BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 3 2006, pages 278–284
doi:10.1093/bioinformatics/bti810

Sequence analysis

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*

Susan Idicula-Thomas[1,†], Abhijit J. Kulkarni[2,†], Bhaskar D. Kulkarni[2], Valadi K. Jayaraman[2,*] and Petety V. Balaji[1,*]

[1]School of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India and [2]Chemical Engineering and Process Development Division, National Chemical Laboratory, Dr Homi Bhabha Road, Pune 411 008, India

Received on April 30, 2005; revised on November 26, 2005; accepted on December 1, 2005
Advance Access publication December 6, 2005
Associate Editor: Charlie Hodgman

ABSTRACT

Motivation: Inclusion body formation has been a major deterrent for overexpression studies since a large number of proteins form insoluble inclusion bodies when overexpressed in *Escherichia coli*. The formation of inclusion bodies is known to be an outcome of improper protein folding; thus the composition and arrangement of amino acids in the proteins would be a major influencing factor in deciding its aggregation propensity. There is a significant need for a prediction algorithm that would enable the rational identification of both mutants and also the ideal protein candidates for mutations that would confer higher solubility-on-overexpression instead of the presently used trial-and-error procedures.

Results: Six physicochemical properties together with residue and dipeptide-compositions have been used to develop a support vector machine-based classifier to predict the overexpression status in *E.coli*. The prediction accuracy is ~72% suggesting that it performs reasonably well in predicting the propensity of a protein to be soluble or to form inclusion bodies. The algorithm could also correctly predict the change in solubility for most of the point mutations reported in literature. This algorithm can be a useful tool in screening protein libraries to identify soluble variants of proteins.

Avalibility: Software is available on request from the authors.

Contact: balaji@iitcb.ac.in; vk.jayaraman@ncl.res.in

Supplementary information: Supplementary data are available at *Bioinformatics* Online web site.

INTRODUCTION

Only some proteins are soluble upon overexpression in *Escherichia coli*; most of the proteins form inclusion bodies on overexpression. Several strategies have been reported to sidestep the problem associated with the solubilization and refolding of the overexpressed proteins from inclusion bodies (Hammarstrom *et al*., 2002; Tresaugues *et al*., 2004; Yang *et al*., 2003). These include (1) using a different host or strain of *E.coli*, (2) reducing the level of expression either by decreasing the induction temperature or by using weak promoters (Clark, 1998; Georgiou and Valax, 1999), (3) using small-molecule additives such as glycylglycine, L-arginine and sorbitol (Ghosh *et al*., 2004; Schein, 1990; Winter *et al*., 2001), (4) co-expressing chaperones (Machida *et al*., 1998) and (5) overexpressing the protein as a fusion protein (Makrides, 1996; Stevens, 2000). It is not very clear why only some, but not all, proteins are soluble on overexpression. A cursory look at the proteins, which are and are not soluble on overexpression, reveals that the primary sequence of the protein is the most important determinant of the solubility status of the overexpressed protein. The determinative role of the primary sequence is further confirmed by observations that point mutations alter the solubility status of the overexpressed protein under identical conditions (Dale *et al*., 1994; Jenkins *et al*., 1995; Malissard and Berger, 2001; Murby *et al*., 1995; Pedelacq *et al*., 2002; Timson and Reece, 2003). Sequence-independent factors such as the kinetics of translation (Cortazzo *et al*., 2002; Komar *et al*., 1999; Makrides, 1996), absence of certain post-translational modifications (Zhang *et al*., 1998) and reducing environment of the cytoplasm (Lilie *et al*., 1998; Makrides, 1996) have also been found to contribute to the solubility status in some cases.

An early attempt to determine the relationship between the amino acid sequence with the solubility status of the overexpressed protein was performed by Wilkinson and Harrison; they observed that inclusion body formation is correlated, in decreasing order of correlation, to charge average, turn-forming residue fraction, cysteine fraction, proline fraction, hydrophilicity and molecular weight (Davis *et al*., 1999; Wilkinson and Harrison, 1991). Spurred by structural genomics initiatives, additional investigations have been undertaken in this direction. Gerstein and coworkers analyzed 562 proteins of *Methanobacterium thermoautotrophicum* using genetic algorithms and several machine learning algorithms including decision tress and support vector machines (SVMs) (Bertone *et al*., 2001): the parameters found critical by this study are residues Glu, Ile, Thr and Tyr, combined composition of basic (Arg, Lys), acidic (Asp, Glu) and aromatic (Phe, Trp, Tyr) residues, acidic residues with their amides (Asn, Asp, Gln, Glu), presence of signal sequence and hydrophobic residues, secondary structural features, and low complexity regions. In a subsequent study, Gerstein and coworkers analyzed 27 267 protein targets selected for structural genomics studies and found serine percentage composition to be the

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

major determinant of solubility (Goh *et al*., 2004). Luan and coworkers performed expression experiments on 10 167 ORFs of *Caenorhabditis elegans* (Luan *et al*., 2004): with one expression vector and one *E.coli* strain, expression was observed for 4854 ORFs of which only 1536 were soluble. Analysis of these 1536 sequences revealed that the hydrophobicity is a key-determining factor for an ORF to yield a soluble expression product. It is to be noted here that the expression conditions and the nature of proteins (e.g. thermophilic/mesophilic, cytosolic/membrane-bound, etc.) analyzed are different in these studies. This could probably be the reason for the apparent dissimilarity between the results of the various studies.

Recently, a study was undertaken by Idicula-Thomas and Balaji to identify the sequence-dependent features that correlate to solubility of proteins when overexpressed in *E.coli* under normal growth conditions (Idicula-Thomas and Balaji, 2005). The proteins used in this study were based on literature reports on the solubility on overexpression in *E.coli* and included both thermophilic and mesophilic proteins from viruses, prokaryotes and eukaryotes. The study revealed that the aliphatic index, the frequency of occurrence of Asn, Thr and Tyr, and the dipeptide- and tripeptide-compositions significantly vary between the soluble and inclusion body-forming proteins.

SVM (Vapnik, 1995) has gained popularity over other machine learning methods for interpreting biological data (Bhasin and Raghava, 2004; Brown *et al*., 2000; Byvatov and Schneider, 2003; Ding and Dubchak, 2001; Furey *et al*., 2000; Jaakkola *et al*., 2000; Natt *et al*., 2004, Zien *et al*, 2000) because of their ability to very effectively handle noise and large datasets/input spaces (Zavaljevski *et al*., 2002). In the present study, an SVM-based algorithm has been developed to predict the solubility-on-overexpression based on the features identified by Idicula-Thomas and Balaji (2005). In addition, the effect of using the residue-, dipeptide- and tripeptide-compositions on classification has also been investigated. The use of SVM is especially appropriate in this case since the use of dipeptide- and tripeptide-compositions for classification results in a large increase in the number of features (by 400 and 8000, respectively).

SYSTEMS AND METHODS

Datasets

The proteins for the analyses were chosen based on literature reports on their solubility on overexpression in *E.coli* under normal growth conditions i.e. 37°C, without the use of solubility enhancing fusion tags or chaperone co-expression. These criteria were used to ensure that the observed solubility on overexpression is mainly owing to its sequence features rather than sequence-independent factors. Only 62 proteins could be obtained with these criteria for the dataset of soluble proteins (dataset **S**). It is to be noted that most of the proteins form inclusion bodies on overexpression in *E.coli*; only few proteins are found in the soluble fraction. Hence, the number of proteins that are available to populate the dataset of soluble proteins is meager. Even though a large number of proteins have been reported to form inclusion bodies on overexpression, the size of dataset **I** was restricted to 130, since a large difference in the number of proteins between the two datasets may hamper the SVM training procedure (Lin *et al*., 2002). Both the datasets **S** and **I** included thermophilic as well as mesophilic proteins from viruses, prokaryotes and eukaryotes. The accession numbers of the proteins included in the two datasets are given in Table S1.

Table 1. Summary of the reduced alphabet sets used in the study

Property	Reduced class	Vector size	References
Conformational similarity	[CMQLEKRA], [P], [ND], [G], [HWFY], [S], [TIV]	7	Chakrabarti and Pal (2001)
Hydrophobicity	[CFILMVW], [AG], [PH], [EDRK], [NQSTY]	5	Rose *et al*. (1985)
BLOSUM50 substitution matrix	[FWY], [CILMV], [H], [AG], [ST], [EDNQ], [KR], [P]	8	Murphy *et al*. (2000)

The proteins were randomly split into training and test datasets. The training dataset comprised 87 inclusion body-forming and 41 soluble proteins (total 128). The test dataset comprised 43 inclusion body-forming and 21 soluble proteins (total 64). The inclusion body-forming and soluble proteins are in approximately 2:1 ratio in the each of the two datasets. This ratio is approximately same as the ratio of the sizes of datasets **I** and **S**.

Representation of the proteins as vectors of fixed length

Pattern recognition algorithms require the proteins to be represented as fixed length vectors. Three different models, incorporating sequence information at various levels, were considered while performing the simulations. In order to investigate the effect of a particular class of amino acids on solubility, the 20 amino acids were grouped into various classes based on certain common properties and the composition of the reduced sets of amino acids were also considered for classification (Table 1). The six physicochemical properties of the protein, viz., length of the protein (L), hydropathic index (GRAVY), aliphatic Index (AI), instability index of the entire protein (II_P), instability index of N-terminus (II_N) and net charge (NC) were included along with single residue and dipeptide frequencies. This resulted in a dataset size of 446 features: 20 reduced alphabet sets (Table 1), 6 physicochemical properties (as above), 20 residues, 400 dipeptides. Details of calculating the various physicochemical parameters are included in Supplementary Data. Simulations were also performed by including tripeptide-composition (i.e. by including 8000 additional features) or by deleting dipeptide-composition (i.e. by decreasing 400 features). These three models were trained using the SVM algorithm.

ALGORITHM

Support vector machines for classification

SVMs are a class of machine learning algorithms that can perform pattern recognition and regression based on the theory of statistical learning and the principle of structural risk minimization (Vapnik, 1995, Muller *et al*., 2001). SVM tries to locate the hyperplane that maximally separates the training datasets by maximizing the margin between them. It non-linearly transforms the original input space into a higher-dimensional feature space by means of kernel functions. By doing so, the data become linearly separable in a high-dimensional feature space (Gunn, 1997; Kulkarni *et al*., 2004).

The training dataset is of the form $\{(\mathbf{x_i}, y_i)\}_{i=1,2,\ldots,N}$. Here $\mathbf{x_i}$ is the vector representing the sequence-dependent features for the i-th protein in the training dataset, y_i is the corresponding class of the protein and N is the total number of proteins in the training dataset. For soluble proteins $y_i = +1$ and for inclusion body-forming proteins $y_i = -1$; this assignment regards the soluble proteins as the

positive class and hence the number of true positives for the algorithm is the number of soluble proteins getting classified correctly in the test dataset. The SVM-based classification is dependent on the sign of $f(x)$, which is calculated as

$$f(\mathbf{x}) = \sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b$$

where m is the number of input data having non-zero values of Lagrange multipliers (α_i) (usually less than N) obtained by solving a quadratic optimization problem, $K(\mathbf{x_i}, \mathbf{x})$ is the kernel matrix and b is the bias term. Kernel matrix calculations are performed with kernel functions. The Gaussian Radial Basis Function kernel was exclusively used in the computations (Gunn, 1997; Burges, 1998).

The codes for all the algorithms were developed in-house. SVM performance was compared with one of the standard packages i.e. LIBSVM-2.6 (Chang and Lin, 2001, www.csie.ntu.edu.tw/~cjlin/libsvm) and since both the codes were performing equally well except for the speed of computations, in-house codes were used for all the computations. Use of in-house code, in addition, enabled the extraction of the magnitude of $f(x)$.

The following stepwise procedure was employed in the simulations to implement the algorithm:

(1) Get the protein sequence data.
(2) Assign labels—soluble proteins: positive class; insoluble proteins: negative class.
(3) Convert all the sequences to their numerical equivalents.
(4) Scale the features to zero mean and SD 1.
(5) Partition the data as training and test sets.
(6) Run SVM classifier on training set.
(7) Run SVM classifier on the test set to assess the generalization.

In a separate simulation, we tried steps 5–7 with only 20 features that were ranked at the top (for SVM model with 446 features) with unbalanced correlation score method. We found that classification accuracy for this is more or less same (with 70 ± 1% classification).

In another variation, we employed the following procedure:

(1) Steps (1)–(6) are same as earlier.
(2) Add random Gaussian Noise in a feature.
(3) Observe the change in SVM discriminant function value f(x) to check the sensitivity to solubility.
(4) Repeat steps (2) and (3) for all the features.

IMPLEMENTATION

Performance of SVM classifier

SVM classifier was applied to discriminate between soluble and inclusion body-forming proteins based on the sequence-dependent features. All the features were scaled to zero mean and SD 1. The various user-defined parameters e.g. kernel width parameter σ and regularization parameter C were selected using 5-fold stratified cross validation on the training dataset. Various values of σ and C were tried in the range of [0.1–5] and [0.1–1000], respectively. The original class distribution (ratio of 1:2 between the soluble and inclusion body-forming proteins) is approximately maintained during the stratification procedure. The performance of the algorithm was tested on the unseen test dataset after training it with optimal parameters.

Table 2. Classification result on test dataset

Algorithm	Number of features	Prediction accuracy[a] (%)	Specificity[b] (%)	Sensitivity[c] (%)	Enrichment factor[d]
SVM	446	72	76	55	1.68
	46	66	48	48	1.45
	8446	67	67	50	1.52

[a]Prediction accuracy is defined as $[(cs + ci)/t] \times 100$, where cs and ci are the number of correctly classified soluble and inclusion body-forming proteins, respectively, and t is the total number of proteins in the soluble and inclusion body-forming test datasets.
[b]Specificity is defined as $[cs/(cs + wi)] \times 100$, where cs is as above and wi is the number of inclusion body-forming proteins wrongly classified as soluble proteins.
[c]Sensitivity is defined as $[cs/(cs + ws)] \times 100$, where cs is as above and ws is the number of soluble proteins wrongly classified as inclusion body-forming proteins.
[d]Enrichment factor (Ef) is defined as $\{[cs/(cs + ws)]/[s/(s + i)]\}$, where cs and ws are as above and s and i are the total number of soluble and inclusion body-forming proteins, respectively, in the respective test datasets. Ef is especially suitable for the unbalanced datasets, where class distribution is not even. Ef > 1 indicates enrichment in the classifier performance whereas Ef = 0 indicates no effect and Ef < 1 indicates impairment to the classifier performance (Fechner *et al*., 2003).

To start with, the SVM classifier was trained with 46 features comprising 20 reduced alphabet sets (Table 1), 6 physicochemical properties (described in the last paragraph under the Systems and methods section) and 20 residues. The test accuracy of the classifier was only 66%. The use of additional information in the form of dipeptide-composition improved the accuracy of classification yielding a value of 72% (Table 2). There was no improvement in classification accuracy when tripeptide-composition was used as additional information (Table 2). A possible reason for this is that many tripeptides are underrepresented or not represented at all, owing to the small size of the datasets. This could have created a redundancy in the training procedure and thus had a negative impact on the prediction accuracy (Burges, 1998). Other performance parameters are as shown in Table 2. ROC curve for the best classifier (with 446 features) is shown in Figure 1. Area under ROC curve and area under convex hull of ROC curve indicate that the classifier is not a random classifier and the classification accuracies, which we are getting, are reasonable.

To rule out the effect of sampling of proteins into the training and test datasets, 50 random splits of the datasets **S** and **I** into training and test datasets (with the same ratio of nearly 1:2 between the two classes of proteins) were created and the three models were trained and tested separately for each of the splits. No marked changes in the prediction accuracy were observed in each of these cases (data not shown) suggesting that the nature of the training and test datasets has not biased the prediction accuracy of the classifier.

Other than SVM, we also tried two conventional classifiers i.e. linear logistic regression and k-nearest neighbor classifier (with $k = 5$, by cross validation). Both these classifiers resulted in base level accuracy of 67% (i.e. predicted only inclusion body-forming proteins correctly). SVM with linear kernel also resulted in baseline accuracy of 67%. We did not include these results in Table 2, since other than classification accuracy, all the performance parameters are either zero or meaningless. We did not plot ROC

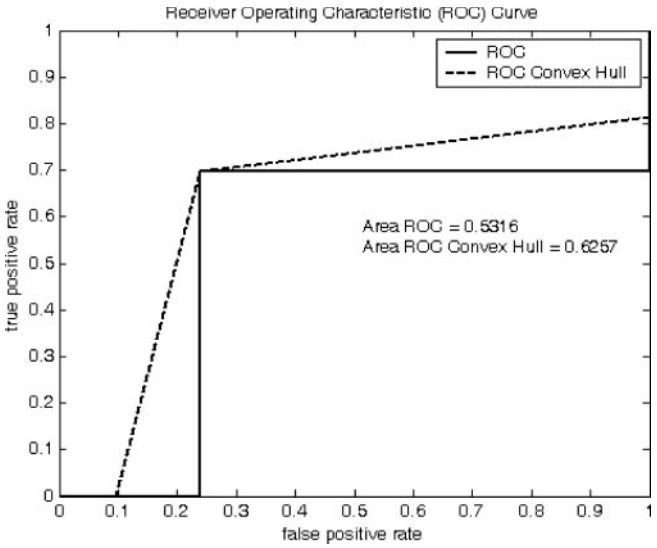

Figure 1. ROC Curve for SVM classifier (with 446 features).

curves for these classifiers (*k*-nn, linear logistic regression) since none of them predicted any of the observations from the positive class (soluble proteins). In case of *K*-NN, we further optimized the number of neighbors and found that with increased neighbors (and subsequently increase in classifier complexity) it is able to classify few of the positive points with classification accuracy 67%, sensitivity 50%, specificity 38% and enrichment factor as 1.56.

Weighted classifiers

Owing to the fact that classes are imbalanced in the dataset, we performed some simulations adding class-dependent weights to regularize the learning process in K-NN and SVM. Here we considered the dataset containing 446 features. We call these classifiers as weighted_KNN and weighted_SVM (Lin *et al*., 2002). The results of both these classifiers are improved as compared with their non-weighted counterparts. The results are as follows:

Weighted_KNN: 72% classification accuracy, 57% sensitivity, 57% specificity and enrichment factor as 1.78.

Weighted_SVM: 74% classification accuracy, 57% sensitivity, 81% specificity and enrichment factor as 1.78.

Feature extraction

The results of our experiments have categorically indicated that SVMs with 46 features and 8446 features do not perform well and weighted_SVM does not show substantial improvement. So we resort to unweighted counterpart for further analysis taking into account the extra computational cost owing to addition of weights. It is evident that the SVM classifier with 446 features provides good classification performance. With a view to ascertain whether there is a subset of most informative features among these 446 features, we employed different feature selection methods like principal component analysis, genetic algorithm, etc. Our simulations indicated that none of these algorithms could pick out informative subsets of features exhibiting satisfactory prediction accuracy. Noting that our problem consists of a dissimilar number of positive and negative samples and the number of features is quite high, a feature selection algorithm, viz., unbalanced correlation score (Weston *et al*., 2003) was then used. In this method, the features in the models are ranked for their contribution to prediction of the expression status by the following criterion:

$$f_j = \sum_{y_i=1} \mathbf{X}_{ij} - \lambda \sum_{y_i=-1} \mathbf{X}_{ij},$$

where f_j represents the score (called as unbalanced correlation score) for feature j, $\mathbf{X}$ is a training data matrix where rows represent the proteins and the columns represent the sequence-dependent features of the corresponding proteins. λ is a regularization parameter and can be optimized using cross validation. It is evident from the mathematical expression that the sign of the score value is the sign of correlation with solubility. Further, a sensitivity analysis (Daae and Ison, 1999) was performed to determine how well the features, which have a high unbalanced correlation score, correlate with solubility. The correlation of a feature to solubility was determined by perturbing the feature (keeping the other features constant) and monitoring the effect of the perturbation on the number of true positives detected by the SVM classifier. Perturbation is added in the form of random Gaussian Noise. If the number of proteins identified as soluble increases with positive changes in the feature then it is inferred that the feature has positive correlation with solubility and vice versa.

Sequence-dependent features that correlate to solubility

Based on the unbalanced correlation score and sensitivity analysis, the correlations of 20 highest-ranked-features (for the SVM model with 446 features) to solubility were determined (Table 3). We ran the SVM classifier with only these 20 top-ranked features and found that classification accuracy is more or less same (with 70 ± 1% classification). Thus, unbalanced correlation score feature extraction method performed better than conventional methods but could not increase the prediction accuracy of the SVM with 446 features. The knowledge of physicochemical parameters and the compositions of residues and dipeptides that favor/disfavor solubility (Table 3) could, however, be used in designing mutations that would bring about the desired change in their solubility, thus warranting further analysis and discussion.

Among the physicochemical properties considered, thermostability of proteins, as represented by its aliphatic index AI, is found to be the most crucial determinant of solubility (Table 3). This suggests that an increase in thermostability of the proteins would enhance the propensity for the protein to be overexpressed in a soluble form. It has been noted that thermolabile folding intermediates, being chaperonin substrates, could increase the propensity to aggregate and form inclusion bodies by exhausting the *in vivo* supply of chaperones (King *et al*., 1996).

A higher instability index also appears to increase the propensity of a protein to be overexpressed in the soluble form (Table 3). This observation suggests that a protein with a lower *in vivo* half-life has a lesser propensity to form inclusion bodies compared with proteins with a higher *in vivo* half-life. The correlation between *in vivo* half-life of a protein and solubility could be rationalized by the role played by longer-lived partially folded intermediates of a protein. These long-lived intermediates can interact with a greater chance

Table 3. Top 20 features selected by feature selection algorithm

Rank[a]	SVM model with 446 features	Correlation score value	Correlation with solubility[b]
1	AI	0.55	P
2	Glu	0.32	P
3	His-His	0.28	P
4	Arg-Gly	0.26	P
5	Arg	0.25	P
6	Gly	−0.38	N
7	II_P	0.24	P
8	NC	0.24	P
9	Asn-Thr	−0.35	N
10	Arg-Ala	0.23	P
11	Cys	−0.32	N
12	Met	−0.3	N
13	Gln	0.22	P
14	Phe	−0.30	N
15	Ile	0.22	P
16	Gly-Ala	0.21	P
17	II_N	0.21	P
18	Ser	−0.29	N
19	Leu	0.20	P
20	Pro	−0.29	N

[a]The features are ranked in the descending order in relation to correlation to solubility. The feature that is ranked one for a model was found to have the highest correlation to solubility for the respective model.
[b]P denotes positive correlation with solubility and N denotes negative correlation with solubility.

with other partially folded intermediates and also exhaust the limited *in vivo* supply of chaperones (Fink, 1998), thus contributing to inclusion body formation.

It is observed that a higher net charge favors solubility on overexpression (Table 3). There have been previous reports of the favorable role played by a higher net charge in reducing inclusion body formation (Dale *et al*., 1994; Malissard and Berger, 2001; Davis *et al*., 1999) and protein aggregation (Chiti *et al*., 2003; Monti *et al*., 2004) in general. Among the positively charged residues, Arg seems to have a positive effect on solubility. Interestingly, it is observed that, of the negatively charged amino acids, Glu exerts a positive influence on solubility (Table 3) whereas Asp has an opposite effect. It is to be noted that Glu has a higher helix propensity compared with Asp (Kallberg *et al*., 2001). The significance of this observation comes in light of the helix to sheet transition seen to accompany inclusion body formation in certain cases (Przybycien *et al*., 1994).

An enrichment of hydrophobic residues in proteins has been reported to be associated with an increased propensity to form inclusion body (Luan *et al*., 2004). As expected, the hydrophobic residues, viz. Cys, Met and Phe were found to hinder solubility. The hydrophobic residues (Ile and Leu) that enhance the aliphatic index were however found to favor solubility (Table 3) and this is also reflected by the higher aliphatic index of soluble proteins (Idicula-Thomas and Balaji, 2005). Apart from the hydrophobic residues, the turn-forming residues Asn, Gly, Pro and Ser have also been reported to favor inclusion body formation (Davis *et al*., 1999) and in accordance with this, the residues Gly, Pro and Ser were found to be negatively correlated to solubility in the present study also (Table 3).

Although Asn was not selected among the top 20 features to influence solubility at the residue level, the dipeptide (Asn-Thr) was found to negatively correlate to solubility. The dipeptides that were found to positively correlate to solubility are His-His, Arg-Gly, Arg-Ala and Gly-Ala (Table 3). The amino acid composition is known to influence the folding kinetics of the protein (Chan and Dill, 1994; Socci and Onuchic, 1994), which, in turn, plays a crucial role in deciding the propensity of the protein to form inclusion bodies (Finke *et al*., 2000; Hoffmann *et al*., 2001). The knowledge of the mechanism by which the above-mentioned dipeptides influence the folding kinetics of the proteins would probably help in rationalizing their correlation to solubility.

Use of the SVM-based algorithm in engineering mutations for enhanced solubility

In order to validate the use of the algorithm in designing mutations for improved/altered solubility, the SVM-based predictions were compared with the experimental observations for certain point mutations reported in the literature. For this analysis, we employed SVM with 446 features. It was observed that in 22 of the 23 point mutations studied, the SVM-based algorithm could correctly predict the change in solubility brought about by the mutation (Table 4 and Table S2). When the results of the SVM-based prediction were compared with the two existing models for prediction of solubility, viz., Wilkinson-Harrison (WH) model (Davis *et al*., 1999) and the solubility index (SI) based model proposed by Idicula-Thomas and Balaji (2005), it was observed that these models could correctly predict the change in solubility only in 7 and 16 cases, respectively, out of the 23 point mutations studied (Table 4). Since the quantitative measure of solubility of proteins in the training datasets is not known, the changes predicted for solubility are mentioned in a qualitative sense only. The actual prediction accuracy is very low (21.73%) if the sign of the SVM discriminant function value is used. This is not surprising since there is a very strong overlap between these sequences and the classifier could predict only wild-type proteins. However, a positive (or a negative) change in SVM discriminant function value indicates an increase (or a decrease) in the solubility of proteins. We used SVM owing to its strong generalization abilities. Other classifiers (k-NN and logistic regression) resulted in baseline accuracies only and hence were not used in mutation studies.

DISCUSSION

Pattern classification algorithms based on statistical learning paradigms are gaining popularity in the field of biological sciences. In the present work, SVM-based classifier has been trained to predict the expression status of proteins in *E.coli* based on the sequence-based features of these proteins. The most critical sequence-based features for prediction of the expression status of the proteins could be inferred from the unbalanced correlation score of these features and the effect of these features on solubility were inferred based on the sensitivity analysis.

The proteins considered in the present study are based on the overexpression status of proteins as reported in the literature by various research groups. The fraction of proteins overexpressed in the soluble form varies for various proteins and the information on the extent of solubility or the formation of inclusion bodies is not available for most of the proteins. With an increase in the accurate

Table 4. Effect of mutation on solubility on overexpression: comparison of SVM-based prediction and experimental observation

Mutation	Effect of mutation on solubility			
	Experimental observation	SVM-based prediction[a]	SI-based prediction[b]	WH scheme-based prediction[c]
HIV-1 Integrase Core Domain (Jenkins *et al*., 1995)				
F185K	Increase	Increase	Increase	Decrease[d]
V165K	Increase	Increase	Increase	Decrease[d]
Nucleoside diphosphate kinase (Pedelacq *et al*., 2002)				
A10D	Increase	Increase	Decrease[d]	Increase
E40K	Increase	Increase	Increase	Decrease[d]
Dihydrofolate reductase (Dale *et al*., 1994)				
N130D	Increase	Increase	Increase	Increase
N48E	Increase	Increase	Increase	Increase
N48E/N130D	Increase	Increase	Increase	Increase
Human interferon gamma (Wetzel *et al*., 1991)				
P122S	Increase	Increase	Increase	No change[d]
K128R	Increase	Increase	Decrease[d]	No change[d]
K128Q/R129P	Increase	Increase	Decrease[d]	Decrease[d]
A123S/K125I/K130E/S132G	Increase	Increase	Decrease[d]	Decrease[d]
K87Q	Increase	Increase	Increase	Decrease[d]
V79A	Increase	Decrease[d]	Decrease[d]	No change[d]
Human galactokinase (Timson and Reece, 2003)				
P28K	Decrease	Decrease	Increase[d]	Decrease
V32M	Decrease	Decrease	Decrease	No change[d]
G36R	Decrease	Decrease	Decrease	Decrease
T288M	Decrease	Decrease	Decrease	No change[d]
A384P	Decrease	Decrease	Increase[d]	Decrease
H44Y	Decrease	Decrease	Decrease	No change[d]
R68C	Decrease	Decrease	Decrease	Increase[d]
G346S	Decrease	Decrease	Decrease	No change[d]
G349S	Decrease	Decrease	Decrease	No change[d]
A198V	Decrease	Decrease	Decrease	No change[d]

The wild-type protein is soluble and the mutants are insoluble.
[a]The solubility of a mutant is inferred to increase (or decrease) if the value of $f(x)$ for the mutant is higher (or lower) than that of the corresponding wild-type protein.
[b]This is the prediction model for solubility suggested by Idicula-Thomas and Balaji (2005).
[c]This is the prediction model for solubility adopted by Davis *et al*. (1999).
[d]In this case there is disagreement between the experimental observation and the prediction of change in solubility.

information of the expression status of proteins, one can expect improvements in the prediction accuracies using the SVM classifier based on the sequence-based features of the proteins.

A few structural genomics studies aimed at the identification of 'soluble' proteins based on amino acid sequences have employed large datasets, e.g. 562 (Bertone *et al*., 2001), 27 267 (Goh *et al*., 2004) and 4854 (Luan *et al*., 2004) proteins. The experimental conditions for expression of proteins such as the host strain, temperature, media, etc. are the same within each of these studies. However, these studies identified different features as crucial for solubility-on-expression of a protein. The nature of proteins (e.g. thermophilic or mesophilic, cytosolic or membrane-bound, etc.) and the expression/overexpression conditions are critical in dictating solubility of the protein. But, the role of the 3D structures of the folding intermediates and that of the native protein in governing the expression status of the proteins cannot be undermined. The sequence-based features contributing to the solubility status are quite subtle. Hence, a certain level of misclassification is expected owing to the absence of the structure-based features in the SVM-based prediction algorithm. In light of this, the 72% accuracy (Table 2) achieved by the present SVM-based algorithm for predicting the solubility status of proteins on overexpression in *E.coli* under normal growth conditions based on sequence-based features alone is quite significant in the absence of 3D structure-based features. The algorithm is also able to rationalize the experimentally observed effect of certain point mutations on solubility of the proteins.

The SVM-based algorithm developed in the present study is especially important for the large-scale structural genomics initiatives wherein high-throughput methods are currently being developed to identify proteins that would be overexpressed in soluble fraction in *E.coli* (Knaust and Nordlund, 2001). The prediction algorithm can also be helpful in replacing directed evolution methods currently undertaken for screening protein libraries for soluble variants (Waldo, 2003).

ACKNOWLEDGEMENTS

S.Idicula-Thomas and A.J.K. are grateful to the Council of Scientific and Industrial Research, India for research fellowships. Financial support from the Department of Science and Technology, India is gratefully acknowledged by V.K.J. and B.D.K.

Conflict of Interest statement: none declared.

REFERENCES

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.

Bhasin,M. and Raghava,G.P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **32** (Web Server issue), W414–W419.

Brown,M.P. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA.*, **97**, 262–267.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.*, **2**, 121–167.

Byvatov,E. and Schneider,G. (2003) Support vector machine applications in bioinformatics. *Appl. Bioinformatics*, **2**, 67–77.

Chakrabarti,P. and Pal,D. (2001) The interrelationships of side-chain and main-chain conformations in proteins. *Prog. Biophys. Mol. Biol.*, **76**, 1–102.

Chan,H.S. and Dill,K.A. (1994) Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.*, **100**, 9238–9257.

Chang,C. and Lin,C.J. (2001) LIBSVM: a library for support vector machines. .

Chiti,F. *et al.* (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **424**, 805–808.

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

Cortazzo,P. *et al.* (2002) Silent mutations affect *in vivo* protein folding in *Escherichia coli*. *Biochem. Biophys. Res. Commun.*, **293**, 537–541.

Daae,E.B. and Ison,A.P. (1999) Classification and sensitivity analysis of a proposed primary metabolic reaction network for *Streptomyces lividans*. *Metab. Eng.*, **1**, 153–165.

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

Ding,C.H. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Fechner,U. *et al.* (2003) Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput. Aided Mol. Des.*, **17**, 687–698.

Fink,A.L. (1998) Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des.*, **3**, R9–R23.

Finke,J.M. *et al.* (2000) Aggregation events occur prior to stable intermediate formation during refolding of interleukin 1beta. *Biochemistry*, **39**, 575–583.

Furey,T.S. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Georgiou,G. and Valax,P. (1999) Isolating inclusion bodies from bacteria. *Methods Enzymol.*, **309**, 48–58.

Ghosh,S. *et al.* (2004) Method for enhancing solubility of the expressed recombinant proteins in *Escherichia coli*. *Biotechniques*, **37**, 418, 420, 422–423.

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.

Gunn,S. (1997) Support vector machines for classification and regression. *ISIS technical report*.

Hammarstrom,M. *et al.* (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci.*, **11**, 313–321.

Hoffmann,F. *et al.* (2001) Kinetic model of *in vivo* folding and inclusion body formation in recombinant *Escherichia coli*. *Biotechnol Bioeng.*, **72**, 315–322.

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.

Jaakkola,T. *et al.* (2000) A discriminative framework for detecting remote protein homologies. *J Comput Biol.*, **7**, 95–114.

Jenkins,T.M. *et al.* (1995) Catalytic domain of human immunodeficiency virus type 1 integrase: identification of a soluble mutant by systematic replacement of hydrophobic residues. *Proc. Natl Acad. Sci. USA.*, **92**, 6057–6061.

Kallberg,Y. *et al.* (2001) Prediction of amyloid fibril-forming proteins. *J. Biol. Chem.*, **276**, 12945–12950.

King,J. *et al.* (1996) Thermolabile folding intermediates: inclusion body precursors and chaperonin substrates. *FASEB J.*, **10**, 57–66.

Knaust,R.K. and Nordlund,P. (2001) Screening for soluble expression of recombinant proteins in a 96-well format. *Anal. Biochem.*, **297**, 79–85.

Komar,A.A. *et al.* (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during *in vitro* translation. *FEBS Lett.*, **462**, 387–391.

Kulkarni,A. *et al.* (2004) Support vector classification with parameter tuning assisted by agent-based technique. *Comput. Chem. Eng.*, **28**, 311–318.

Lilie,H. *et al.* (1998) Advances in refolding of proteins produced in *E. coli*.. *Curr. Opin. Biotechnol.*, **9**, 497–501.

Lin,Y., Lee,Y. and Wahba,G. (2002) Support vector machines for classification in nonstandard situations. *Machine Learning*, **46**, 191–202.

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

Machida,S. *et al.* (1998) Overproduction of beta-glucosidase in active form by an *Escherichia coli* system coexpressing the chaperonin GroEL/ES. *FEMS Microbiol Lett.*, **159**, 41–46.

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.

Monti,M. *et al.* (2004) The regions of the sequence most exposed to the solvent within the amyloidogenic state of a protein initiate the aggregation process. *J Mol Biol.*, **336**, 253–262.

Muller,K.R. *et al.* (2001) An Introduction to Kernel-Based Learning Algorithms. *IEEE Trans Neural Netw.*, **2**, 181–201.

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.

Murphy,L.R. *et al.* (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.*, **13**, 149–152.

Natt,N.K. *et al.* (2004) Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins*, **56**, 11–18.

Pedelacq,J.D. *et al.* (2002) Engineering soluble proteins for structural genomics. *Nat. Biotechnol.*, **20**, 927–932.

Przybycien,T.M. *et al.* (1994) Secondary structure characterization of beta-lactamase inclusion bodies. *Protein Eng.*, **7**, 131–136.

Rose,G.D. *et al.* (1985) Hydrophobicity of amino acid residues in globular proteins. *Science*, **229**, 834–838.

Schein,C.H. (1990) Solubility as a function of protein structure and solvent components. *Biotechnology*, **8**, 308–317.

Socci,N.D. and Onuchic,J.N. (1994) Folding kinetics of protein-like heteropolymers. *J. Chem. Phys.*, **100**, 1519–1528.

Stevens,R.C. (2000) Design of high-throughput methods of protein production for structural biology. *Structure*, **8**, R177–R185.

Timson,D.J. and Reece,R.J. (2003) Functional analysis of disease-causing mutations in human galactokinase. *Eur. J. Biochem.*, **270**, 1767–1774.

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

Vapnik,V. (1995) *The nature of statistical learning theory*. Springer, 1st edn. NY.

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

Weston,J. *et al.* (2003) Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics*, **19**, 764–771.

Wetzel,R. *et al.* (1991) Mutations in human interferon gamma affecting inclusion body formation identified by a general immunochemical screen. *Biotechnology*, **9**, 731–737.

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

Winter,J. *et al.* (2001) Increased production of human proinsulin in the periplasmic space of *Escherichia coli* by fusion to DsbA. *J. Biotechnol.*, **84**, 175–185.

Yang,J.K. *et al.* (2003) Directed evolution approach to a structural genomics project: Rv2002 from Mycobacterium tuberculosis. *Proc. Natl Acad. Sci. USA.*, **100**, 455–460.

Zavaljevski,N. *et al.* (2002) Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**, 689–696.

Zhang,Y. *et al.* (1998) Expression of eukaryotic proteins in soluble form in *Escherichia coli*. *Protein Expr. Purif.*, **12**, 159–165.

Zien,A. *et al.* (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.