Prediction of Essential Proteins in Prokaryotes by Incorporating Various Physico-chemical Features into the General form of Chou's Pseudo Amino Acid Composition

Aditya Narayan Sarangi[1], Mohtashim Lohani[2] and Rakesh Aggarwal*,[1]

[1]Biomedical Informatics Centre, School of Telemedicine and Biomedical Informatics, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Raebareli Road, Lucknow 226014, India; [2]Department of Biotechnology, Integral University, Dasauli. P.O. Bas-ha Kursi Road, Lucknow 226026, India

Abstract: Prediction of essential proteins of a pathogenic organism is the key for the potential drug target identification, because inhibition of these would be fatal for the pathogen. Identification of these proteins requires the use of complex experimental techniques which are quite expensive and time consuming. We implemented Support Vector Machine algorithm to develop a classifier model for in silico prediction of prokaryotic essential proteins based on the physico-chemical properties of the amino acid sequences. This classifier was designed based on a set of 10 physico-chemical descriptor vectors (DVs) and 4 hybrid DVs calculated from amino acid sequences using PROFEAT and PseAAC servers. The classifier was trained using data sets consisting of 500 known essential and 500 non-essential proteins (n=1,000) and evaluated using an external validation set consisting of 3,462 essential proteins and 5,538 non-essential proteins (n=9,000). The performances of individual DV sets were evaluated. DV set 13, which is the combination of composition, transition and distribution descriptor set and hybrid autocorrelation descriptor set, provided accuracy of 91.2% in 10-fold cross-validation of the training set and an accuracy of 89.7% in external validation set and of 91.8% and 88.1% using a different yeast protein dataset. Our result indicates that this classification model can be used for identification of novel prokaryotic essential proteins.

Keywords: Machine learning, support vector machine, essential protein, classification

1. INTRODUCTION

Knowledge of complete genome sequences and proteome composition of several pathogenic organisms has made it possible to determine potential drug targets in these organisms using computer-aided tools. These tools are based on the assumption that the proteins used as drug targets must be essential for the survival of the pathogen, but not of the human host [1-3]. Interference with these essential proteins, which are metabolically active and thus crucial for the sustenance of cellular function, may be expected to be fatal for the pathogenic microorganism.

Identification and characterization of microbial essential proteins by traditional means requires the use of complex techniques, such as high-throughput gene disruption systems, anti-sense RNA technology, genome-wide mutagenesis, global transposon mutagenesis, systematic single-gene knockout experiments, etc. [4]; each of these approaches is however quite expensive and time consuming. Sequence-similarity-based in silico tools can also be used to identify highly conserved genes or proteins which are essential for survival, growth and replication of pathogenic organisms. However, this approach has some limitations, including (i) the use of arbitrary cut-off scores to differentiate between putative essential and non-essential proteins, (ii) failure to recognize that some proteins may be essential under some specific growth conditions but not others [2, 3], and (iii) inability to reliably distinguish between related and unrelated proteins at low pairwise sequence identity, e.g. below ≈25% [5, 6].

Recent progress in machine learning techniques has led to the development of tools that provide a better classification of objects based on their features. One such tool is Support Vector Machine (SVM) [7]. It is a supervised machine learning technique which has been extensively applied for solving classification problems in several biological and biomedical fields. Because of their robust predictive and highly accurate classification ability, SVM classification algorithms are considered to be superior to other supervised learning methods [8], and well suited for inductive inference, i.e. prediction based on prior observations. The SVM algorithms have thus been applied to a wide range of prediction problems, such as identification of drug targets [9-14], analysis of microarray gene expression data [15], prediction of sub-cellular localization of proteins [16-19], assignment of proteins to functional families [20-22], assessment of propensity of a protein to crystallize [23-25], prediction of protein solubility [26] likelihood of protein-protein interactions [27, 28], identification of nucleosome [29], classification and

*Address correspondence to this author at the Department of Gastroenterology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Raebareli Road, Lucknow 226014, India; Tel: 91-522-249 4431; Fax: 91-522-266 8017; E-mail: aggarwal.ra@gmail.com

1875-5305/13 $58.00+.00

analysis of regulatory pathways [30], and prediction of protein domains [31].

Since essential proteins can be expected to share certain features distinct from those of non-essential proteins, it should theoretically be possible to develop an SVM algorithm to predict whether a protein belongs to one of these two categories. This approach can significantly reduce the time required for identification and characterization of microbial essential proteins. In the current study, we therefore implemented SVM to develop a classification model for in silico prediction of 'prokaryotic essential proteins' in newly-sequenced microbial proteomes, based solely on the available aminoacid sequence information, and tested the validity of this approach.

2. MATERIALS AND METHODS

2.1. Construction of Datasets

2.1.1. Construction of a Positive Dataset

Aminoacid sequences of all the prokaryotic essential proteins included in the database of essential genes (n=7,643; version 6.8, accessed on November 4, 2011) [4] were downloaded. Sequences with length less than 50 aminoacid residues (n=216) were excluded. The remaining 7,427 protein sequences were processed using the CD-HIT program [32] to remove redundant proteins with amino acid sequence identity exceeding 40%. The remaining 3,962 non-redundant protein sequences formed our essential protein (gold-standard positive) dataset.

2.1.2. Construction of Negative Dataset

Since no dataset of experimentally-confirmed non-essential (gold-standard negative) proteins is available, we constructed one using an in silico approach. Positive protein sequences in the previous step were subjected to BLASTP against the Pfam v24 database [33] using the software JFeature [34] at expectation (E-value) cut-off of 10^{-8}, gap penalty of 11, gap extension penalty of 1and amino acid identity of less than 30%. This procedure of negative dataset preparation is implemented in references [20, 21, 35]. The program JFeature was instructed to randomly select 6,038 protein sequences from among the proteins fulfilling the above criteria. The amino acid sequences of these proteins were used as a negative dataset. The number of items in the negative dataset was so chosen that the total number of items in the two datasets was exactly 10,000. A flow diagram for construction of positive and negative datasets is shown in (Fig. **1**).

2.1.3. Construction of Training and Testing Datasets

From the above datasets of essential (positive) and non-essential (negative) protein sequences, smaller separate training and testing sets were generated using the following procedures.

2.1.3.1. Generation of Training Sets

Machine learning techniques are the most efficient when the training dataset contains an equal number of positive and negative feature vectors (items)[36]. Hence, we used a training dataset containing feature vectors of equal number of essential and non-essential proteins. For our initial experiments, we randomly selected 500 sequences each from the 3,962 positive and 6,038 negative protein sequences. For sensitivity analysis, alternative training sets ware constructed by repeating the process of random selection of 500 positive and 500 negative vectors three times. We thus generated four separate training sets, each containing 500 positive and 500 negative descriptor vectors (DVs).

2.1.3.2. Generation of Test Sets

Test sets are used as datasets for external validation, i.e. to measure the true predictive ability of the classifier models. In nature, the number of non-essential proteins far exceeds that of essential proteins. To simulate this situation, we created external validation datasets comprising of a larger num-

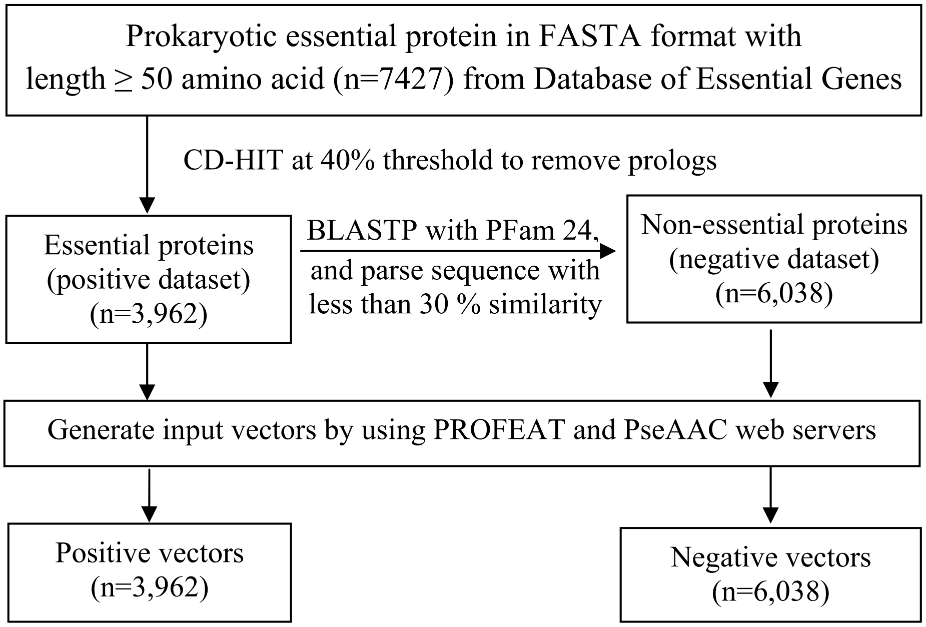

Prokaryotic essential protein in FASTA format with length ≥ 50 amino acid (n=7427) from Database of Essential Genes
CD-HIT at 40% threshold to remove prologs
Essential proteins (positive dataset) (n=3,962)
BLASTP with PFam 24, and parse sequence with less than 30 % similarity
Non-essential proteins (negative dataset) (n=6,038)
Generate input vectors by using PROFEAT and PseAAC web servers
Positive vectors (n=3,962)
Negative vectors (n=6,038)

Figure 1. Diagram representing positive and negative dataset construction.

ber of non-essential proteins than of essential proteins. In all, four such test datasets of 9,000 proteins each were generated, by excluding the proteins selected for each of the four training sets from the overall pool of 10,000 positive and negative proteins. Each test dataset thus contained 3,462 essential and 5,538 non-essential proteins.

2.1.3.3. External Validation Using Yeast Proteins

To further validate our approach, we downloaded the total proteome of a yeast, *Saccharomyces cerevisiae* strain S288c (n=6,627) from UniProt and obtained the yeast essential protein dataset from DEG (n=1,110). The complete proteome of *Saccharomyces cerevisiae* was subjected to BlastP against yeast essential protein database and *S. cerevisiae* protein sequences having no significant similarity with essential proteins were parsed from the BlastP output and considered as non-essential protein sequences (n=688).

Training Dataset (balanced dataset). We randomly selected 250 each of the 1,110 essential proteins and 688 non-essential *S. cerevisiae* proteins and constructed a balanced training dataset. This procedure was repeated four times to generate four balanced training datasets.

External validation set. We randomly selected 200 essential proteins and 500 non-essential proteins and constructed an external validation set (n=700). This procedure was repeated four times to generate four external validation sets.

2.1.3.4. Dataset for Comparison of Prediction Performance

We downloaded the raw yeast dataset (*Saccharomyces cerevisiae)* used by Gustafson et.al [37] which contains information on 61DVs of 966 yeast essential proteins and 3762 non-essential proteins. Of these DVs, we selected the top 42 DVs that were identified by these authors using conditional mutual information maximization criteria [37].

Training datasets (balanced dataset). We randomly selected 250 each of the 966 essential proteins and 3762 non-essential *S. cerevisiae* proteins, and constructed a balanced training dataset. This procedure was repeated four times to generate four balanced training datasets.

External validation sets. From among the proteins not included in the training set, we randomly selected 200 essential proteins and 500 non-essential proteins to obtain an external validation set (n=700); there were four external validation sets, one corresponding to each training set.

2.2. Extraction of Features from Protein Sequences

For both training and prediction steps, SVM techniques require each data instance (item) to be represented as a set of vectors of real numbers, with each vector representing one feature. Our hypothesis was that the SVM classifiers for prediction of essential proteins can be based on features that can be extracted from amino acid sequences. For the current study, we extracted 10 sets of widely used protein feature descriptor vectors from the amino acid sequences of proteins, using PROFEAT (Protein Feature) [38] and PseAAC (pseudo amino acid composition) [39] web servers.

PROFEAT has been used frequently for several prediction and classification tasks, including prediction of protein structural and functional classes[40], identification of N-glycosylation sites [41], classification of lung cancers [42], prediction of drug-target interaction networks [43], and prediction of ligands for orphan targets [44]. The concept of PseAAC (pseudo aminoacid composition) was proposed by Chou[45]. This method avoids the loss of important encrypted information that is hidden in protein primary structure. PseAAC has been used in prediction of several protein attributes, such as outer membrane proteins [46], metalloproteinase family [47], protein folding rates [48], protein-protein interactions [49], subcellular localization of virus proteins [50], protein solubility [51], and allergenic proteins [52]. PseAAC has also been used in identification of DNA-binding proteins [53], G-protein coupled receptors and their types [54, 55], risk type of human papillomaviruses [56], protein quaternary structural attributes [57], bacterial virulent proteins [58], etc. Due to wide implementation of PseAAC, a standalone tool called PseAAC-Builder has been implemented [59], which allows for calculation of various Chou's pseudo-amino acid composition vectors, in addition to the web-server PseAAC [39].

We used PROFEAT and PseAAC servers to obtain the following DVs:

2.2.1. Descriptor Vectors Based on Physico-chemical Properties

DV1: Amino acid composition (AAC). This descriptor specifies the fraction of each amino acid type in the amino acid sequence [60]. Thus, this descriptor vector, generated for each protein using the PROFEAT server, consisted of a total of 20 descriptor values, one for each amino acid.

DV2: Pseudo amino acid composition (also called parallel-correlation type). This descriptor vector was generated using PseAAC web server. The pseudo aminoacid composition of a protein sequence can be defined by the equation:

$$P = [p_1\, p_2\, p_3 \ldots\ldots\ldots\ldots\, P_{20}\, p_{20+1} \ldots\ldots\ldots\, p_{20+\lambda}]^T \qquad (1)$$

The first 20 elements are frequencies of the 20 native amino acids in protein P, and additional λ factors are used to incorporate some sequence-order information. According to [61], PseAAC for a protein P can be generally formulated as

$$P = [\psi_1\, \psi_2\, \psi_3 \ldots\ldots\ldots\, \Psi_u \ldots\ldots\ \ldots\ldots \psi_\Omega]^T \qquad (2)$$

Where Ω is an integer, the value of Ω and the components $\psi_1,\ \psi_2,\ \psi_3$ depend on how to extract the desired information from the aminoacid sequence P. Equation 2 can represent various different modes of PseAAC when the elements are given by

$$\Psi_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j}, & 1 \leq \mathrm{u} \leq 20 \\[2ex] \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j}, & 20+1 \leq \mathrm{u} \leq 20+\lambda=\Omega;\ \lambda < \mathrm{L} \end{cases} \qquad (3)$$

This method generates (20+λ) dimensional vectors for each protein sequence [62, 63]. We specified λ as 20. Six attributes were taken into consideration i.e. hydrophobicity, hydrophilicity, mass, pK1 (alpha-COOH), pK2 (NH_3) and pI (at 25°C), and a total of 40 descriptor values were computed for each protein.

DV3: Amphiphilic pseudo amino acid composition (APAAC); (also called as series-correlation type). This descriptor vector was generated using PseAAC web server. This method generates (20+ i*λ) dimensional vector, where i is the number of attributes selected [39], and λ is a constant. As for DV2, the value of λ was set to 20 and the same 6 attributes were used. Thus, a total of 140 descriptor values were computed for each protein.

DV4: Di-peptide composition (DPC). This descriptor vector provides information about the fraction of amino acids as well as their local order. A total of 400 descriptor values, one for each possible dipeptide for 20 amino acids, were computed for each protein.

DV5: Normalized Moreau–Broto autocorrelation. Given an AA-index set [64], the normalized Moreau-Broto autocorrelation coefficient for protein sequence is defined by the equation

$$AC(d) = \frac{1}{N-d} \times \sum_{i=1}^{N-d} P(i)P(i+d) \tag{4}$$

where N is the length of the amino acid sequence, d is the lag of autocorrelation i.e distance in the number of residues separated in the protein sequence, and P(*i*), P(*i*+*d*) are the amino acid property at position *i* and *i* + *d*, respectively. The value of *d* must be less than the number of amino acid residues in the shortest protein chain in the dataset. The value of *d* was set to 30 in this study. Amino acid properties used were eight types of amino acid indices [65, 66] *i.e.* hydrophobicity scale, average flexibility index, polarizability parameter, free energy of amino acid solution in water, residue accessible surface area, amino acid residue volume, steric parameters and relative mutability. This descriptor, computed using PROFEAT, thus consisted of a total of 30×8 = 240 descriptor values.

DV6: Moran autocorrelation descriptors. These are defined by the equation

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(Pi-\bar{P})(Pi+d-\bar{P})}{\frac{1}{N}\sum_{i=1}^{N}(Pi-\bar{P})^2} \tag{5}$$

Where *N, d, P i* and *P i + d* have the same meanings as defined for Normalized Moreau-Broto autocorrelation. It also consists of 240 descriptor values.

DV7: Geary autocorrelation descriptors can be defined as

$$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(Pi-Pi+d)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(Pi-\bar{P})^2} \tag{6}$$

Where *N, d, P i* and *P i + d*are similar to those above. It also used the same 8 properties and consists of a total of 240 descriptor values.

DV8: Composition (C), Transition (T) and Distribution (D). These three descriptors are used to describe the global composition of a particular amino acid property in a protein, the percent frequencies with which the attribute changes its index along the entire length of the protein and the distribution pattern of the attribute along the sequence, respectively [67]. Thus, 21 attributes (3 for C, 3 for T and 15 for D) for each of these three attributes were generated using PROFEAT, leading to a vector with 147 descriptor values (21 for composition, 21 for transition, and 105 for distribution).

DV9: Quasi Sequence Order Descriptors. These descriptor vectors are derived from the Schneider-Wrede physicochemical distance matrix and Grantham chemical distance matrix between the 20 amino acids [68]. A total of 160 descriptor values were computed using PROFEAT, for each protein sequence.

DV10: Total amino acid property (TAAP) descriptor for a property *i* is defined as

$$P_{tot_{(i)}} = \sum_{J=1}^{N} P_j^i \tag{7}$$

Where P_j^i is the property *i* of amino acid R_j and N is the length of the sequence. The Amino Acid Index Database [64] is a database that provides 544 properties associated with each of the 20 aminoacids. Out of the 544 indices, 13 have incomplete data. Therefore, among all 531 features in AA index database, only 216 features relevant to the study of protein sequence, structure and function [69] taken into consideration for the design of prediction model. A total of 216 descriptor values were computed for each protein sequences using PROFEAT.

2.2.2. Hybrid Feature Descriptor Vectors

To assess whether combination of the above descriptor vectors can provide better prediction than individual descriptor vectors, we also constructed our hybrid feature descriptor vector sets as follows:

DV11: This descriptor vector combined three descriptors related to amino acid composition of proteins, namely amino acid composition (DV1), pseudo amino acid composition (DV2) and amphiphilic pseudo amino acid composition (DV3). This vector contained a total of 200 descriptor values (20 for DV1, 40 for DV2 and 140 for DV3) for each protein.

DV12: This descriptor vector combined the three autocorrelation descriptors, namely Normalized Moreau–Broto autocorrelation (DV5), Moran autocorrelation (DV6) and Geary autocorrelation (DV7), and contained a total of 720 descriptor values for each protein.

DV13: This descriptor vector combined composition, transition and distribution descriptors (DV8) and autocorrelation descriptors (DV12). It covered 45 attributes (21 for DV8 and 24 for DV12) with a total of 867 descriptor values for each protein.

DV14: This descriptor vector combined all the ten individual descriptor vectors (DV1 to DV10), and included a total of 279 attributes and 1843 descriptor values for each protein.

2.3. Support Vector Machine

The SVM algorithm optimally classifies objects located in a n-dimensional space into two classes by introducing a ‘n-1’ dimensional separating hyperplane which maximizes the separation between the two data clusters. To achieve a robust discriminative power, the algorithm uses special nonlinear functions called as kernels tricks to transform the input space into a multidimensional vector space. This is done by

(i) allowing some data points to fall on the wrong side of the hyperplane by introducing a user-specified parameter 'C' called regularization parameter that specifies the trade-off between good classification and large margin; (ii) using different kernel types i.e. linear, polynomial, sigmoid, and radial basis functions (RBF); and (iii) using a kernel width parameter γ, that fine tunes the discrimination between two classes in a multi-dimensional space.

Previous experimental results [70, 71] have shown that the RBF kernel performs better than the polynomial and linear kernels. We therefore used SVM with RBF function and the WEKA machine learning package [72] to identify classifiers that could distinguish prokaryotic essential proteins from non-essential proteins based on feature vectors calculated from protein sequences.

2.4. Model Generation

For the generation of an optimal SVM model, two key parameters for the RBF kernel referred to above, namely the regularization parameter '*C*' and the kernel width parameter 'γ', need to be pre-selected. Optimization of both these parameters for each training dataset was done by using the grid search utility programme included in LIBSVM 3.1 package [37]. The grid search was performed using 5-fold cross-validation of each training dataset to optimize *C* and γ. The training datasets were trained by applying LIBSVM classifier implemented in WEKA machine learning tool with optimized *C* and γ determined by the 5-fold cross-validation step. The LIBSVM classifier employed with optimized *C* and γ on the external validation data set to assess the prediction potential of the models.

2.5. Model Evaluations

2.5.1. 10-fold Cross Validation of Training Set

In machine learning, cross-validation is frequently used for evaluating and comparing the performances of different predictive modeling algorithms. Three cross-validation methods are available, viz, independent or external validation test, subsampling or k-fold cross validation test and Jackknife test [73]. Of these, though Jackknife test is the most widely used since it considered as the most unbiased and rigorous method for assessing the performance of a classifier [29, 34, 54-56, 74-79], however, this technique is computationally intensive, and we therefore used the external validation dataset and the 10-fold cross-validation techniques to assess the effectiveness of SVM as the prediction engine [80].

For 10-fold cross-validation, each training dataset of 1,000 feature vectors was partitioned automatically and randomly into 10 equal-sized subsets of 100 each. Nine of these subsets were combined and used as training set and the remaining one subset was used for testing; this cross-validation process was repeated 10 times, with each subset serving once as the test dataset. This cross validation was done using LIBSVM classifier of WEKA machine learning tool. The prediction rate of the SVM classifiers was evaluated using four measures, namely, sensitivity, specificity, overall accuracy and Mathew's correlation coefficient.

2.5.2. Evaluation of External Validation Sets

The test datasets were not used to influence the model buildup process. Instead these were used as external validation sets to assess the true predictive potential of the classifier models developed using the test sets. The first column of the test set indicated the class status and labeled with +1 (for essential protein) or -1 (for non-essential protein). Labels in the test file were used only for comparison with predicted category for identification of accuracy.

2.5.3. Evaluation of Prediction Performance

The above validation processes generated data on numbers of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) objects; these parameters were used to calculate sensitivity, specificity, overall prediction accuracy, F-measure and Matthew's correlation coefficient to assess the overall prediction performance of each SVM model.

2.5.3.1. Sensitivity

Sensitivity measures the proportion of actual positives which are correctly identified.

$$\text{Sensitivity } (Q_p) = TP/(TP + FN) \quad (8)$$

2.5.3.2. Specificity

Specificity measures the proportion of actual negatives which are correctly identified.

$$\text{Specificity } (Q_n) = TN / (TN + FP) \quad (9)$$

2.5.3.3. Accuracy

Accuracy is the proportion of objects correctly identified, and includes either true positives or true negatives, divided by the total number of objects.

$$\text{Accuracy } (Q_a) = (TP + TN)/(TP + TN + FP + FN) \quad (10)$$

2.5.3.4. F-measure

F-measure is a measure of accuracy for binary classification function. It is the harmonic mean of precision (*p*) and recall (*r*).

Precision determines the fraction of results that actually turns out to be positive in the group the classifier has declared as a positive class.

$$\text{Precision } (p) = (TP)/(TP + FP) \quad (11)$$

Recall (*r*) measures the fraction of positive examples correctly predicted by the classifier.

$$\text{Recall } (r) = (TP)/(TP + FN) \quad (12)$$

$$\text{F-measure} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (13)$$

Putting the value of precision and recall in equation 10

$$\text{F-measure} = (2 \times TP)/(2 \times TP + FP + FN) \quad (14)$$

2.5.3.5. Matthew's Correlation Coefficient (MCC)

MCC is considered as a balanced measure of quality of binary classification models built using unbalanced datasets. It takes a value between −1 and +1, with+1 representing a perfect prediction, 0 an average random prediction and −1 an inverse prediction.

$$MCC = (TP \times TN - FP \times FN)/\sqrt{((TP+FN) \times (TP+FP) \times (TN+FP) \times (TN+FN))} \quad (15)$$

2.5.3.6. Root Means Squared Error (RMSE)

RMSE is calculated as the square root of the sum of squared errors in modeling or prediction divided by their corresponding total number.

$$RMSE = \sqrt{((p_{1-} a_1)^2 + ... + (p_n - a_n)^2))/n)} \quad (16)$$

where p_1, p_2...p_n are the predicted value of test instances and a_1, a_2...a_n are the actual values.

2.5.3.7. Area under ROC Curve

Receiver-operating characteristics (ROC) curves assess the relationship between sensitivity and specificity in a two-dimensional plane. Area under an ROC curve (AUC) [81] is a sensitive measure of classifier performance. The values of AUC range between 0 and 1, with 1 indicating a perfect classification and 0.5 representing random guessing. AUC values between 0.8 and 0.9 are often considered as good, and those between 0.9 and 1.0 as excellent.

2.5.3.8. Kappa Statistic

Kappa statistic [56] is used to measure the agreement between predicted values with the values which may be expected by chance. Its values tend to zero when there is no agreement beyond chance and approaches 1.0 for very strong statistical relation between the predicted and true category labels

A good generalizable prediction model should have high values for kappa statistic and AUC, and a small value for RMSE.

3. RESULTS

3.1. Results of 10-fold Cross Validation of Training Sets

Each combination of training set and descriptor vectors had widely different optimum values of C and γ. These are shown in Supplementary Table **1**.

The number of true positive, true negative, false positive, and false negative sequences predicted during the 10-fold cross-validation of each training set are shown in Supplementary Table **2**. The number of essential protein sequences predicted correctly (true positives) was the highest using hybrid autocorrelation vector (DV8), whereas the number of non-essential protein sequences predicted correctly (true negatives) was the highest using combined hybrid autocorrelation vector and composition, transition, distribution vector (DV13) (Supplementary Table **2**). Thus, DV8 had an average sensitivity of 94.4% and DV13 had an average specificity of

Table 1. Performance Characteristics of Prediction Using Sensitivity (Q_p), Specificity (Q_n), Accuracy (Q_a), Kappa Statistic (K), Root Mean Square Error (RMSE), Area Under Receiver Operating Characteristic Curve (AUC), F-measure (F), Matthew's Correlation Coefficient (MCC) of 10-fold Cross Validation

Descriptor vector	Q_p	Q_n	Q_a	*K*	RMSE	AUC	F	MCC
DV1	76.6±4.1	72.0±3.5	74.3±2.9	0.49±0.06	0.51±0.03	0.82±0.04	0.75±0.03	0.49±0.06
DV2	79.3±5.1	77.9±6.7	78.6±5.8	0.57±0.12	0.46±0.06	0.86±0.05	0.79±0.06	0.57±0.12
DV3	81.5±6.9	76.1±2.3	78.8±4.6	0.58±0.09	0.46±0.05	0.86±0.05	0.79±0.05	0.58±0.09
DV4	74.0±2.6	74.5±3.9	74.2±1.1	0.48±0.02	0.51±0.01	0.82±0.02	0.74±0.01	0.49±0.02
DV5	63.1±4.4	63.8±2.6	63.4±1.7	0.27±0.03	0.60±0.01	0.69±0.03	0.63±0.03	0.27±0.03
DV6	58.2±2.8	54.6±3.5	56.4±3.1	0.13±0.06	0.66±0.02	0.58±0.04	0.57±0.03	0.13±0.06
DV7	59.8±3.2	57.5±4.0	58.6±3.0	0.17±0.06	0.64±0.02	0.60±0.06	0.59±0.03	0.17±0.06
DV8	94.4±1.2	81.0±2.5	87.7±1.5	0.75±0.03	0.35±0.02	0.93±0.01	0.88±0.01	0.76±0.03
DV9	77.0±7.2	68.7±3.8	72.8±3.5	0.46±0.07	0.52±0.03	0.80±0.04	0.74±0.04	0.46±0.07
DV10	79.6±3.7	70.6±2.2	75.1±2.8	0.50±0.06	0.50±0.03	0.83±0.03	0.76±0.03	0.50±0.06
DV11	79.4±5.7	81.7±3.8	80.5±4.7	0.61±0.09	0.44±0.05	0.88±0.05	0.80±0.05	0.61±0.09
DV12	87.2±2.2	86.3±2.1	86.7±1.0	0.73±0.02	0.36±0.01	0.93±0.01	0.87±0.01	0.73±0.02
DV13	92.9±1.6	89.5±1.9	91.2±1.2	0.82±0.02	0.30±0.02	0.97±0.01	0.91±0.01	0.82±0.02
DV14	84.6±2.5	80.8±0.7	82.7±1.5	0.65±0.03	0.42±0.02	0.90±0.01	0.83±0.02	0.65±0.03

Data are shown as average observed value ± standard deviation.
DV1: Amino acid composition; DV2: Pseudo amino acid composition;
DV3: Amphiphilic pseudo amino acid composition; DV4: Dipeptide composition;
DV5: Normalized Moreau–Broto autocorrelation; DV6: Moran autocorrelation;
DV7: Geray autocorrelation; DV8: Composition, transition and distribution;
DV9: Quasi sequence order; DV10: Total amino acid properties;
DV11: Combination of DV1, DV2 and DV3; DV12: Combination of DV5, DV6 and DV7;
DV13: Combination of DV8 and DV12; DV14: Combination of DV1 to DV10

Table 2. Performance Characteristics of Prediction Using Sensitivity (Q_p), Specificity (Q_n), Accuracy (Q_a), Kappa Statistic (K), Root Mean Square Error (RMSE), Area Under Receiver Operating Characteristic Curve (AUC), F-measure (F), Matthew's Correlation Coefficient (MCC) of External Validation Sets

Descriptor vector	Q_p	Q_n	Q_a	*K*	RMSE	AUC	F	MCC
DV1	52.5±5.9	71.4±3.6	64.1±1.6	0.24±0.04	0.60±0.01	0.62±0.02	0.53±0.04	0.24±0.04
DV2	52.8±6.5	75.3±7.1	66.8±3.1	0.29±0.05	0.58±0.03	0.64±0.02	0.55±0.03	0.29±0.05
DV3	59.3±6.7	73.6±4.8	68.1±1.7	0.33±0.04	0.56±0.02	0.66±0.02	0.59±0.03	0.33±0.04
DV4	51.6±15	75.7±5.5	66.4±2.6	0.28±0.09	0.58±0.02	0.64±0.05	0.53±0.10	0.28±0.08
DV5	54.4±2.6	63.7±2.6	60.1±1.6	0.18±0.03	0.63±0.01	0.59±0.02	0.51±0.02	0.18±0.03
DV6	54.4±1.5	54.4±1.7	54.4±0.6	0.08±0.01	0.68±0.00	0.54±0.00	0.48±0.01	0.09±0.01
DV7	58.7±2.7	56.8±2.8	57.5±0.7	0.15±0.00	0.65±0.01	0.58±0.00	0.51±0.01	0.15±0.00
DV8	88.8±2.9	81.1±2.3	84.1±0.5	0.68±0.01	0.40±0.01	0.85±0.01	0.81±0.01	0.68±0.01
DV9	56.3±8.7	68.1±5.6	63.6±1.9	0.24±0.05	0.60±0.02	0.62±0.03	0.54±0.05	0.24±0.05
DV10	56.6±7.8	69.5±5.0	64.5±0.6	0.26±0.02	0.60±0.01	0.63±0.02	0.55±0.04	0.26±0.02
DV11	56.2±6.6	78.8±5.8	70.1±1.8	0.36±0.03	0.55±0.02	0.67±0.02	0.59±0.03	0.36±0.03
DV12	82.3±1.8	87.6±2.6	85.6±1.2	0.70±0.02	0.38±0.02	0.85±0.01	0.81±0.01	0.70±0.02
DV13	**89.1±1.2**	**90.0±1.1**	**89.6±0.5**	**0.78±0.01**	**0.32±0.01**	**0.90±0.00**	**0.87±0.01**	**0.78±0.01**
DV14	68.1±6.4	80.9±3.3	76.0±1.4	0.49±0.04	0.49±0.01	0.75±0.02	0.68±0.03	0.49±0.03

Data are shown as average observed value ± standard deviation, DV1 to DV14 are as described in footnote to Table **1**

89.5% (Table **1**). Table **1** shows the mean (± standard deviation) of performance characteristics of the prediction achieved using each of the 14 descriptor vectors with the four training sets. Prediction ability of each of the four training sets in terms of 10-fold cross validation accuracy, kappa statistic, RMSE, AUC, F-measure, MCC are shown in Figure **2**. Data on prediction accuracy of individual training sets are shown in Supplementary Table **3**.

3.2. Evaluation of External Validation Sets

External validation datasets were employed to further evaluate the performance of classification models. Data on the number of true positive, true negative, false positive, and false negative sequences predicted during evaluation of test datasets are shown in Supplementary Table **4**. Data for prediction accuracy using each individual test dataset are shown in Supplementary Table **5**. The average prediction accuracy in terms of kappa statistic, RMSE, sensitivity, specificity, accuracy and F-measure (Fig. **3**) obtained from the evaluation of test sets (Table **2**) was quite similar to the results of 10-fold cross validation; this consistency of predictive ability indicates the validity of our SVM classification models.

3.3. Ranking of Descriptor Sets Based on Different Model Evaluation Parameters

Predictive potential of classifier models generated using different protein descriptor vector sets were ranked based on different model evaluation parameters. Of the 10 individual descriptor vector sets (DV1 to DV10), the composition, transition and distribution descriptor set (DV8) showed the best classification accuracy in the 10-fold cross-validation of training datasets (87.7±1.6%; Table **1**) as well as for prediction using the external validation datasets (84.1±0.5%; Table **2**). DV11, which is a composite descriptor vector set based on amino acid composition (DV1), pseudo amino acid composition (DV2) and amphiphilic pseudo amino acid composition (DV3), provided better classification accuracy than that of its three components taken individually. The three autocorrelation descriptor sets (DV5-DV7), make use of the same eight physico-chemical properties but differ in the underlying correlation algorithm. Combination of these three autocorrelation descriptor sets (DV12) provided a better classification accuracy (86.9±1.2%) than that of individual autocorrelation descriptor sets. The best classification accuracy (91.2±1.2)% was achieved using DV13, i.e. a combination of composition, transition and distribution descriptor vector set (DV8) and hybrid autocorrelation descriptor set (DV12).

3.4. Result of External Validation Using Yeast Proteins

To assess the generalizability of the predictive model identified using prokaryotic proteins, we tested the most effective model i.e. the model based on DV13, on an entirely different dataset, i.e. yeast proteins.

Performance characteristics including sensitivity (Q_p), specificity (Q_n), accuracy (Q_a), kappa statistic (K), root mean square error (RMSE), area under ROC, F-measure (F), Matthew's correlation coefficient (MCC) of DV13 for essential vs. non-essential proteins using the yeast protein training dataset in 10-fold cross-validation experiments are shown in (Fig. **4a**). Similar data for a test dataset of yeast protein are shown in (Fig. **4b**). Table **4** shows the comparison of results

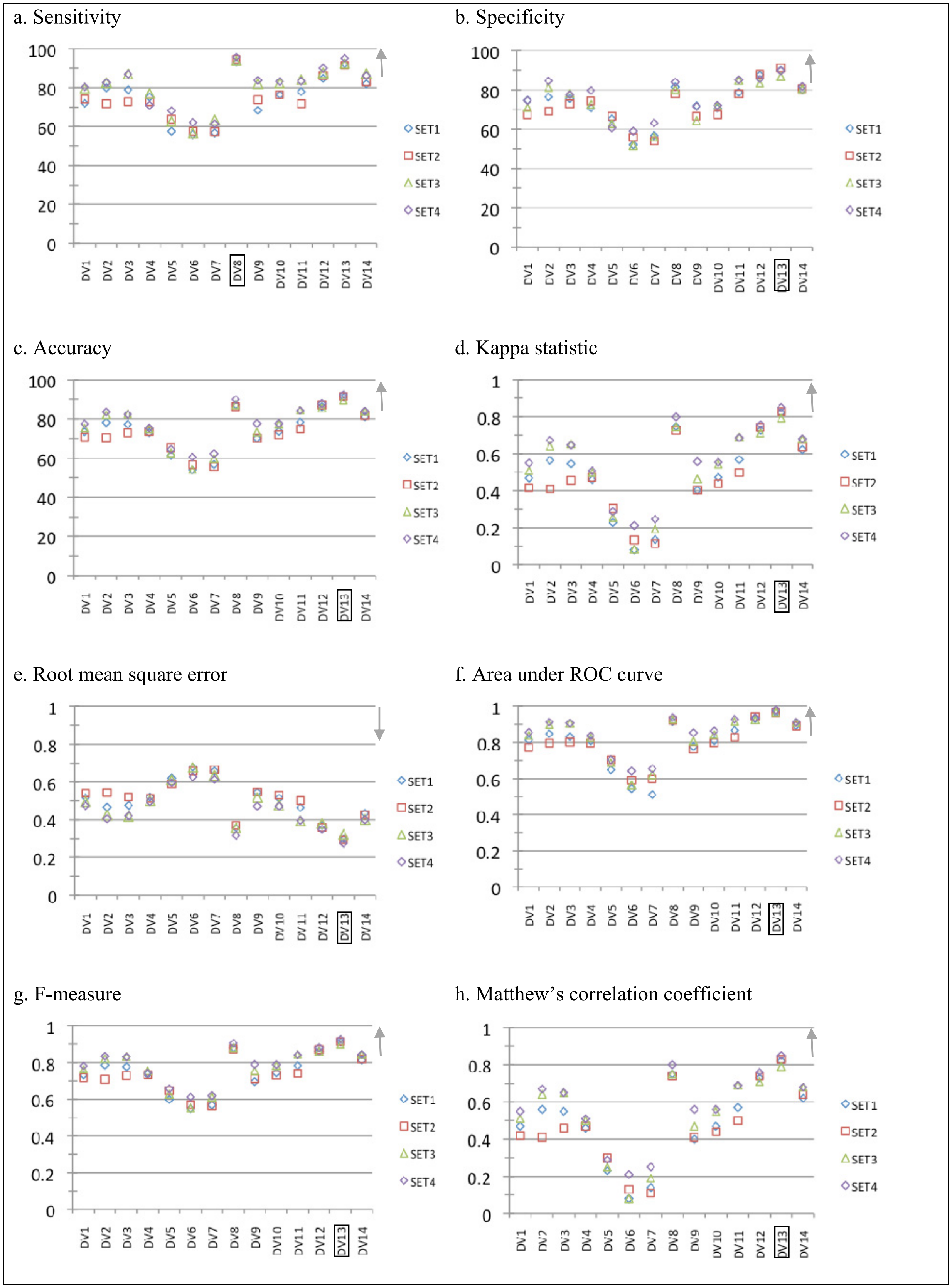


Figure 2. Prediction performance of various SVM models on training dataset evaluated using 10-fold cross-validation. The upward arrows indicate that higher values are more desirable and downward arrow indicates that lower values are more desirable. SET1- SET4 represents four different training sets.

Table 3. Ranking of descriptor sets based on different model evaluation parameters

Evaluation parameter	Average predictive potential of descriptor sets ranked by descending order		
	Training Set		Test Set
Sensitivity	DV8, DV13, DV12, DV14		DV13, DV8, DV12, DV14
Specificity	DV13, DV12, DV11, DV8		DV13, DV12, DV8, DV14
Accuracy	DV13, DV8, DV12, DV14		DV13, DV12, DV8, DV14
Kappa statistic	DV13, DV8, DV12, DV14		DV13, DV12, DV8, DV14
RMSE	DV13, DV8, DV12, DV14		DV13, DV12, DV8, DV14
AUC	DV13, DV8, DV12, DV14		DV13, DV8, DV12, DV14
F-measure	DV13, DV8, DV12, DV14		DV13, DV8, DV12, DV14
MCC	DV13, DV8, DV12, DV14		DV13, DV12, DV8, DV14

RMSE, AUC, MCC, DV1 to DV14 are as described in heading of Table **1**.

Table 4. Prediction Performance Comparison

Evaluation	Parameter	Hwang *et al.*	Acencio *et al.*	Yang *et al.*	Gustafson *et al.*	Our Model
10-fold cross-validation	AUC	0.78	0.72	0.85	0.78±0.05	0.91±0.2
	Precision	0.77	0.70	0.77	0.72±0.03	0.87±0.03
	Recall	0.70	0.71	0.77	0.78±0.04	0.96±0.01
	F-measure	0.74	0.70	0.77	0.75±0.00	0.91±0.2
	MCC	0.50	0.40	0.55	0.48±0.02	0.82±0.04
External validation	AUC	0.79	0.72	0.85	0.82±0.03	0.91±0.01
	Precision	0.74	0.66	0.79	0.90±0.01	0.72±0.02
	Recall	0.33	0.17	0.76	0.74±0.05	0.95±0.03
	F-measure	0.45	0.27	0.77	0.81±0.02	0.82±0.02
	MCC	0.42	0.28	0.55	0.50±0.02	0.75±0.03

AUC, MCC are as described in heading of Table **1**

of 10-fold cross-validation for yeast training dataset (balanced dataset) and prediction performance of external validation set (imbalanced dataset), with those reported in previous studies [82]

3.5. Prediction of Yeast Proteins

The performance characteristics of SVM for predicting essential versus non-essential yeast proteins [sensitivity (Q_p), specificity (Q_n), accuracy (Q_a), kappa statistic (K), root mean square error (RMSE), area under ROC, F-measure (F), Matthew's correlation coefficient (MCC)] in 10-fold cross-validation experiment are shown in (Fig. **5a**). Similar data for test dataset are shown in (Fig. **5b**).

DISCUSSION

Machine learning algorithms have proved to be useful for prediction of essential vs non-essential nature of a protein [82-84]. Most of the previous attempts at prediction of essential genes or proteins were done using the essential protein dataset of yeast, i.e. an eukaryote [82-84]. The yeast essential proteins dataset (n=1,110) is relatively smaller as compared to the prokaryotic essential protein dataset (n=7,643), which contains information on essential proteins for 20 bacterial species. Thus the prediction of essential vs. non-essential proteins in prokaryotes may be expected to be a more challenging task than that for yeasts. In this study, we developed an SVM model for the prediction of 'prokaryotic essential

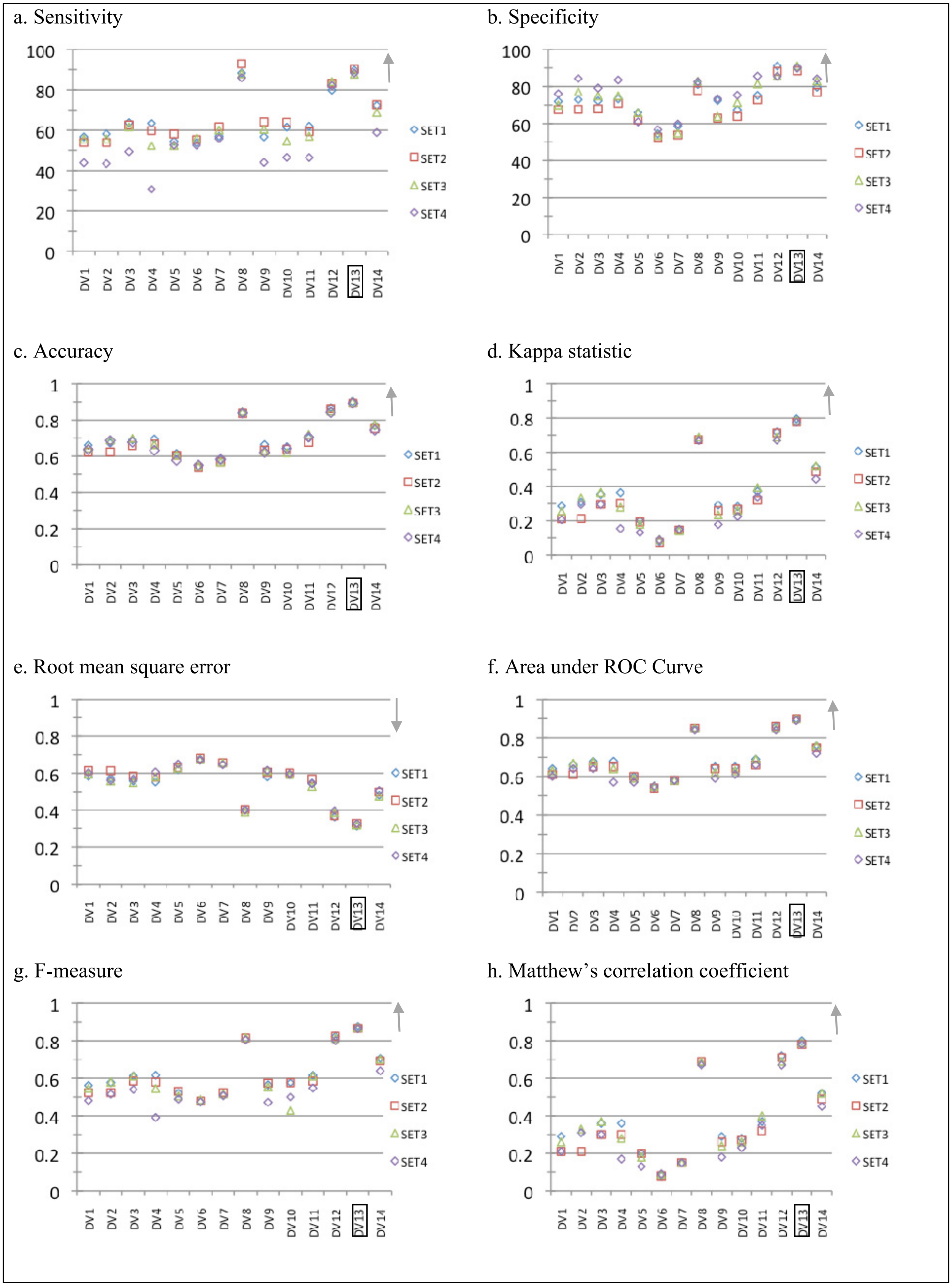
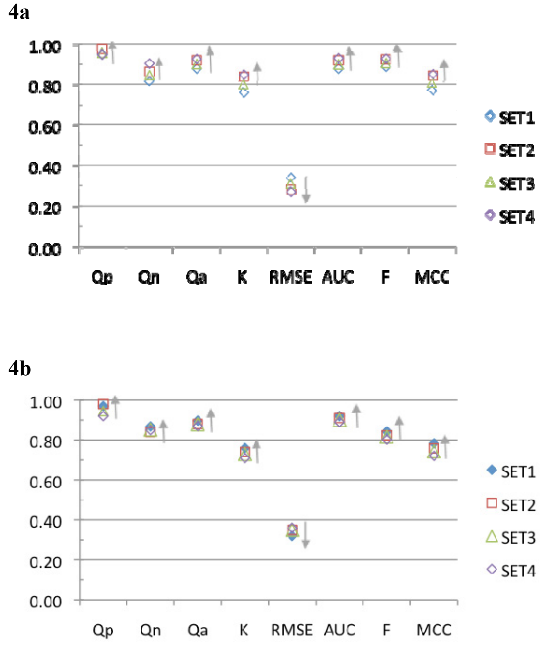
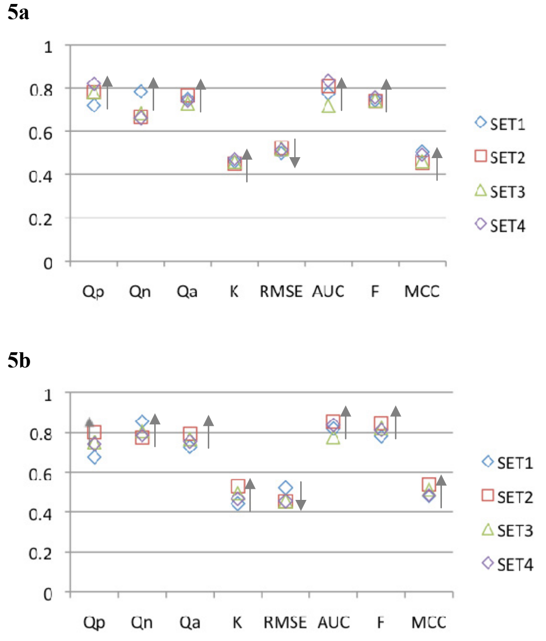


Figure 3. Prediction performance of various SVM models on test dataset evaluated using external validation set. The upward arrow indicates that higher values are more desirable and downward arrow indicates that lower values are more desirable.SET1-SET4 represents four different external validation sets.



The upward arrow indicates that higher values are more desirable and downward arrow indicates that lower values are more desirable.

Figure 4. Performance characteristics of prediction using sensitivity (Q_p), specificity (Q_n), accuracy (Q_a), kappa statistic (K), root mean square error (RMSE), area under receiver operating characteristic curve (AUC), F-measure (F), Matthew's correlation coefficient (MCC) of Yeast training dataset (**4a**) and Yeast test dataset (**4b**). SET1–SET4 (**4a**) represents four different Yeast training dataset each containing 250 positive and 250 negative vectors and SET1-SET4 (**4b**) represents four different Yeast test dataset each containing 200 positive and 500 negative vectors.



The upward arrow indicates that higher values are more desirable and downward arrow indicates that lower values are more desirable.

Figure 5. Performance characteristics of prediction using sensitivity (Q_p), specificity (Q_n), accuracy (Q_a), kappa statistic (K), root mean square error (RMSE), area under receiver operating characteristic curve (AUC), F-measure (F), Matthew's correlation coefficient (MCC) of Yeast training dataset (Gustafson *et al.*) (**5a**) and Yeast test dataset (Gustafson *et al.*)(**5b**). SET1–SET4 (5a) represents four different Yeast training dataset each containing 250 positive and 250 negative vectors and SET1-SET4 (**5b**) represents four different Yeast test dataset each containing 200 positive and 500 negative vectors.

proteins' based solely on the available amino acid sequence information, and tested the predictive ability of this approach using a large number of prokaryotic proteins that are known to be essential and those that may be expected to be non-essential. The SVM algorithm was found to have a high degree of accuracy in predicting essential versus non-essential nature of proteins in both internal cross-validation and external validation using the prokaryotic dataset. More importantly we found that the descriptor vector that performed the best for prediction of essential vs. non-essential prokaryotic proteins, also performed very well for similar prediction among yeast proteins, lending further credibility to our approach.

Several computer-based approaches have been used for identification of essential proteins. Thus, Acencio and Lemke [84] used a J48 algorithm, which integrated data on network topological feature, cellular localization and biological process of each protein and used these for prediction of essential nature of genes. They found that on integrated classifier, that contained data on 12 topological features, 5 cellular localization and 6 biological processes was able to predict essential vs. non-essential nature of the gene product with an AUC of 0.808. Also the integrated classifier performed significantly better than individual predictor. In another study, Hwang, Lin, Chang, Mori, Juan and Huang [83] used an SVM algorithm for essential gene identification in yeast based on network and sequence analysis. They used 14 different attributes to construct SVM classification models to predict gene essentiality. They found the integrated classifier comprising network information with sequence data boosted the performance of the prediction model. Yang, Yangy, Tseng and Ma [82] used SVM to predict essential proteins in a protein-protein network. These workers used a classifier with 45 attributes that included 90 descriptor values and compared its performance with the method used by Acencio and Lemke [84], and Hwang, *et al.* [83]. However, prediction of gene essentiality based on network topology is restricted to organisms whose integrated molecular network has already been constructed [85]. Such data, which need collection of extensive experiments, are not available for several organisms, in particular those that have been recently sequenced.

Gustafson *et al*. [37], in contrast, used naïve Bayes classifiers for essential gene identification in yeast and *E. coli* based on integrated genomic and protein features, without using the data about protein-protein interaction. This approach has the advantage that it can be used even for organisms for which either no or limited protein-protein interaction data are available. Because our study is directed at prediction of protein essentiality in less-studied organisms, we focused on the use of aminoacid physico-chemical property based DVs that can be easily generated without prior extensive laboratory data. Of course, addition of network topology information, when this becomes available, to the DVs that we used would surely result in better performance.

Our data show that a vector set based on combination of composition, transition, distribution (CTD) and hybrid autocorrelation (DV13) comprising of 45 attributes, including 867 descriptor values had the best predictive ability. For instance, we found the AUC, precision, recall, F-measure and MCC using DV13 of 0.91±0.02, 0.87±0.03, 0.96±0.01, 0.91±0.02 and 0.82±0.04, respectively, for internal 10-fold cross-validation. Even for external validation dataset, these parameters were estimated to be very good, i.e. 0.91±0.01, 0.72±0.02, 0.95±0.03, 0.82±0.02 and 0.75±0.03, respectively. These values are better than those achieved by classifier used in previous studies. These performance measures of our model in the training dataset and external validation sets are better than those reported. This suggests that our model may provide a better prediction of essential proteins than the methods available previously. It would be useful to test our model using essential protein databases for other organisms, as these become available. This should help determine the robustness of our model.

For any predictive model, the performance is best when it is used for the data from which it is derived. This phenomenon is termed over-fitting. Thus, a stiffer test for the performance of a predictor system is the assessment of its performance using an external dataset. The ability of the vector set that we found useful for prokaryotes to provide good prediction even for an unrelated yeast protein dataset provides an important external validation of our approach and confirms its robustness.

In our study, the use of hybrid feature descriptor sets tended to provide a better accuracy in classification of proteins into essential and non-essential than that achieved with the use of individual feature descriptor sets. However, at times, use of a combination of features did not provide better results than those obtained with its subsets; for instance, use of DV14, which is the combination of DV1 to DV10, yielded an overall accuracy of only 82.7±1.5%, which is lower than that of DV8, one of its component, alone or of DV12 (composed of DV5, DV6 and DV7). This indicates that there is no clear relationship between the number of vectors used and the predictive ability. Thus, the better performance observed in our study with DV13 was not merely related to the larger number of vectors included in this descriptor vector, but is likely related to the inherent better predictive ability of its components.

4. CONCLUSION

In the present study, a method to predict prokaryotic essential proteins based on SVM has been developed. This method employed a set of 10 physico-chemical descriptor vectors and 4 hybrid descriptor vectors calculated from amino acid sequences using PROFEAT and PseAAC servers. A hybrid descriptor set, DV13, provided the best classification accuracy among the 14 such descriptor sets tried. The DV13 descriptor vector, which is a combination of composition, transition and distribution descriptor set and hybrid autocorrelation descriptor set, provided an accuracy (91.2±1.2)% in 10-fold internal cross-validation and of (89.7±0.5)% in external validation using the prokaryotic protein datasets, and of (91.8±2)% and (88.1±1.1)% using a different yeast protein dataset. We believe that this prediction approach can be used for identification of novel essential proteins in prokaryotes. Since such essential proteins provide opportunities to disrupt the prokaryotic cell function, the SVM approach may provide a useful method for rapid screening of whole proteomes of various pathogens for potential drug targets in future. Going by the current trend towards construction of open source web servers [86], we propose to design a web server for the method presented in this paper for use in future by other workers involved in developing more promising prediction systems.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

Biomedical Informatics Centre at Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow is supported by a research grant from Indian Council of Medical Research, New Delhi. ANS was also supported during this work by this grant.

REFERENCES

[1] Sakharkar, K.R.; Sakharkar, M.K.; Chow, V.T. Biocomputational strategies for microbial drug target identification. *Methods Mol. Med*., **2008**, *142*, 1-9.

[2] Galperin, M.Y.; Koonin, E.V. Searching for drug targets in microbial genomes. *Curr. Opin. Biotechnol*., **1999**, *10*(6), 571-8.

[3] Sakharkar, K.R.; Sakharkar, M.K.; Chow, V.T. A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa. *In Silico Biol*., **2004**, *4*(3), 355-60.

[4] Zhang, R.; Lin, Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res*., **2009**, *37*(Database issue), D455-8.

[5] Rost, B. Twilight zone of protein sequence alignments. *Protein Eng*., **1999**, *12*(2), 85-94.

[6] Teichert, F.; Minning, J.; Bastolla, U.; Porto, M. High quality protein sequence alignment by combining structural profile prediction and profile alignment using SABER-TOOTH. *BMC Bioinform*., **2010**, *11*, 251.

[7] Ben-Hur, A.; Ong, C.S.; Sonnenburg, S.; Scholkopf, B.; Ratsch, G. Support vector machines and kernels for computational biology. *PLoS Comput. Biol*., **2008**, *4*(10), e1000173.

[8] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem*., **2001**, *26*(1), 5-14.

[9] Bakheet, T.M.; Doig, A.J. Properties and identification of human protein drug targets. *Bioinformatics*, **2009**, *25*(4), 451-7.

[10] Bakheet, T.M.; Doig, A.J. Properties and identification of antibiotic drug targets. *BMC Bioinform*., **2010**, *11*, 195.

[11] Huang, C.; Zhang, R.; Chen, Z.; Jiang, Y.; Shang, Z.; Sun, P.; Zhang, X.; Li, X. Predict potential drug targets from the ion channel proteins based on SVM. *J. Theor. Biol*., **2010**, *262*(4), 750-6.

[12] Li, Q.; Lai, L. Prediction of potential drug targets based on simple sequence properties. *BMC Bioinform*., **2007**, *8*, 353.

[13] Hyun, B.R.; Jung, H.; Jang, W.H.; Jung, S.H.; Han, D.S. Weighted feature value based Drug Target Protein prediction. *Int. J. Comput. Biol. Drug. Des*., **2008**, *1*(4), 422-33.

[14] Zernov, V.V.; Balakin, K.V.; Ivaschenko, A.A.; Savchuk, N.P.; Pletnev, I.V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci*., **2003**, *43*(6), 2048-56.

[15] Pirooznia, M.; Yang, J.Y.; Yang, M.Q.; Deng, Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, **2008**, *9 Suppl 1*, S13.

[16] Cai, Y.D.; Liu, X.J.; Xu, X.B.; Chou, K.C. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. *J. Cell. Biochem*., **2002**, *84*(2), 343-8.

[17] Bhasin, M.; Raghava, G.P. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res*., **2004**, *32*(Web Server issue), W414-9.

[18] Gao, Q.B.; Wang, Z.Z.; Yan, C.; Du, Y.H. Prediction of protein subcellular location using a combined feature of sequence. *FEBS Lett*., **2005**, *579*(16), 3444-8.

[19] Liao, B.; Jiang, J.B.; Zeng, Q.G.; Zhu, W. Predicting Apoptosis Protein Subcellular Location with PseAAC by Incorporating Tripeptide Composition. *Protein Pept. Lett*., **2011**, *18*(11), 1086-92.

[20] Cai, C.Z.; Han, L.Y.; Ji, Z.L.; Chen, X.; Chen, Y.Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res*., **2003**, *31*(13), 3692-7.

[21] Han, L.Y.; Cai, C.Z.; Ji, Z.L.; Cao, Z.W.; Cui, J.; Chen, Y.Z. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res*., **2004**, *32*(21), 6437-44.

[22] Tang, Z.Q.; Lin, H.H.; Zhang, H.L.; Han, L.Y.; Chen, X.; Chen, Y.Z. Prediction of functional class of proteins and peptides irrespective of sequence homology by support vector machines. *Bioinform. Biol. Insights*, **2009**, *1*, 19-47.

[23] Babnigg, G.; Joachimiak, A. Predicting protein crystallization propensity from protein sequence. *J. Struct. Funct. Genomics*, **2010**, *11*(1), 71-80.

[24] Kandaswamy, K.K.; Pugalenthi, G.; Suganthan, P.N.; Gangal, R. SVMCRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence. *Protein Pept. Lett*., **2010**, *17*(4), 423-30.

[25] Mizianty, M.J.; Kurgan, L.A. CRYSpred: Accurate sequence-based protein crystallization propensity prediction using sequence-derived structural characteristics. *Protein Pept. Lett*., **2012**, *19*(1), 40-9.

[26] Xiaohui, N.; Nana, L.; Feng, S.; Xuehai, H.; Jingbo, X.; Huijuan, X. Predicting protein solubility with a hybrid approach by pseudo amino acid composition. *Protein Pept. Lett*., **2010**, *17*(12), 1466-72.

[27] Aziz, M.M.; Maleki, M.; Rueda, L.; Raza, M.; Banerjee, S. Prediction of biological protein-protein interactions using atom-type and amino acid properties. *Proteomics*, **2011**, *11*(19), 3802-10.

[28] Zellner, H.; Staudigel, M.; Trenner, T.; Bittkowski, M.; Wolowski, V.; Icking, C.; Merkl, R. Prescont: Predicting protein-protein interfaces utilizing four residue properties. *Proteins*, **2011**, *80*(1), 154-68.

[29] Chen, W.; Lin, H.; Feng, P.M.; Ding, C.; Zuo, Y.C.; Chou, K.C. iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. *PloS ONE*, **2012**, *7*(10), e47843.

[30] Huang, T.; Chen, L.; Cai, Y.D.; Chou, K.C. Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. *PloS ONE*, **2011**, *6*(9), e25297.

[31] Li, B.Q.; Hu, L.L.; Chen, L.; Feng, K.Y.; Cai, Y.D.; Chou, K.C. Prediction of protein domain with mRMR feature selection and analysis. *PloS ONE*, **2012**, *7*(6), e39308.

[32] Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **2006**, *22*(13), 1658-9.

[33] Sonnhammer, E.L.; Eddy, S.R.; Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **1997**, *28*(3), 405-20.

[34] Hayat, M.; Khan, A. Discriminating outer membrane proteins with Fuzzy K-nearest Neighbor algorithms based on the general form of Chou's PseAAC. *Protein Pept. Lett*., **2012**, *19*(4), 411-21.

[35] Chen, X.; Fang, Y.; Yao, L.; Chen, Y.; Xu, H. Does drug-target have a likeness? *Methods Inf. Med*., **2007**, *46*(3), 360-6.

[36] Chen, L.; Zeng, W.M.; Cai, Y.D.; Feng, K.Y.; Chou, K.C. Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PloS ONE*, **2012**, *7*(4), e35254.

[37] Gustafson, A.M.; Snitkin, E.S.; Parker, S.C.; DeLisi, C.; Kasif, S. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics*, **2006**, *7*, 265.

[38] Rao, H.B.; Zhu, F.; Yang, G.B.; Li, Z.R.; Chen, Y.Z. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid

sequence. *Nucleic Acids Res*., **2011**, *39*(Web Server issue), W385-90.

[39] Shen, H.B.; Chou, K.C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem*., **2008**, *373*(2), 386-8.

[40] Chen, C.; Chen, L.X.; Zou, X.Y.; Cai, P.X. Predicting protein structural class based on multi-features fusion. *J. Theor. Biol*., **2008**, *253*(2), 388-92.

[41] Karnik, S.; Mitra, J.; Singh, A.; Kulkarni, B.D.; Sundarajan, V.; Jayaraman, V.K., Identification of N-Glycosylation Sites with Sequence and Structural Features Employing Random Forests. In: *Pattern Recognition and Machine Intelligence*, Springer Berlin Heidelberg: **2009**; Vol. 5909, pp 146-151.

[42] Hosseinzadeh, F.; Ebrahimi, M.; Goliaei, B.; Shamabadi, N. Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. *PloS ONE*, **2012**, *7*(7), e40017.

[43] Yu, W.; Jiang, Z.; Wang, J.; Tao, R. Using Feature Selection Technique for Drug-Target Interaction Networks Prediction. *Curr. Med. Chem*., **2011**, *18*(36), 5687-5693.

[44] Wassermann, A.M.; Geppert, H.; Bajorath, J. Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. *J. Chem. Inf. Model*., **2009**, *49*(10), 2155-2167.

[45] Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **2001**, *43*(3), 246-255.

[46] Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theor. Biol*., **2008**, *252*(2), 350-356.

[47] Mohammad Beigi, M.; Behjati, M.; Mohabatkar, H. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genomics*, **2011**, *12*(4), 191-7.

[48] Guo, J.; Rao, N.; Liu, G.; Yang, Y.; Wang, G. Predicting protein folding rates using the concept of Chou's pseudo amino acid composition. *J. Comput. Chem*., **2011**, *32*(8), 1612-7.

[49] Zhao, X.W.; Ma, Z.Q.; Yin, M.H. Predicting protein-protein interactions by combing various sequence- derived features into the general form of Chou's Pseudo amino acid composition. *Protein Pept. Lett*., **2012**, *19*(5), 492-500.

[50] Wang, X.; Li, G.Z.; Lu, W.C. Virus-ECC-mPLoc: A Multi-Label Predictor for Predicting the Subcellular Localization of Virus Proteins with Both Single and Multiple Sites Based on a General Form of Chou's Pseudo Amino Acid Composition. *Protein Pept. Lett*., **2013**, *20*(3, 309-17.

[51] Niu, X.H.; Hu, X.H.; Shi, F.; Xia, J.B. Predicting protein solubility by the general form of Chou's pseudo amino acid composition: approached from chaos game representation and fractal dimension. *Protein Pept. Lett.*, **2012**, *19*(9), 940-8.

[52] Mohabatkar, H.; Beigi, M.; Abdolahi, K.; Mohsenzadeh, S. Prediction of allergenic proteins by means of the concept of chou's pseudo amino acid composition and a machine learning approach. *Med. Chem*., **2013**, *9*(1), 133-7.

[53] Zhao, X.W.; Li, X.T.; Ma, Z.Q.; Yin, M.H. Identify DNA-binding proteins with optimal Chou's amino acid composition. *Protein Pept. Lett*., **2012**, *19*(4), 398-405.

[54] Zia Ur, R.; Khan, A. Identifying GPCRs and their types with Chou's pseudo amino acid composition: an approach from multi-scale energy representation and position specific scoring matrix. *Protein Pept. Lett.*, **2012**, *19*(8), 890-903.

[55] Mohabatkar, H.; Mohammad Beigi, M.; Esmaeili, A. Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol*., **2011**, *281*(1), 18-23.

[56] Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol*., **2010**, *263*(2), 203-9.

[57] Sun, X.; Shi, S.; Qiu, J.; Suo, S.; Huang, S.; Liang, R. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol. Biosyst*., **2012**, *8*(12), 3178-84.

[58] Nanni, L.; Lumini, A.; Gupta, D.; Garg, A. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. *IEEE/ACM. Trans. Comput. Biol. Bioinform*., **2012**, *9*(2), 467-475.

[59] Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem*., **2012**, *425*(2), 117-9.

[60] Bhasin, M.; Raghava, G.P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem*., **2004**, *279*(22), 23262-6.

[61] Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol*., **2011**, *273*(1), 236-47.

[62] Lin, H.; Ding, H.; Guo, F.B.; Zhang, A.Y.; Huang, J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein Pept. Lett*., **2008**, *15*(7), 739-44.

[63] Nanni, L.; Brahnam, S.; Lumini, A. High performance set of PseAAC and sequence based descriptors for protein classification. *J. Theor. Biol*., **2010**, *266*(1), 1-10.

[64] Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*., **2008**, *36*(Database issue), D202-5.

[65] Feng, Z.P.; Zhang, C.T. Prediction of membrane protein types based on the hydrophobic index of amino acids. *J. Protein. Chem*., **2000**, *19*(4), 269-75.

[66] Lin, Z.; Pan, X.M. Accurate prediction of protein secondary structural content. *J. Protein. Chem*., **2001**, *20*(3), 217-20.

[67] Dubchak, I.; Muchnik, I.; Holbrook, S.R.; Kim, S.H. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA*, **1995**, *92*(19), 8700-4.

[68] Chou, K.C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun*., **2000**, *278*(2), 477-83.

[69] Mathura, V.S.; Kolippakkam, D. APDbase: Amino acid physico-chemical properties database. *Bioinformation*, **2005**, *1*(1), 2-4.

[70] Habib, T.; Zhang, C.; Yang, J.Y.; Yang, M.Q.; Deng, Y. Supervised learning method for the prediction of subcellular localization of proteins using amino acid and amino acid pair composition. *BMC Genomics*, **2008**, *9 Suppl 1*, S16.

[71] Wang, J.; Chen, Q.; Chen, Y., RBF Kernel Based Support Vector Machine with Universal Approximation and Its Application. Advances in Neural Networks – ISNN 2004. Yin, F.-L.; Wang, J.; Guo, C., Eds. Springer Berlin / Heidelberg: **2004**; Vol. 3173, pp 512-517.

[72] Zheng, Z.; Kramer, S.; Schmidt, B. DySC: Software for greedy clustering of 16S rRNA reads. *Bioinformatics*, **2012**, *28*(16), 2182-3.

[73] Chou, K.C.; Zhang, C.T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **1995**, *30*(4), 275-349.

[74] Ren, L.Y.; Zhang, Y.S.; Gutman, I. Predicting the Classification of Transcription Factors by Incorporating their Binding Site Properties into a Novel Mode of Chou's Pseudo Amino Acid Composition. *Protein Pept. Lett*., **2012**, *19*(11), 1170-6.

[75] Sun, X.Y.; Shi, S.P.; Qiu, J.D.; Suo, S.B.; Huang, S.Y.; Liang, R.P. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol. Biosyst*., **2012**, *8*(12), 3178-84.

[76] Mohabatkar, H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett*., **2010**, *17*(10), 1207-14.

[77] Cao, J.Z.; Liu, W.Q.; Gu, H. Predicting viral protein subcellular localization with Chou's pseudo amino acid composition and imbalance-weighted multi-label K-nearest neighbor algorithm. *Protein Pept. Lett.*, **2012**, *19*(11), 1163-9.

[78] Chou, K.C.; Wu, Z.C.; Xiao, X. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst*., **2012**, *8*(2), 629-41.

[79] Chou, K.C.; Shen, H.B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc*., **2008**, *3*(2), 153-62.

[80] Martin, S.; Roe, D.; Faulon, J.L. Predicting protein-protein interactions using signature products. *Bioinformatics*, **2005**, *21*(2), 218-26.

[81] Bardley, A.P. Use of the area under ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*., **1997**, *30*, 1145-1159.

[82] Yang, Z.H.; Yangy, C.B.; Tseng, C.T.; Ma, X.H. In *Prediction for essential proteins with the support vector machine*, National Computer Symposium, National Chiayi University, **2011**; pp 26-33.

[83] Hwang, Y.C.; Lin, C.C.; Chang, J.Y.; Mori, H.; Juan, H.F.; Huang, H.C. Predicting essential genes based on network and sequence analysis. *Mol. Biosyst*., **2009**, *5*(12), 1672-8.

[84] Acencio, M.L.; Lemke, N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinform*., **2009**, *10*, 290.

[85] da Silva, J.P.M.; Acencio, M.L.; Mombach, J.C.M.; Vieira, R.; da Silva, J.C.; Lemke, N.; Sinigaglia, M. In silico network topology-based prediction of gene essentiality. *Physica A*, **2008**, *387*(4), 1049-1055.

[86] Chou, K.C.; Shen, H.B. Review: recent advances in developing web-servers for predicting protein attributes. *Nat. Sci*., **2009**, *1*(2), 63-92.

Received: September 25, 2012 Revised: December 24, 2012 Accepted: December 24, 2012