamate and kainate receptor and to lysozyme and carbonic anhydrase, *Protein Sci*, **8**, 1475-1483.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank, *Nucleic Acids Res*, **28**, 235-242.

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T. and Gerstein, M. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics, *Nucleic Acids Res*, **29**, 2884-2898.

Biou, V., Gibrat, J.F., Levin, J.M., Robson, B. and Garnier, J. (1988) Secondary structure prediction: combination of three different methods, *Protein Eng*, **2**, 185-191.

Charton, M. and Charton, B.I. (1982) The structural dependence of amino acid hydrophobicity parameters, *J Theor Biol*, **99**, 629-644.

Chen, L., Oughtred, R., Berman, H.M. and Westbrook, J. (2004) TargetDB: a target registration database for structural genomics projects, *Bioinformatics*, **20**, 2860-2862.

Chow, M.K., Amin, A.A., Fulton, K.F., Fernando, T., Kamau, L., Batty, C., Louca, M., Ho, S., Whisstock, J.C., Bottomley, S.P. and Buckle, A.M. (2006) The REFOLD database: a tool for the optimization of protein expression and refolding, *Nucleic Acids Res*, **34**, D207-212.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M.,

Edwards, A.M. and Arrowsmith, C.H. (2000) Structural proteomics of an archaeon, *Nat Struct Biol*, **7**, 903-909.

Dale, G.E., Broger, C., Langen, H., D'Arcy, A. and Stuber, D. (1994) Improving protein solubility through rationally designed amino acid replacements: solubilization of the trimethoprim-resistant type S1 dihydrofolate reductase, *Protein Eng*, **7**, 933-939.

Davis, G.D., Elisee, C., Newham, D.M. and Harrison, R.G. (1999) New fusion protein systems designed to give soluble expression in Escherichia coli, *Biotechnol Bioeng*, **65**, 382-388.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likehood from incomplete data via the EM algorithm, *J Roy Statist Soc*, **39**, 1-38.

Domingos, P. and Pazzani, M. (1996) Beyond independence: conditions for the optimality of the simple bayesian classifier. In Saitta, L. (ed), *International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 105-112.

Edwards, A.M., Arrowsmith, C.H., Christendat, D., Dharamsi, A., Friesen, J.D., Greenblatt, J.F. and Vedadi, M. (2000) Protein production: feeding the crystallographers and NMR spectroscopists, *Nat Struct Biol*, **7 Suppl**, 970-972.

Fan, R.E., Chen, P.H. and Lin, C.J. (2005) Working set selection using second order information for training SVM, *Journal of Machine Learning Research* **6**, 1889-1918.

Frydman, J. (2001) Folding of newly translated proteins in vivo: the role of molecular chaperones, *Annu Rev Biochem*, **70**, 603-647.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D. and Bairoch, A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis, *Nucleic Acids Res*, **31**, 3784-3788.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D. and Bairoch, A. (2005) Protein Identification and Analysis Tools on the ExPASy Server. In Walker, J.M. (ed), *The Proteomics Protocols Handbook*. Humana Press, Totowa, New Jersey.

Georgiou, G. and Valax, P. (1996) Expression of correctly folded proteins in Escherichia coli, *Curr Opin Biotechnol*, **7**, 190-197.

Gill, S. and Hippel, P. (1989) Calculation of protein extinction coefficients from amino acid sequence data. , *Anal. Biochem.*, **182**, 319-326.

Goh, C.S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G.T., Zhao, H. and Gerstein, M. (2004) Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis, *J Mol Biol*, **336**, 115-130.

Golovanov, A.P., Hautbergue, G.M., Wilson, S.A. and Lian, L.Y. (2004) A simple method for improving protein solubility and long-term stability, *J Am Chem Soc*, **126**, 8933-8939.

Guruprasad, K., Reddy, B.V. and Pandit, M.W. (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*, **4**, 155-161.

Hall, M.A. and Holmes, G. (2003) Benchmarking Attribute Selection Techniques for Data Mining, *IEEE Transactions on Knowledge and Data Engineering*, **15**, 1437-1447.

Harrison, R.G. (1999) Recombinant Protein Solubility Prediction, http://www.biotech.ou.edu/. University of Oklahoma.

Idicula-Thomas, S. and Balaji, P.V. (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli, *Protein Sci*, **14**, 582-592.

Idicula-Thomas, S., Kulkarni, A.J., Kulkarni, B.D., Jayaraman, V.K. and Balaji, P.V. (2006) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli, *Bioinformatics*, **22**, 278-284.

Ikai, A. (1980) Thermostability and aliphatic index of globular proteins, *J Biochem (Tokyo)*, **88**, 1895-1898.

Janin, J. and Wodak, S. (1978) Conformation of amino acid side-chains in proteins, *J Mol Biol*, **125**, 357-386.

Jones, D.T., Bryson, K., Coleman, A., McGuffin, L.J., Sadowski, M.I., Sodhi, J.S. and Ward, J.J. (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition, *Proteins*, **61 Suppl 7**, 143-151.

Kapust, R.B. and Waugh, D.S. (1999) Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused, *Protein Sci*, **8**, 1668-1674.

Kawashima, S. and Kanehisa, M. (2000) AAindex: amino acid index database, *Nucleic Acids Res*, **28**, 374.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C. and Murthy, K.R.K. (2001) Improvements to Platt's SMO Algorithm for SVM Classifier Design, *Neural Computation*, **13**, 637-649.

Kohavi, R. and John, G. (1997) Wrappers for feature subset selection., *Artificial Intelligence journal*, **97**, 273-324.

Krigbaum, W.R. and Komoriya, A. (1979a) Local interactions as a structure determinant for protein molecules: II, *Biochim Biophys Acta*, **576**, 204-248.

Krigbaum, W.R. and Komoriya, A. (1979b) Local interactions as a structure determinant for protein molecules: III, *Biochim Biophys Acta*, **576**, 229-246.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J Mol Biol*, **305**, 567-580.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein, *J Mol Biol*, **157**, 105-132.

Layne, E. (1957) Spectrophotometric and Turbidimetric Methods for Measuring Proteins., *Methods in Enzymology*, **3**, 447-455.

Li, W., Jaroszewski, L. and Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases, *Bioinformatics*, **17**, 282-283.

Li, W., Jaroszewski, L. and Godzik, A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases, *Bioinformatics*, **18**, 77-82.

Liu, J. and Rost, B. (2004) Sequence-based prediction of protein domains, *Nucleic Acids Res*, **32**, 3522-3530.

Luan, C.H., Qiu, S., Finley, J.B., Carson, M., Gray, R.J., Huang, W., Johnson, D., Tsao, J., Reboul, J., Vaglio, P., Hill, D.E., Vidal, M., Delucas, L.J. and Luo, M. (2004) High-throughput expression of C. elegans proteins, *Genome Res*, **14**, 2102-2110.

Makrides, S.C. (1996) Strategies for achieving high-level expression of genes in Escherichia coli, *Microbiol Rev*, **60**, 512-538.

Meila, M. and Heckerman, D. (2001) An Experimental Comparison of Model-Based Clustering Methods., *Machine Learning*, **42**, 9-29.

Motulsky, H. (1995) *Intuitive Biostatistics*. Oxford University Press, New York.

Nozaki, Y. and Tanford, C. (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale, *J Biol Chem*, **246**, 2211-2217.

Onuchic, J.N. and Wolynes, P.G. (2004) Theory of protein folding, *Curr Opin Struct Biol*, **14**, 70-75.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J. and Orengo, C. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis, *Nucleic Acids Res*, **33**, D247-251.

Pedelacq, J.D., Piltch, E., Liong, E.C., Berendzen, J., Kim, C.Y., Rho, B.S., Park, M.S., Terwilliger, T.C. and Waldo, G.S. (2002) Engineering soluble proteins for structural genomics, *Nat Biotechnol*, **20**, 927-932.

Platt, J. (1999) Fast training of support vector machines using sequential minimal optimization. In Scholkopf, B., Burges, C.J.C. and Smola, A.J. (eds), *Advances in Kernel Methods: Support Vector Learning* MIT Press, Cambridge, MA, 182-208.

Ponnuswamy, P.K., Prabhakaran, M. and Manavalan, P. (1980) Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins, *Biochim Biophys Acta*, **623**, 301-316.

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I. and Sussman, J.L. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded, *Bioinformatics*, **21**, 3435-3438.

Roseman, M.A. (1988) Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds, *J Mol Biol*, **200**, 513-522.

Seckler, R. and Jaenicke, R. (1992) Protein folding and protein refolding, *Faseb J*, **6**, 2545-2552.

Singh, S.M. and Panda, A.K. (2005) Solubilization and refolding of bacterial inclusion body proteins, *J Biosci Bioeng*, **99**, 303-310.

Smialowski, P., Schmidt, T., Cox, J., Kirschner, A. and Frishman, D. (2006) Will my protein crystallize? A sequence-based predictor, *Proteins*, **62**, 343-355.

Stoscheck, C. (1990) Quantitation of Protein, *Methods in Enzymology*, **182**, 50-69.

Tresaugues, L., Collinet, B., Minard, P., Henckes, G., Aufrere, R., Blondeau, K., Liger, D., Zhou, C.Z., Janin, J., Van Tilbeurgh, H. and Quevillon-Cheruel, S. (2004) Refolding strategies from inclusion bodies in a structural genomics project, *J Struct Funct Genomics*, **5**, 195-204.

Tsumoto, K., Umetsu, M., Kumagai, I., Ejima, D., Philo, J.S. and Arakawa, T. (2004) Role of arginine in protein refolding, solubilization, and purification, *Biotechnol Prog*, **20**, 1301-1308.

Waldo, G.S. (2003) Genetic screens and directed evolution for protein solubility, *Curr Opin Chem Biol*, **7**, 33-38.

Wilkinson, D.L. and Harrison, R.G. (1991) Predicting the solubility of recombinant proteins in Escherichia coli, *Biotechnology (N Y)*, **9**, 443-448.

Zhou, H. and Zhou, Y. (2004) Quantifying the effect of burial of amino acid residues on protein stability, *Proteins*, **54**, 315-322.