BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 17 2009, pages 2200–2207
doi:10.1093/bioinformatics/btp386

*Sequence analysis*

# SOLpro: accurate sequence-based prediction of protein solubility

Christophe N. Magnan, Arlo Randall and Pierre Baldi*

Institute for Genomics and Bioinformatics, School of Information and Computer Sciences, University of California, Irvine, CA, USA

Received on April 13, 2009; revised on June 9, 2009; accepted on June 17, 2009

Advance Access publication June 23, 2009

Associate Editor: Burkhard Rost

**ABSTRACT**

**Motivation:** Protein insolubility is a major obstacle for many experimental studies. A sequence-based prediction method able to accurately predict the propensity of a protein to be soluble on overexpression could be used, for instance, to prioritize targets in large-scale proteomics projects and to identify mutations likely to increase the solubility of insoluble proteins.

**Results:** Here, we first curate a large, non-redundant and balanced training set of more than 17 000 proteins. Next, we extract and study 23 groups of features computed directly or predicted (e.g. secondary structure) from the primary sequence. The data and the features are used to train a two-stage support vector machine (SVM) architecture. The resulting predictor, SOLpro, is compared directly with existing methods and shows significant improvement according to standard evaluation metrics, with an overall accuracy of over 74% estimated using multiple runs of 10-fold cross-validation.

**Availability:** SOLpro is integrated in the SCRATCH suite of predictors and is available for download as a standalone application and as a web server at: http://scratch.proteomics.ics.uci.edu.

**Contact:** pfbaldi@ics.uci.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

Producing soluble recombinant proteins on overexpression is a prerequisite for many structural, functional and biochemical studies. *Escherichia coli* is the most commonly used host for the expression of recombinant proteins, because it is relatively fast and inexpensive to manipulate genetically, and it usually results in high protein yields (Idicula-Thomas and Balaji, 2005; Ventura, 2005). However, even in *E.coli* overexpression often yields insoluble proteins. For instance, overexpressed proteins can form misfolded aggregates called *inclusion bodies* (Clark, 1998; Ventura, 2005; Wilkinson and Harrison, 1991). In these cases, solubilization and refolding strategies, such as the use of fusion proteins (Davis *et al*., 1999), co-expression of chaperones (Trésaugues *et al*., 2004), weak promoters (Makrides, 1996) or protein engineering methods (Izard *et al*., 1994; Murby *et al*., 1995), are attempted. However, these difficult and expensive procedures do not ensure solubility (Singh and Panda, 2005). Thus, for many proteomic projects, it would be helpful to have computational methods capable of: (i) identifying proteins that are likely to be problematic from a solubility standpoint; and (ii) guiding protein engineering approaches aimed at addressing solubility issues.

Over the past two decades several research groups have developed methods for predicting protein solubility from the sequence. One of the first studies (Wilkinson and Harrison, 1991) analyzed six sequence-based features (average charge, turn-forming residue fraction, cysteine fraction, proline fraction, hydrophilicity and total number of residues) and the solubility of the corresponding proteins on overexpression. These authors found a strong correlation of solubility with average charge and turn-forming residue fraction and proposed a binary prediction model as well as a ranking model. Since then several other projects have found strong relationships between primary sequence characteristics and solubility on overexpression in *E.coli* (Bertone *et al*., 2001; Christendat *et al*., 2000; Goh *et al*., 2004; Idicula-Thomas and Balaji, 2005; Luan *et al*., 2004). There is a significant overlap among the features identified by these studies, and the differences can be largely attributed to differences in the datasets. These projects provide evidence that primary sequence is the main determinant of solubility, given the same host and same experimental conditions. Furthermore, several studies of the impact of single-site mutations on protein solubility (Malissard and Berger, 2001; Murby *et al*., 1995) provide an additional evidence that primary sequence determines solubility.

The model proposed in Wilkinson and Harrison (1991) was initially evaluated on a set of 81 proteins, with a reported accuracy of 88%. In spite of the prediction models proposed by several authors (Bertone *et al*., 2001; Christendat *et al*., 2000; Goh *et al*., 2004; Idicula-Thomas and Balaji, 2005; Luan *et al*., 2004), the slightly revised Wilkinson–Harrison solubility model (Davis *et al*., 1999) was widely considered to be the most accurate method up to fairly recently (Ahuja *et al*., 2006; Koschorreck *et al*., 2005). However, Idicula-Thomas and Balaji (2005) evaluated the Wilkinson–Harrison model on a small set and reported a prediction accuracy <50%. More recently, Smialowski *et al*. (2007) evaluated the Wilkinson–Harrison model on a balanced and redundancy-reduced set of 14 200 proteins and reported an accuracy of only 56.2%. Similarly, Smialowski *et al*. (2007) evaluated the prediction model proposed in Idicula-Thomas and Balaji (2005) on their set of 14 200 proteins and reported an accuracy of only 53.1%. The accuracy of the model proposed in Idicula-Thomas *et al*. (2006) was estimated at 72%, but using an unbalanced test set containing 67% insoluble proteins. Indeed, issues of small, or unbalanced, or redundant training sets have hampered the field for quite some time.

In part to address these issues, Smialowski *et al*. (2007) prepared a large balanced training set of proteins overexpressed in *E.coli*.

*To whom correspondence should be addressed.

 © The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Using this training set, they built the PROSO predictor and reported an accuracy of 71.7%, achieving significant improvements over the Wilkinson–Harrison solubility model. However, the redundancy of the sequences in this dataset was only reduced at the 50% sequence identity level using the CD-HIT program (Li *et al*., 2001), when a 25% or 30% level is considered necessary to sufficiently reduce the bias introduced by the presence of homologs in the training and test sets (Rost, 1999). The estimated accuracy of 71.7% decreases to 59.3%, when evaluated on a dataset curated with a 25% similarity cutoff (see Section 3.3).

To curate appropriate datasets for training machine learning models and obtain realistic estimates of the performance of these models, the issues of dataset size, redundancy and balance must be dealt with rigorously. Thus, here we begin by curating a large, balanced and non-redundant set of proteins expressed in *E.coli* (**SOLP**). From this dataset, we extract and study several sequence-based feature sets and finally develop a two-stage support vector machine (SVM) architecture to predict the propensity of a protein to be soluble on overexpression in *E.coli*. The resulting method, SOLpro, is evaluated by repeated 10-fold cross-validations, and by direct comparison to previous methods on our training set.

## 2 DATASETS AND METHODS

In this section, we describe the three main methodological steps: (i) the preparation of a rigorous set of soluble and insoluble proteins; (ii) the extraction and analysis of several sequence-based feature sets; and (iii) the derivation of SOLpro, to predict protein solubility. For convenience, the names assigned to datasets always appear in bold.

### 2.1 SOLP: a new set of proteins for solubility prediction

Only a few public databases provide information on the solubility of recombinant proteins. In addition, even when solubility status is provided, the experimental details under which solubility was assessed are often not reported. As a result, it is challenging to extract a consistent dataset, in terms of expression system and experimental conditions (Idicula-Thomas *et al*., 2006). The most widely used expression system is *E.coli*, thus we focus on this system in our data preparation. As in most other solubility studies, we make the reasonable assumption that the experimental conditions, which have become a standard protocol in *E.coli*, are relatively homogeneous.

To prepare our dataset **SOLP**, we: (i) extracted and selected proteins from the PDB, SwissProt and TargetDB databases; (ii) merged these protein subsets with the proteins used in Idicula-Thomas and Balaji (2005); (iii) reduced the redundancy of the sequences with a rigorous threshold (25% similarity); and (iv) balanced the resulting protein set (Sections 2.1.1–2.1.5). Basic statistics on **SOLP** and its subsets are given in Table 1.

*2.1.1 PDB subset of relevant soluble proteins* The Protein Data Bank (Berman *et al*., 2000) contains about 55 000 protein structures. Annotations available for each protein in the database make it possible to identify which proteins were expressed in *E.coli*. We first extracted proteins expressed in *E.coli* (with the PDB annotation 'EXPRESSION_SYSTEM: ESCHERICHIA COLI') using a plasmid vector, the most common vector type (with the annotation 'EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID'). If a protein met these conditions, each one of its chains was included as an independent member of our dataset. The sequence of each chain was extracted using the PDB 'SEQRES' annotation. This protocol resulted in the **PDB-Ecoli** subset, containing almost half of the chains in the PDB database. The final set of PDB relevant proteins, **PDB-RP**, was derived by removing from **PDB-Ecoli** sequences with any of the following characteristics:

- The protein is likely a membrane protein according to annotation or prediction by TMHMM (Krogh *et al*., 2001). Membrane proteins are not soluble on overexpression without particular solubilization strategies (Sanders *et al*., 2004).
- The primary sequence has two or more contiguous unknown amino acids. Some proteins in the PDB contain long stretches of unknown residues and these regions provide no information for sequence based predictors.
- The sequence length is outside the range [10:2000]. The extremely short peptides do not correspond to independently folded proteins. The extremely long sequences, which represent <0.05% of the PDB sequences, are not handled appropriately by the external tools used to predict features. Both the extremely short and extremely long sequences create outliers in the training set that tend to harm the training process.

**Table 1.** **SOLP**: size and composition of the various datasets (**-RP** indicates relevant proteins obtained after filtering)

| Subset | Size | Soluble (%) | Insoluble (%) |
|---|---|---|---|
| **PDB-Ecoli** | 44 450 | 100 | 0 |
| **PDB-RP** | 38 572 | 100 | 0 |
| **SP-enzymes** | 3306 | 100 | 0 |
| **SP-RP** | 3045 | 100 | 0 |
| **TDB-expressed** | 76 503 | 36.84 | 63.16 |
| **TDB-RP** | 70 707 | 37.68 | 62.32 |
| **ITB-train** | 175 | 22.86 | 77.14 |
| **ITB-RP** | 158 | 22.78 | 77.22 |
| **SOLP-redundant** | 112 482 | 60.72 | 39.28 |
| **SOLP-unbalanced** | 19 793 | 43.98 | 56.02 |
| **SOLP** | 17 408 | 50.00 | 50.00 |

*2.1.2 SwissProt subset of E.coli enzymes* The SwissProt database (The UniProt Consortium, 2007) contains numerous *E.coli* proteins. Their solubility in *E.coli* is not systematically documented. However, the *E.coli* enzymes can reasonably be assumed to be soluble in *E.coli*. To ensure a rigorous set of soluble proteins, we limited our selection to SwissProt proteins annotated as: '*E.coli*', 'Enzyme' and 'Reviewed'. Sequences in the resulting set, noted **SP-enzymes**, were then filtered using the same rules as the proteins of the PDB database (Section 2.1.1). We finally obtained a set of 3045 enzymes, called **SP-RP**. Note that some of these proteins are included in the PDB subset **PDB-RP**, and were removed during the redundancy reduction step (Section 2.1.5).

*2.1.3 TargetDB subset of relevant proteins* The PDB organizers created TargetDB (Chen *et al*., 2004) to centralize target sequences and progress status from essentially all of the worldwide structural genomics projects. For each target, the feature 'status' provides the list of achieved preparation steps, such as 'cloned', 'expressed', 'soluble' and 'purified'. The primary shortcoming of the 'status' annotation is that it does not explicitly indicate if a protein is found to be insoluble. Proteins found to be soluble are indicated explicitly with the 'soluble' tag, but a missing annotation does not necessarily mean an insoluble protein.

Another significant shortcoming of the TargetDB annotation is that it does not indicate which expression system was used in the experiments. We know that 77.2% of the 6536 proteins found in the intersection of the TargetDB and

**Table 2.** Amino acids alphabets used to compute frequencies of monomers, dimmers and trimers

| Name | Amino acid groups | Description |
|---|---|---|
| Natural-20 | A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y | Natural amino acid alphabet |
| Hydropho-5 | CFILMVW, NQSTY, DEKR, AG, HP | Grouped by hydrophobicity (Idicula-Thomas *et al*., 2006) |
| ConfSimi-7 | ACEKLMQR, FHWY, ITV, DN, G, P, S | Grouped by conformational similarity (Idicula-Thomas *et al*., 2006) |
| BlosumSM-8 | CILMV, DENQ, FWY, AG, KR, ST, H, P | Grouped according to BLOSUM50 substitution matrix (Idicula-Thomas *et al*., 2006) |
| ClustEM-14 | DEQ, AH, FW, IV, ST, C, G, K, L, M, N, P, R, Y | Grouped from eight numeric scales using EM algorithm (Smialowski *et al*., 2007) |
| ClustEM-17 | DE, IL, NQ, A, C, F, G, H, K, M, P, R, S, T, V, W, Y | Grouped from eight numeric scales using EM algorithm (Smialowski *et al*., 2007) |
| PhysChem-7 | AGILPV, FWY, HKR, CM, DE, NQ, ST | Grouped according to physico-chemical properties: aliphatic, aromatic, positively charged, sulphurated, negatively charged, amide, alcohol. |

Amino acids groups are separated by commas.

PDB databases were expressed in *E.coli*. If we assume that the proportion is similar for the rest of TargetDB, then ∼20% of the proteins in the final dataset **SOLP** were expressed using a host other than *E.coli*. This problem is mitigated by that fact that many of the proteins expressed in other hosts would have similar solubility properties if expressed in *E.coli*. However, the problem cannot be avoided entirely, since the TargetDB is the most important source of data for this prediction problem and the only source of insoluble proteins.

To create a dataset from TargetDB, we started by extracting the subset of proteins annotated as 'Cloned' and 'Expressed' (**TDB-expressed**), the minimum preparation steps in order to observe the solubility, and not annotated as 'Work Stopped' since the reason for stopping work on a target is not reported. Of the remaining proteins, we classified those annotated as 'Soluble' as soluble, and all others as insoluble. Then, we removed the irrelevant or undesirable sequences following the rules described in Section 2.1.1. In addition, we removed proteins classified as insoluble that have an identical match to a PDB chain, finally, resulting in the **TDB-RP** dataset.

*2.1.4 Datasets from previous studies* Most of the datasets used in previous works have a significant overlap with the union of **PDB-RP** and **TDB-RP**. One notable exception is the training set used in Idicula-Thomas and Balaji (2005). We also considered this set of 175 sequences, referred to as **ITB-train**, and filtered it, using the rules described in Section 2.1.1, resulting in **ITB-RP** containing 158 sequences.

*2.1.5 Final training set of proteins* As discussed in Section 1, previous studies were based on datasets that were either too small, or contained a high level of redundancy, or were unbalanced. Evaluation performed on small datasets have low confidence and the corresponding models have poor generalization capabilities. Since we collected large sets of relevant proteins, the dataset size is not an issue here. Evaluation performed on redundant datasets is biased due to the presence of clear homologs in the training and test sets. To overcome the issue of redundancy, we first merged together the sets of relevant proteins **PDB-RP**, **SP-RP**, **TDB-RP** and **ITB-RP** to form **SOLP-redundant**. With 112 482 proteins, **SOLP-redundant** provides a solid basis for generating a large, non-redundant and balanced dataset. The redundancy of **SOLP-redundant** was then reduced using BLASTCLUST (Altschul *et al*., 1997). Two proteins were considered redundant, and thus in the same cluster, when the aligned parts of the sequences have >25% similarity with a 50% minimum sequence length coverage required for at least one of the two sequences. Due to the low similarity cutoff, many clusters contain both soluble and insoluble proteins. We randomly selected one member from each cluster to form the **SOLP-unbalanced** dataset. Using this protocol, 82.4% of the sequences in **SOLP-redundant** were removed. Finally, in order to clarify the interpretation of the experimental results and to make a comparison with previous prediction methods possible, we balanced the soluble and insoluble protein subsets of **SOLP-unbalanced** by randomly removing insoluble proteins until the subsets were equal in size. Our final training set **SOLP** contains 17 408 proteins (8704 soluble and 8704 insoluble). Supplementary Table 1 provides a breakdown of **SOLP** by protein database origin.

## 2.2 Sequence-based features

Strong relationships between several primary sequence characteristics and protein solubility have been previously reported, as described in Section 1. However, no clear consensus has emerged from these studies. In this work, **SOLP** is used to independently evaluate the correlation with solubility of 23 distinct feature sets. The details of these feature sets are provided below. For convenience, the names assigned to each set always appear in square brackets.

*2.2.1 Description of the sequence-based feature sets* The sequences in **SOLP** are described by 23 distinct feature sets, of which 19 were previously described and four are novel. The origin of each feature set is clearly indicated in the description below.

Of the 23 feature sets, 21 are frequencies of amino acid monomers, dimers and trimers using seven different alphabets, including the natural 20 amino acid alphabet and six reduced alphabets described in Table 2. The 21 sets are denoted by `[Name-S:X]`, where `Name` is the name given to the alphabet in Table 2, and `S` is the size of the corresponding alphabet. `X` takes the value M, D or T associated with the frequencies of monomers, dimers and trimers over the corresponding alphabet (e.g. `[Hydropho-5:M]`, `[Hydropho-5:D]` and `[Hydropho-5:T]`). The three sets computed from the alphabet `PhysChem-7` are novel. The two remaining feature sets are described below.

- `[Computed]`: features directly computed from the sequence (Ahuja *et al*., 2006; Idicula-Thomas and Balaji, 2005; Idicula-Thomas *et al*., 2006; Wilkinson and Harrison, 1991) consist of:
  (1) Sequence length $n$.
  (2) Turn-forming residues fraction: $(N+G+P+S)/n$, where, for instance $N$ is the number of asparagine residues in the sequence.
  (3) Absolute charge per residue: $|\frac{R+K-D-E}{n} - 0.03|$.
  (4) Molecular weight.
  (5) GRAVY index, defined as the averaged hydropathy value (Kyte and Doolittle, 1982) of the amino acids in the primary sequence.
  (6) Aliphatic index: $(A+2.9V+3.9I+3.9L)/n$ (Ikai, 1980).
- `[Predicted]`: this novel feature set consists of features predicted from the sequence using the SCRATCH suite of predictors (Cheng *et al*., 2005):
  (1) Beta residues fraction, as predicted by SSpro (Cheng *et al*., 2005).
  (2) Alpha residues fraction, as predicted by SSpro.
  (3) Number of domains, as predicted by DOMpro (Cheng *et al*., 2006).

**Table 3.** Size of the feature sets corresponding to dimer and trimer frequencies before and after the feature selection process

| Feature set | Initial features | Selected features |
|---|---|---|
| [Natural-20:D] | 400 | 13 |
| [Natural-20:T] | 8000 | 24 |
| [Hydropho-5:D] | 25 | 10 |
| [Hydropho-5:T] | 125 | 12 |
| [ConfSimi-7:D] | 49 | 20 |
| [ConfSimi-7:T] | 343 | 15 |
| [BlosumSM-8:D] | 64 | 25 |
| [BlosumSM-8:T] | 512 | 21 |
| [ClustEM-14:D] | 196 | 16 |
| [ClustEM-14:T] | 2744 | 22 |
| [ClustEM-17:D] | 289 | 27 |
| [ClustEM-17:T] | 4913 | 42 |
| [PhysChem-7:D] | 49 | 21 |
| [PhysChem-7:T] | 343 | 12 |

The specific features selected are provided in Supplementary Table 3.

(4) Exposed residues fraction, as predicted by ACCpro (Cheng *et al*., 2005), using a 25% relative solvent accessibility cutoff.

In all the following experiments, each feature is normalized to [−1,+1].

*2.2.2 Feature selection* The sizes of the feature sets, corresponding to the frequencies of dimers and trimers, range from 25 ([Hydropho-5:D]) to 8000 ([Natural-20:T]). Irrelevant or redundant features are well known to affect machine learning algorithms and increase computation time drastically. We use the wrapper method described in Kohavi and John (1997) to find relevant feature subsets, because their heuristic selection method is appropriate for large datasets and feature spaces. We use Naive Bayes as the induction algorithm and a depth-first search as the selection algorithm. The evaluation function to be optimized is the accuracy estimated by 10-fold cross-validation. The selection process is stopped when the standard deviation (SD) of the accuracies computed during the last five steps does not exceed 0.01. The numbers of selected features are shown in Table 3 and the selected features are reported in Supplementary Table 3.

## 2.3 Solubility prediction

The 23 feature sets calculated on proteins in **SOLP** were first studied independently of each other to compare empirically their correlations with solubility. Then, standard techniques for combining several representations of a training dataset into a global prediction method were applied. Ensemble methods (Dietterich, 2000) performed better than single classifiers during these experiments, so we designed a two-stage architecture to predict the propensity of a protein to be soluble on overexpression in *E.coli*. The remainder of this section presents the comparative study of the feature sets as well as the final architecture retained and the corresponding evaluation protocol.

*2.3.1 Comparative study of the feature sets* Each feature set described in Section 2.2 was applied to the 17 408 sequences in **SOLP**. Some of these sequence features have been found to be correlated with solubility in the past, specifically those in [Computed] as well as some of the monomer, dimer and trimer frequencies (Idicula-Thomas and Balaji, 2005; Idicula-Thomas *et al*., 2006; Smialowski *et al*., 2007; Wilkinson and Harrison, 1991). However, because of the small size or redundancy of the training sets used in these studies, and because we propose to include new features, here we perform a comparative analysis of all the feature sets using **SOLP**.

For this comparative analysis, we applied three machine learning algorithms to the data associated with each feature set: *k*-nearest neighbors (*k*NN), neural networks (NN) and SVMs. For *k*NN and NN, we used Weka (Witten and Frank, 2005), and for SVM we used LIBSVM (Chang and Lin, 2001), which implements the sequential minimal optimization (SMO) algorithm proposed in Fan *et al*. (2005). We tuned the hyperparameters of each algorithm to maximize the accuracy computed by 10-fold cross-validation. Results are reported in Table 4. Note that the accuracy estimated by 10-fold cross-validation for each training set is reported only for the most accurate model computed by each algorithm.

**Table 4.** Accuracies of the individual prediction models computed from the individual feature sets

| Feature Set | *k*NN | NN | SVM |
|---|---|---|---|
| Monomer frequencies | | | |
| [Natural-20:M] | 60.99 (2) | 62.63 (1) | **64.39** (1) |
| [ClustEM-17:M] | 60.87 (3) | 61.94 (2) | **64.05** (2) |
| [ClustEM-14:M] | 61.83 (1) | 61.70 (3) | **63.65** (3) |
| [PhysChem-7:M] | 60.01 (4) | 59.12 (7) | **61.89** (4) |
| [BlosumSM-8:M] | 58.15 (14) | 57.13 (12) | **60.20** (10) |
| [ConfSimi-7:M] | 57.17 (17) | 57.11 (13) | **59.63** (12) |
| [Hydropho-5:M] | 56.56 (20) | 55.58 (20) | **58.75** (20) |
| Dimer frequencies | | | |
| [PhysChem-7:D] | 58.67 (10) | 59.40 (4) | **61.60** (6) |
| [ClustEM-14:D] | 59.21 (6) | 58.11 (9) | **60.75** (7) |
| [ClustEM-17:D] | 59.15 (7) | 58.13 (8) | **60.71** (8) |
| [BlosumSM-8:D] | 57.37 (16) | 57.44 (10) | **60.41** (9) |
| [Natural-20:D] | 58.50 (13) | 56.23 (16) | **59.65** (11) |
| [ConfSimi-7:D] | 56.49 (22) | 56.84 (15) | **59.56** (13) |
| [Hydropho-5:D] | 56.52 (21) | 56.91 (14) | **58.86** (19) |
| Trimer frequencies | | | |
| [ClustEM-17:T] | 58.60 (11) | 57.42 (11) | **59.55** (14) |
| [PhysChem-7:T] | 58.54 (12) | 55.72 (18) | **59.25** (16) |
| [Hydropho-5:T] | 57.12 (18) | 55.50 (21) | **58.96** (17) |
| [ConfSimi-7:T] | 57.85 (15) | 55.60 (19) | **58.91** (18) |
| [ClustEM-14:T] | **58.69** (9) | 53.98 (22) | 58.59 (21) |
| [BlosumSM-8:T] | 57.01 (19) | 56.03 (17) | **58.58** (22) |
| [Natural-20:T] | **55.81** (23) | 50.67 (23) | 54.79 (23) |
| Other feature sets | | | |
| [Computed] | 59.57 (5) | 59.38 (5) | **61.67** (5) |
| [Predicted] | 59.12 (8) | 59.17 (6) | **59.38** (15) |

Overall accuracy ranking for each machine learning method is shown in parentheses and the highest accuracy for each feature set is shown in bold.

*2.3.2 Single classifiers* Several techniques to combine the various feature sets and individual features into a global prediction method were applied. The first approach was to build a single classifier that used all, or a subset, of the features considered in the 23 feature sets described in Section 2.2. For this aim, all of the individual features from the 23 feature sets were grouped together into a single large feature set. A variety of feature selection procedures were performed on the large feature set and we experimented with various architectures. The highest accuracy obtained by any single classifier trained on multiple feature sets was 68%, which is only a few percent higher than the best classifier trained on an individual feature set. (Table 4).

*2.3.3 Ensemble classifiers* Ensemble methods aim to improve performance by pooling the predictions made by many individual predictors in some way (Dietterich, 2000). We tested various ensemble methods to evaluate their potential as global classifiers taking as input, for instance, the 23 binary predictions or probability estimates computed by the primary classifiers on each sequence. Most of the ensemble methods obtained an overall accuracy >70% (for brevity's sake detailed results of the various combinations are not reported). Based on the results of

these experiments the probability estimates produced by the 23 first-layer SVMs were selected as the input to the second-stage ensemble predictor, implemented also as an SVM (although NNs performed equally well). The accuracy of the resulting architecture was ∼73%.

Next, we tested each single feature as an additional input to the second-layer classifier and found that the normalized sequence length was the only feature to improve the global performance of the method (by ∼1%), thus this feature was included as an input to the second-layer classifier. Finally, we tested for irrelevant or redundant first-layer classifiers using the feature selection method described in Kohavi and John (1997). To apply their method, we used an SVM for the induction step and a best-first-backward search for the selection step. Three primary classifiers were removed, specifically those computed from the sets `[Hydropho-5:D]`, `[BlosumSM-8:T]` and `[PhysChem-7:T]`. Removing these classifier's results in a simpler architecture with marginally improved performance (+0.3% in accuracy).

*2.3.4 SOLpro: a two-stage SVM-based architecture* The final SOLpro architecture is summarized here. After experimentation and feature selection, 20 primary SVM predictors are retained, associated with 20 different feature sets. The 20 probability estimates produced by the primary predictors and the normalized sequence length make up the 21 final inputs to the second stage SVM combiner. The probability estimate produced by the second stage SVM is the final SOLpro prediction.

*2.3.5 Evaluation and comparison with previous methods* Repeated 10-fold cross-validations were launched, using different randomized balanced splits of **SOLP**, to derive a reliable estimate of the accuracy (Dietterich, 1998; Kohavi, 1995). Note that for each cross-validation experiment, the classifiers of the two layers are computed without using any information about the test proteins. We mention this explicitly because multi-layered methods sometimes use different splits of data for different layers, which can bias the accuracy estimates.

The following standard evaluation criteria were computed: accuracy, precision, recall, Matthews correlation coefficient, area under the receiver operating characteristic (ROC) curve and gain [Specificity(class)/$P$(class)] for each class. When the classes are balanced, as in this work, the gain is equal to two times the specificity. In spite of its redundance, we include it for the sake of direct comparison with previous work. Average values for each metric computed over 10 runs of 10-fold cross-validations (100 values) are given in Table 5.

We also evaluated our own implementation of the revised Wilkinson–Harrison model (Davis *et al*., 1999) on **SOLP**, as well as PROSO (Smialowski *et al*., 2007) by directly applying the PROSO web server to **SOLP**. PROSO was proposed recently and shown to outperform previous methods in a comparative study led by the authors. Results are reported in Table 5 and discussed in Section 3.3.

*2.3.6 Evaluation on homologous proteins* In order to test SOLpro's predictive abilities on clashing protein pairs, i.e. pairs of proteins that are highly homologous with different solubility annotation, we started from **SOLP-redundant** (Table 1) and computed the clusters of homologous proteins defined by a 95% identity cutoff, with minimum sequence length coverage of 95% for both sequences. We only kept those clusters containing both soluble and insoluble proteins (804 clusters). These clusters contained 3075 proteins (1490 soluble and 1585 insoluble), representing 3598 clashing pairs of proteins. The associated probability estimates predicted by SOLpro were then compared for each clashing pair. Results are given and discussed in Section 3.4.

## 3 RESULTS AND DISCUSSION

### 3.1 Features sets and machine learning methods

*3.1.1 Comparative study of the feature sets* It is informative to study the relationships between feature sets and solubility. The 23 feature sets, described in Section 2.2, were analyzed using the protocol described in Section 2.3.1. Results are reported in Table 4 where the feature sets are grouped into four categories: monomer, dimer, trimer frequencies and the two remaining sets (`[Computed]` and `[Predicted]`). It is worth noting that our approach is global, since we studied the correlation of each entire feature set, and not its individual feature components, with solubility. The various feature sets were compared in terms of the prediction accuracy of the models that were trained on them. While it is possible that performances obtained using other machine learning techniques could vary slightly from what we observed, global tendencies can be determined by comparing the results obtained using the three distinct machine learning algorithms ($k$NN, NN and SVM) that were applied to each dataset.

The models trained on monomer frequencies are the most accurate of any models trained on individual features sets. Specifically, the models trained using `[Natural-20:M]`, `[ClustEM-17:M]` and `[ClustEM-14:M]` are the top three overall using all three machine learning algorithms. Also, `[PhysChem-7:M]` is fourth overall using $k$NN and SVM. These results confirm the conclusions given in Idicula-Thomas *et al*. (2006) and Smialowski *et al*. (2007), and emphasize the importance of amino acid first-order statistics.

The models computed from the sets of dimer frequencies `[PhysChem-7:D]`, `[ClustEM-14:D]` and `[ClustEM-17:D]` are always ranked between the 4th and the 10th positions overall, regardless of the machine learning algorithm used. These models tend to be less accurate than those computed from monomer frequencies.

As a group, the trimer models are clearly the least accurate. The highest ranking trimer model, by any algorithm, is that of `[ClustEM-14:T]` which ranks nineth overall using $k$NN. The trimer models from the native alphabet `[Natural-20:T]` have the worst accuracy overall using all three algorithms. This could be partially explained by the drastic reduction in the number of features

**Table 5.** Evaluation of SOLpro, RevWH and PROSO on the **SOLP** dataset, with SOLpro results in bold

| **Method** | PROSO[a] | PROSO[b] | RevWH[c] | SOLpro |
|---|---|---|---|---|
| **Dataset** | –[a] | **SOLP** | **SOLP** | **SOLP** |
| **# Proteins** | *14200* | 16901[d] | 17408 | 17408 |
| Accuracy | *71.70* | 59.28 | 53.75 | **74.15** |
| MCC | *0.434* | 0.184 | 0.076 | **0.487** |
| Recall (soluble) | *0.685* | 0.506 | 0.471 | **0.681** |
| Recall (insoluble) | *0.749* | 0.674 | 0.604 | **0.803** |
| Precision (soluble) | *0.732* | 0.595 | 0.543 | **0.775** |
| Precision (insoluble) | *0.704* | 0.591 | 0.533 | **0.715** |
| Gain (soluble) | *1.463* | 1.225 | 1.087 | **1.550** |
| Gain (insoluble) | *1.408* | 1.150 | 1.066 | **1.431** |
| ROC area (AUC) | *0.781* | ND | ND | **0.742** |

The published results of PROSO on their own dataset are also indicated in the leftmost column and shown in italics.
[a]Method proposed in Smialowski *et al*. (2007), evaluated by the authors on a dataset they prepared. Results are those given in the reference.
[b]Method proposed in Smialowski *et al*. (2007), evaluated on **SOLP** using the web server available at http://mips.gsf.de/proso/proso.seam.
[c]Revised Wilkinson–Harrison solubility model (Davis *et al*., 1999).
[d]The 507 proteins in **SOLP** were rejected by the PROSO web server due to the presence of unknown amino acids in the primary sequence.

in this feature set resulting from the feature selection procedure (from 8000 down to 24 features, see Table 3).

The models computed from the features that were found strongly correlated with solubility in previous studies, grouped in `[Computed]`, are always ranked fifth overall. These results confirm that these features are relevant for this problem.

Surprisingly, predicted properties of the proteins, grouped in the set `[Predicted]`, provide classifiers with performances very close to those obtained from `[Computed]`, despite the noise introduced by prediction errors. It is interesting to note that `[Predicted]` has only four features, the fewest of the 23 feature sets, and still produces accurate models. Except the classifier obtained with the SVM algorithm, these classifiers are almost as accurate as those obtained from `[Computed]`, and are ranked, respectively, sixth and eighth overall for the NNs and *k*NNs methods.

Although the models calculated from some of the feature sets are significantly less accurate than others, we decided to consider all the feature sets in the second part of our work (Sections 2.3.3 and 2.3.4) since ensemble methods are well known to take advantage of weak classifiers.

*3.1.2 Comparison of the machine learning methods* For each feature set, the performance of the models resulting from the *k*NN, NN and SVM machine learning methods are reported in the columns of Table 4. For 21 of the 23 feature sets, the resulting SVM model outperformed the corresponding *k*NN and NN models. The exceptions were the kNN models calculated from the `[ClustEM-14:T]` and `[Natural-20:T]` feature sets, which had only slightly higher accuracies than the corresponding SVM models. Overall, the SVM models were the most accurate, thus, classifiers of the first-layer in our final architecture were trained using SVMs.

## 3.2 Evaluation of SOLpro

The two-stage SVM-based architecture we proposed to predict solubility was evaluated on **SOLP** following the protocol described in Section 2.3.5. Results are reported in the last column of Table 5. The reported evaluation measures are the means of the corresponding 100 values obtained from 10 independently performed 10-fold cross-validation experiments. The SD of the accuracy was 0.044, the SDs of the other measures were also very small and thus, are not reported. The small SDs attest to the stability of the method.

The overall accuracy of SOLpro is 74.15% with a threshold of 0.5. To the best of our knowledge, this is higher than the reported accuracy of any previous method, except the published accuracy of the Wilkinson–Harrison model proposed in 1991 and evaluated on 81 proteins. SOLpro correctly classifies 68.1% of the soluble proteins and 80.3% of the insoluble proteins. This difference results directly from the slight bias of SOLpro towards insoluble-class predictions (43.9% predicted as soluble and 56.1% predicted as insoluble). The precision on the predicted soluble proteins is actually higher (77.5%) than the precision on the predicted insoluble proteins (71.5%). One possible explanation for the slight bias towards predicting proteins as insoluble is the difference in diversity of the soluble and insoluble subsets. Nearly all the insoluble proteins come from the redundant TargetDB database, while the soluble proteins come from diverse sources. Even after performing the rigorous redundancy reduction, as described in Section 2.1.5, the residual redundancy in these sets may still be an issue. This interpretation is supported by the results of our evaluation of PROSO, the prediction model proposed in Smialowski *et al*. (2007), which was trained on a dataset with similar characteristics to **SOLP**, but with redundancy reduction performed using 50% identity as a cutoff. When evaluated on **SOLP**, PROSO predicts 57.2% of the proteins to be insoluble, and its prediction accuracy on soluble proteins is 17% lower than on insoluble proteins. The larger gaps observed in the evaluation of PROSO support the notion that the redundancy of the insoluble proteins plays a role in the bias towards predicting insoluble proteins.

To analyze potential biases related to database origin, we broke down the performance results by database origin (see Supplementary Table 4). The most relevant result is that SOLpro is more accurate on the soluble PDB proteins (82.20%) than on the soluble TargetDB proteins (65.25%). Two important factors must be taken into account when considering this difference. First, the criteria for solubility in PDB are likely to be more stringent than for TargetDB. Second, the weaker 65.25% performance on TargetDB is <5% below the accuracy of SOLpro on soluble proteins. In addition, the higher accuracy of SOLpro on PDB proteins may be a positive characteristic that could be useful for *in silico* screening of protein targets in connection with structural proteomic projects.

The Matthews correlation coefficient of SOLpro is 0.487, which is also higher than the previously reported values. This measure summarizes the confusion matrix in a single value and is considered a relevant indicator of a methods performance. The last measure computed is the area under the ROC curve (AUC). SOLpro achieves an AUC of 0.742.

## 3.3 Comparison with previous methods

Here, we compare SOLpro to the revised Wilkinson–Harrison solubility model (Davis *et al*., 1999) and PROSO (Smialowski *et al*., 2007) because these are two of the most recognized methods in the literature and because PROSO was shown to outperform all other methods. The evaluation is made using **SOLP**, note that the original Wilkinson–Harrison validation set contains only 81 proteins and the dataset from Smialowski *et al*. (2007) is not available. The protocol for comparing the methods is described in Section 2.3.5 and the results are reported in Table 5 and Supplementary Table 4.

The Wilkinson–Harrison solubility model correctly classified 53.75% of the proteins and obtained a MCC of 0.076. These results are consistent with those reported in previous works (Idicula-Thomas and Balaji, 2005; Smialowski *et al*., 2007), indicating that this model does not generalize well.

PROSO correctly classified 59.3% of the proteins in **SOLP** (soluble: 50.6%, insoluble: 67.4%), compared with 74.2% by SOLpro (refer to Table 5). Database growth may partially explain the higher accuracy of SOLpro, since it was prepared using more data. However, PROSO is a recent method and was trained on a dataset likely to have a significant overlap with **SOLP** due to the similar data curation protocol, with the primary difference being the less-rigorous redundancy reduction. In fact, the reported accuracy of PROSO on the Smialowski *et al*. (2007) dataset is 71.7% and it is very likely that the higher degree of redundancy in their dataset contributes to the 12.4% difference between the two evaluations.

In combination, these results demonstrate that the architecture derived in this study is well suited for solubility prediction, given the

complexity of the problem. The significant efforts made to curate a large, non-redundant and balanced dataset and to reduce the potential biases at each step of the derivation provide confidence that SOLpro has good generalization capability. In addition, when compared with the previous methods, SOLpro demonstrates a significant improvement in the prediction of protein solubility.

### 3.4 Evaluation on homologous proteins

Here, we evaluated SOLpro on clashing pairs of highly homologous proteins (Section 2.3.6). It must be noted that predicting the solubility of such clashing pairs is inherently a very challenging problem, and the difficulty is compounded by the fact that there is noise in the data and data annotations. For instance it is likely that some proteins annotated as insoluble are actually soluble (see Section 2.1.3). For 60.12% of the 3598 clashing pairs, the soluble protein is correctly ranked above the insoluble homolog using the predicted probabilities. While not optimal, this performance level is well above random and provides some information about which proteins are more likely to be soluble among very similar sequences. This is significant for protein engineering applications where, for instance, a small number of residues in an insoluble protein are mutated in order to increase its solubility.

## 4 CONCLUSIONS

Prediction of protein solubility is an important but difficult problem. The complexity is compounded by the ambiguous definition of solubility itself, the many sequence independent factors that affect solubility, and the lack of annotation in databases. These reasons explain, in part, why it is challenging to obtain highly accurate classifiers despite the dominant role of the primary sequence in determining the solubility of a protein. In this work, we have presented a new method, SOLpro, for predicting the propensity of a protein to be soluble on overexpression in *E.coli* from the primary sequence. Throughout the development of SOLpro, we have made every effort to address, as much as possible, the issues discussed above.

First, we have focused our work on predicting solubility in *E.coli* and have carefully curated a large, non-redundant and balanced dataset. Second, we have derived a large set of features comprising both previously described features and novel ones. We have studied these features by feature selection and by analysis of correlation between feature sets and performance. The results on the previously reported features are consistent with previous studies showing, for instance, that the frequencies of monomers are particularly relevant features. The results also show that the novel features, such as predicted secondary structure, can improve prediction performance. Third, we have used machine learning methods to leverage the large training set, robustly handle noise and errors in the data, and accommodate the imprecise definition of solubility. After experimenting with various architectures, we converged on a two-tier SVM-based strategy where the primary classifiers, computed from individual feature sets, provide input to a second-layer ensemble classifier to make final predictions. SOLpro produces both solubility probabilities and binary class predictions. Finally, we have evaluated the binary predictions of the method on the dataset we prepared and compared the performances of SOLpro with those of previous methods. The results show that SOLpro is a suitable method for this problem and significantly outperforms previous methods. Proteomic projects where initial target selection is important could take advantage of SOLpro, as could protein engineering projects seeking to alter solubility.

*Conflict of Interest*: none declared.

## REFERENCES


Ahuja,S. *et al.* (2006) Prediction of solubility on recombinant expression of Plasmodium falciparum erythrocyte membrane protein 1 domains in Escherichia coli. *Malaria J.*, **5**, 52.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Berman,H. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bertone,P. *et al.* (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.

Chang,C.C. and Lin,C.J. (2001) LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm (last accessed date July 3, 2009).

Chen,L. *et al.* (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, **20**, 2860–2862.

Cheng,J. *et al.* (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, W72–W76.

Cheng,J. *et al.* (2006) DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. *Data Min. Knowl. Disc.*, **13**, 1–10.

Christendat,D. *et al.* (2000) Structural proteomics of an archaeon. *Nat. Struct. Mol. Biol.*, **7**, 903–909.

Clark,E.D.B. (1998) Refolding of recombinant proteins. *Cur. Opin. Biol.*, **9**, 157–163.

Davis,G.D. *et al.* (1999) New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol. Bioeng.*, **65**, 382–388.

Dietterich,T.G. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, **10**, 1895–1923.

Dietterich,T.G. (2000) Ensemble methods in machine learning. *Lect. Notes Comput. Sci.*, **1857**, 1–15.

Fan,R.E. *et al.* (2005) Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.*, **6**, 1889–1918.

Goh,C.S. *et al.* (2004) Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J. Mol. Biol.*, **336**, 115–130.

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. *et al.* (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.*, **88**, 1895–1898.

Izard,J. *et al.* (1994) A single amino acid substitution can restore the solubility of aggregated colicin A mutants in Escherichia coli. *Protein Eng.*, **7**, 1495–1500.

Kohavi,R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence 1995*. Morgan Kaufmann, San Francisco, USA, pp. 1137–1143.

Kohavi,R. and John,G.H. (1997) Wrappers for feature subset selection. *Artif. Intell.*, **97**, 273–324.

Koschorreck,M. *et al.* (2005) How to find soluble proteins: a comprehensive analysis of alpha/beta hydrolases for recombinant expression in E. coli. *BMC Genomics*, **6**, 49.

Krogh,A. *et al.* (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.

Li,W. *et al.* (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Luan,C.H. *et al.* (2004) High-throughput expression of C. elegans proteins. *Gen. Res.*, **14**, 2102–2110.

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

Malissard,M. and Berger,E.G. (2001) Improving solubility of catalytic domain of human beta-1,4-galactosyltransferase 1 through rationally designed amino acid replacements. *Eur. J. Biochem.*, **268**, 4352–4358.

Murby,M. *et al.* (1995) Hydrophobicity engineering to increase solubility and stability of a recombinant protein from respiratory syncytial virus. *Eur. J. Biochem.*, **230**, 38–44.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Sanders,C.R. *et al.* (2004) French swimwear for membrane proteins. *ChemBioChem*, **5**, 423–426.

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

Smialowski,P. *et al.* (2007) Protein solubility: sequence based prediction and experimental verification. *Bioinformatics*, **23**, 2536–2542.

The UniProt Consortium (2007) The universal protein resource. *Nucleic Acid Res.*, **35**, D193–D197.

Trésaugues,L. *et al.* (2004) Refolding strategies from inclusion bodies in a structural genomics project. *J. Struct. Funct. Genomics*, **5**, 195–204.

Ventura,S. (2005) Sequence determinants of protein aggregation: tools to increase protein solubility. *Microb. Cell Fact.*, **4**, 11.

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

Witten,I.H. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann Series in Data Management Systems, San Francisco, USA.