*In Silico* Biology 8, 0012 (2008); ©2008, Bioinformation Systems e.V.

# A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search

**Aarti Garg and Gajendra P. S. Raghava***

Institute of Microbial Technology, Sector 39A, Chandigarh, INDIA

* Corresponding author

Email: raghava@imtech.res.in
Phone: +91-172-2690557; Fax: +91-172-2690632
http://www.imtech.res.in/raghava/

**Edited by H. Michael; received July 23, 2007; revised January 30, 2008; accepted February 11, 2008; published March 20, 2008**

## Abstract

Most of the prediction methods for secretory proteins require the presence of a correct N-terminal end of the pre-protein for correct classification. As large scale genome sequencing projects sometimes assign the 5'-end of genes incorrectly, many proteins are encoded without the correct N-terminus leading to incorrect prediction. In this study, a systematic attempt has been made to predict secretory proteins irrespective of presence or absence of N-terminal signal peptides (also known as classical and non-classical secreted proteins respectively), using machine-learning techniques; artificial neural network (ANN) and support vector machine (SVM). We trained and tested our methods on a dataset of 3321 secretory and 3654 non-secretory mammalian proteins using five-fold cross-validation technique. First, ANN-based modules have been developed for predicting secretory proteins using 33 physico-chemical properties, amino acid composition and dipeptide composition and achieved accuracies of 73.1%, 76.1% and 77.1%, respectively. Similarly, SVM-based modules using 33 physico-chemical properties, amino acid, and dipeptide composition have been able to achieve accuracies of 77.4%, 79.4% and 79.9%, respectively. In addition, BLAST and PSI-BLAST modules designed for predicting secretory proteins based on similarity search achieved 23.4% and 26.9% accuracy, respectively. Finally, we developed a hybrid-approach by integrating amino acid and dipeptide composition based SVM modules and PSI-BLAST module that increased the accuracy to 83.2%, which is significantly better than individual modules. We also achieved high sensitivity of 60.4% with low value of 5% false positive predictions using hybrid module. A web server SRTpred has been developed based on above study for predicting classical and non-classical secreted proteins from whole sequence of mammalian proteins, which is available from http://www.imtech.res.in/raghava/srtpred/.

*Keywords*: classical pathway, non-classical pathway, secretory proteins, prediction, SRTpred, redundancy, dataset size, ANN, SVM, BLAST, PSI-BLAST, N-terminal sequence

## Introduction

After translation of messenger RNA by the ribosome, newly formed proteins must be folded, sorted, and delivered to various intracellular and extracellular destinations. The process of directing each newly made polypeptide to their proper final destination is referred as protein targeting or sorting. The targeting of the proteins to the mitochondrion or the chloroplast requires the recognition of a N-terminal peptide, similar to the signal peptide (SP) of secretory proteins. SPs are typically a stretch of 15-30 residues which directs the proteins to the translocation apparatus of the endoplasmic reticulum (ER) and to finally release them into the extracellular space in eukaryotic cell. This pathway of protein secretion in eukaryotic cells is known as classical or ER/Golgi-dependent secretory pathway. However, not all types of secretory proteins require N-terminal SPs for their release into the extracellular space. This includes fibroblast growth factors (FGF-1, FGF-2), interleukins (IL-1), and galectins. These proteins do not follow the classical route of secretion and are secreted following the alternative pathway known as leaderless secretion or the non-classical/conventional secretory pathway [1-3].

Protein secretion is a universal process which occurs in all organisms and has tremendous importance to biological research. In case of pathogenic microorganisms, secretory pathways deliver virulence factors to their sites of action, soluble extracellular enzymes into the surrounding medium, or for specifically targeting proteins to the host cell. In several instances, protein secretion pathways are similar to those involved in assembly of bacterial appendages. Further, several secretory proteins have been identified as major target proteins for the development of drugs [4]. Hence, development of an automatic method for the prediction of secretory proteins would be a help for studies aiming towards deciphering secretory pathways and would also lead to the identification of novel drug targets with greater value for biomedical research.

Until now, many methods have been developed for the classification and prediction of subcellular localizations of proteins based on SPs, mainly SignalP and pTarget [5, 6]. TargetP is a neural-network based method that discriminates between proteins destined for the mitochondrion, the chloroplast, the secretory pathways and other localizations with a success rate of 85.3% (overall) and sensitivity of 0.96 for non-plant secretory proteins. Whereas, the neural network based method SignalP (version 3.0) has been able to achieve high sensitivity of 0.99 and overall accuracy of 0.93 for eukaryotic signal peptide discrimination. Though achieving higher prediction accuracy for classical secreted proteins, these methods unfortunately fail during the prediction of proteins without SP. Hence, non-classical secreted proteins also demand an automated method for the prediction. Recently, a webserver SecretomeP has been developed to predict non-classical secreted proteins, based on an idea that extracellular proteins share certain features regardless of the pathway used to secrete them [7]. It is a neural network based method that has used several features of a protein such as number of atoms, positively charged residues, propeptide cleavage site, protein sorting, low complexity regions, and transmembrane helices as an input to train network. Despite considering large number of protein features, the method has achieved a false positive prediction that is less than 5% at a low sensitivity value of 40%.

Till date, there is not any method available that can predict secretory proteins, irrespective of pathways/SPs, with better accuracy. In the present study, an attempt has been made to develop an automated method that can predict secretory proteins (irrespective of N-terminal SP) based on different features of whole protein sequence. We tried two powerful machine-learning techniques, Artificial Neural Network (ANN) and Support Vector Machines (SVM), for the prediction of proteins destined for secretions. One of the other limitations of existing methods is that they are either similarity based or are solely based on machine learning techniques. In this study, we have integrated both types of approaches in order to achieve a higher accuracy. In addition, we have also analyzed the effect of the presence of similar (redundancy) and of the number of proteins (dataset size) on the prediction performance of ANN and SVM based modules developed in the present study by creating different dataset types.

## Methods

### Data set

The data set used in the present study, consisted of 6975 mammalian protein sequences. Out of them, 3321

sequences were extracellular proteins secreted via classical and non-classical pathways (positive examples), whereas the remaining 3654 proteins were annotated as cytoplasmic and/or the nuclear (negative examples). Previously, the same dataset was used to develop a method SecretomeP [7] and available publicly at http://www.cbs.dtu.dk/services/SecretomeP-1.0/datasets.php. The sequences were extracted from Swiss-Prot database [8] on the basis of subcellular localization annotations in the comment block.

### Dataset types

In order to analyze the effect of redundancy on the prediction performance, different strategies were adopted to make the dataset non-redundant. In this study we used BLAST [9] to define the cut-off E-values to remove redundancy instead commonly used term percent sequence identity (PID). Though PID gave more meaningful value, in practice it is difficult to calculate PID when the similarity is low between sequences (as it requires meaningful alignment). PID itself has a lot of variation particularly when two sequences have local similarity rather than global similarity [10]. This is the reason why we used BLAST instead of PID for detecting local and low similarity. The brief description of different dataset types created in the present study is as follows:

a) Firstly, the redundancy of the sequences was reduced using BLAST at different cut-off $E$-values, such that no two sequences were having a similarity greater than $E$-value of 8e-4, 8e-10 and 4e-40. This created 3 types of alternative datasets (Alt-DS) designated as Alt-DS1, Alt-DS2 and Alt-DS3, left with 924, 1155 and 1876 proteins, respectively. Further, 5-fold cross-validation technique was adopted separately for each Alt-DS types to evaluate the performance. Here, each data set type was divided randomly into five subsets. The training and testing was carried out five times, each time using one distinct subset for testing and remaining four sets for training. The final prediction results were averaged over the number of subsets.

b) Nevertheless, such strategy of making Alt-DS types reduces the similarity significantly, but left with very few numbers of proteins for training. Hence, in order to retain the complete dataset of 6975 protein sequences, we divided the positive and negative examples equally into five subsets (for five-fold cross validation) in such a way that most 'significant' match between these five subsets had BLAST $E$-value of only 8e-4 (26% identity for one sequence pair). Earlier, the same technique was used for the training of SecretomeP [7]. We called this dataset as 'Main-dataset'.

c) Though the Main-dataset created above is ideal for training, the performance may still be biased due to redundant proteins used for testing. For example if test dataset have cluster of 100 similar proteins and the method is tuned/trained to these proteins, then the increase in performance would be artificially amplified (increase significantly) as it is a big cluster; the reverse also true if the method is not suitable for these proteins then a decrease in performance would be found. In order to overcome this problem, we computed the performance of the method after removing redundant proteins from each test set of Main-dataset using a BLAST cut-off value of 8e-4.

### Neural network architecture

For the neural network implementation and to generate the neural network architecture for the learning process, the publicly available free simulation package SNNS, version 4.2, from Stuttgart University has been used [11]. It allows incorporation of the resulting networks into an ANSI C function for use in stand-alone code. A logistic activation function is used. At the start of each simulation, the weights are initialized with the random values. The training was carried out using error back-propagation with a sum of square error function as well as mean square error function [12]. The learning parameter was set to 0.001. The magnitude of the error sum in the test and training set was monitored after each cycle of training. Ultimately, the number of cycles is determined where the network during training converges.

### Support Vector Machines

In the present study, SVM_light, a freely downloadable package of SVM (http://svmlight.joachims.org/old/svm_light_v4.00.html) [13], was used for the classification of secretory

proteins. The software enables the users to define a number of parameters and also allows a choice of inbuilt kernel functions including linear, RBF (Radial Basis Function) and polynomial functions. The machine learning techniques are more successful if input units/patterns are of fixed length. Therefore, in the present study, different approaches based on different features of a protein such as amino acid composition, composition of physico-chemical properties and dipeptide composition are considered which generated fixed length patterns.

### Composition of physico-chemical properties

The 33 physico-chemical properties (e. g. hydrophobicity, hydrophilicity, polarity) were used to represent the proteins as used recently by our group for the prediction of subcellular localization of eukaryotic and prokaryotic proteins [14, 15]. The values of each physico-chemical property for all the 20 amino acids were normalized between 0 and 1 using the standard conversion formula. The input vector has 33 scalar values, each representing the average value of a distinct physico-chemical property of a protein.

### Amino acid composition

Amino acid composition is the fraction of each amino acid in a protein. The fraction of all the 20 natural amino acids was calculated using Equation 1.

$$\text{Fraction of amino acid } i = \frac{\text{total number of amino acid } i}{\text{total number of amino acids in protein}} \quad (1)$$

where $i$ can be any amino acid.

### Dipeptide composition

Dipeptide compositions (e. g. Ala-Ala, Ala-Leu), which give a fixed pattern length of 400 (20 × 20), encapsulate the global information about each protein sequence. This representation encompasses the information about amino acid composition along with the local order of amino acid. The fraction of each dipeptide was calculated using Equation 2.

$$\text{Fraction of dep}(i) = \frac{\text{total number of dep}(i)}{\text{total number of all possible dipeptides}} \quad (2)$$

where dep($i$) is one out of 400 dipeptide.

### BLAST and PSI-BLAST

In the present study, similarity search based modules were also developed to search a query protein against a database of secretory and non-secretory sequences using BLAST and PSI-BLAST, respectively [9]. The PSI-BLAST was used in addition to normal standard BLAST because it has the capability to detect remote homologies. It carries out an iterative search in which the sequences found in one round of search are used to build score model for the next round of searching. Three iterations of PSI-BLAST were carried out at a cut-off $E$-value of 0.001. Depending upon the similarity of the query protein to the proteins present in the database, this module can classify the proteins and returns "unknown classification" if no significant similarity is obtained. In addition, we also performed BLAST search for Alt-DS1 dataset against the complete Swiss-Prot database using cut-off $E$-value of 0.001. The hits so obtained were then used to generate a new larger database. This new larger database was again used to carry out PSI-BLAST search for Alt-DS1 as described above. The five-fold cross-validation technique was used to assess the performance of similarity-search based module.

### Hybrid SVM module

Previously, hybrid approach based SVM modules have achieved remarkable success for the prediction of

subcellular localization of proteins [14-16]. In the present study, the hybrid module was also constructed integrating information about amino acid composition, dipeptide composition, and PSI-BLAST output. SVM was provided with an input vector of 423 dimensions which includes 20 for amino acids composition, 400 for dipeptide composition and 3 for PSI-BLAST output. The PSI-BLAST output was converted to binary variables using the representation such as -1 0 0 (secretory proteins); 0 1 0 (non-secretory); 0 0 1 (Unknown).

### Evaluation of performance

In order to assess the prediction performances; accuracy, Matthews correlation coefficient (MCC) [17], sensitivity and specificity were calculated using Equations 3, 4, 5 and 6, respectively.

$$accuracy = \frac{p+n}{t} \tag{3}$$

$$MCC = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}} \tag{4}$$

$$sensitivity = \frac{p}{p+u} \tag{5}$$

$$specificity = \frac{n}{n+o} \tag{6}$$

where, $t$ = ($p$ + $n$ + $o$ + $u$) is the total number of predictions, $p$ is the number of correctly classified secretory proteins, $n$ is the number of correctly classified non-secretory proteins, $u$ is the number of under-predicted sequences and $o$ is the number of over-predicted sequences.

## Results and discussion

### Performance using BLAST and PSI-BLAST

Using the initial complete data set of 6975 sequences (3321 positive and 3654 negative examples) without reducing redundancy we first developed similarity search based modules using BLAST and PSI-BLAST. The modules were able to achieve striking high accuracy of 96.4% and 95%, respectively, after five-fold cross validation. If similarity based modules were able to classify the proteins with such a high accuracy, then there seems to be no need to go for another machine-learning technique for predictions. However, this high accuracy was due to the presence of enough similarity between proteins. In that case, these results would be biased one.

In order to remove this bias, we tried to make the Alt-DS types such as Alt-DS1, Alt-DS2 and Alt-DS3 using different cut-off $E$-values of 8e-4, 8e-10, and 4e-40 respectively. These dataset types consisted of 924, 1155 and 1876 proteins, respectively. During 5-fold cross-validation for Alt-DS1, Alt-DS2, and Alt-DS3 separately, BLAST gave the performance of 4.5%, 26.8%, and 66.4%, respectively, as shown in Tab. 1. Further, accuracy increased to 9.5%, 31.2%, and 68.4%, respectively, using PSI-BLAST as it has the tendency to search remote homologues also (Tab. 1). Therefore, we can interpret that initial high accuracy of 95% and 96.4% using BLAST and PSI-BLAST, respectively was due to the presence of highly similar proteins in the dataset. Hence, reducing redundancy also resulted in reduction of accuracy of prediction. Further, for Alt-DS1, accuracy increased to 25% from 9.5%, when it was searched against new database (generated by performing BLAST search of Alt-DS1 against complete Swiss-Prot database) using PSI-BLAST. Hence, it proved that performance of BLAST or PSI-BLAST modules can be increased with increase in number of sequences in the database.

**Table 1:** Accuracy (%) obtained using different ANN-based modules, BLAST and PSI-BLAST for Alt-DS1, Alt-DS2 and Alt-DS3.

| | Inputs | Alt-DS1 | Alt-DS2 | Alt-DS3 |
|---|---|---|---|---|
| **Others** | BLAST | 4.5 | 26.8 | 66.4 |
| | PSI-BLAST | 9.5 | 31.2 | 68.4 |
| **Using ANN** | Properties | 68.5 | 70.6 | 74.1 |
| | Amino acid composition | 70.2 | 72.6 | 75.8 |
| | Dipeptide composition | 76.2 | 79.8 | 83.9 |

Hence, it is obvious that this technique of making datasets non-redundant removed the bias, but it also resulted in the reduction of the number of sequences, which further affected the accuracy. In order to overcome this problem, we adopted a new technique to make the dataset non-redundant, while retaining the complete dataset. Using this Main-dataset, first we developed BLAST and PSI-BLAST based modules and an accuracy of 23.4% and 26.9% was attained, respectively, carrying out five-fold cross-validation (Tab. 2). Hence, this technique of retaining the complete dataset while having non-redundancy between different sets yielded better performance in comparison to Alt-DS types.

**Table 2:** Detailed results obtained using different ANN and SVM-based modules, Blast and PSI-BLAST for the Main-dataset.

| | Inputs | Sensitivity (%) | Specificity (%) | Accuracy (%) | MCC |
|---|---|---|---|---|---|
| **Using ANN** | Properties | 73.0 | 73.2 | 73.1 | 0.46 |
| | Amino acid composition | 69.0 | 82.5 | 76.1 | 0.52 |
| | Dipeptide composition | 70.0 | 83.4 | 77.1 | 0.54 |
| **Using SVM** | Properties | 74.7 | 80.1 | 77.4 | 0.60 |
| | Amino acid composition | 76.2 (74.4*) | 82.6 (82.1) | 79.4 (76.8%) | 0.59 (0.57) |
| | Dipeptide composition | 77.0 (75.2) | 82.2 (82.6) | 79.9 (78.7%) | 0.59 (0.58) |
| **Similarity search** | BLAST | 22.4 | 30.9 | 23.4 | ----- |
| | PSI-BLAST | 20.2 | 26.3 | 26.9 | ----- |
| **Hybrid (Amino acid + Dipeptide + PSI-BLAST)** | | 78.9 (78.3) | 87.1 (85.4) | 83.2 (81.8) | 0.66 (0.65) |

* Value in bracket is performance of method after removing redundancy from test dataset

### Performance using ANN

The accuracy achieved by BLAST and PSI-BLAST based modules was very poor for Alt-DS1; therefore, and in order to improve the accuracy, we employed a machine-learning technique such as ANN for training and testing. First, we developed an ANN-based method using amino acid composition as an input feature. For Alt-DS1, Alt-DS2, and Alt-DS3, amino acid composition based-ANN modules were able to achieve accuracies of 70.2%, 72.6%, and 75.8%, respectively (Tab. 1). Further, discrimination between secretory and non-secretory proteins was also assessed at residue level. It was found that compositions of cysteine and leucine amino acids were important for the classification between secretory and non-secretory proteins. For secretory proteins, the average content of cysteine and leucine residues was elevated in comparison to non-secretory proteins as shown in Fig. 1. In addition, physico-chemical properties based-ANN achieved accuracies of 68.5%, 70.6%, and 74.1%, respectively. Further, accuracies increased to 76.2%, 79.8%, and 83.9%, when dipeptide composition was used as an input feature to train the neural network (Tab. 1).

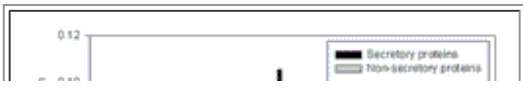
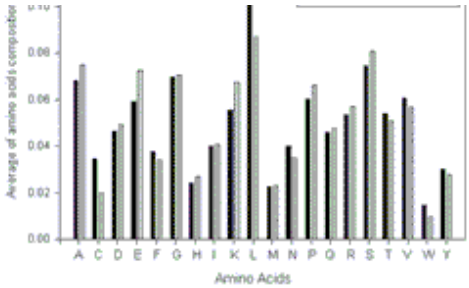


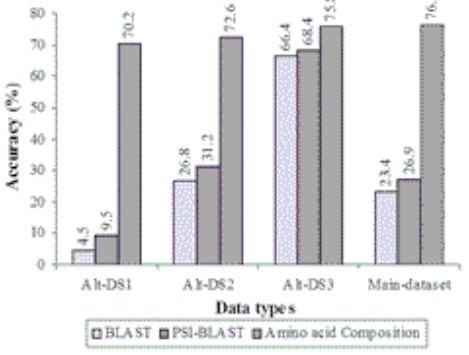

**Figure 1:** Average composition for 20 types of amino acids present in secretory and non-secretory proteins. Here, amino



Click on the thumbnail to enlarge the picture

acids are shown by their single letter code.

The comparison of amino acid composition based networks and similarity search based modules for three Alt-DS types is shown in Fig. 2. It can be clearly seen that composition-based different ANN modules had achieved higher accuracy in comparison to BLAST and PSI-BLAST based modules. Hence, we can interpret that the cases where similarity search based modules are unsuccessful, features such as composition of amino acids, properties and dipeptide can be successfully used. Interestingly, a change in accuracy can also be noticed moving to different data set types. For instance, the accuracies achieved by Alt-DS1 were lower in comparison to Alt-DS2 and Alt-DS3 (Fig. 2). The reason behind this low accuracy might be due to the small number of sequences in Alt-DS1 in comparison to Alt-DS2 and Alt-DS3. The memorization may be one of the reasons for better performance by neural network for composition based method on larger dataset. But this is also complicated by the fact that if a larger dataset is redundant, than chances of training and testing dataset with similar proteins also increases, and hence resulted in overestimation of accuracy.



Click on the thumbnail to enlarge the picture

**Figure 2:** Comparison of amino acid composition based-ANN and similarity search based modules (BLAST and PSI-BLAST) developed using four types of dataset.

Since Alt-DS (Alt-DS1) consisted of a low number of sequences, which was further effecting the prediction accuracy of amino acid composition-based ANN, we hence used Main-dataset (which was also developed using the same $E$-value cut-off but containing 6975 sequences) to develop ANN modules that were based on physico-chemical properties, amino acid and dipeptide composition. Using this, good accuracies of 73.1%, 76.1% and 77.1% were achieved, respectively (Tab. 2).

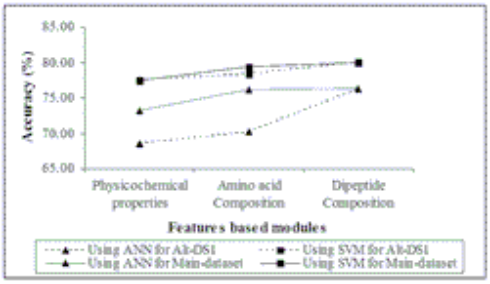
While comparing Main-dataset containing 6975 sequences with Alt-DS1 with 924 sequences (however, generated using the same cut-off $E$-value) for amino acid composition based-ANN modules, depicted an increase of ~6% in accuracy. Hence, we can say that ANN is sensitive to the size of dataset, i. e. accuracy increases with increase in number of sequences as shown in Fig. 2. In addition, for BLAST and PSI-BLAST modules, accuracies were found to be increased by 18% and 16%, respectively, using Main-dataset in comparison to Alt-DS1. Hence, this technique of making a dataset non-redundant seems to be more beneficial, while retaining the complete dataset of 6975 sequences in comparison to making Alt-DS types.

**Performance using SVM**

The Main-dataset was larger as well as non-redundant in comparison to the alternative datasets; therefore further studies were carried out using Main-dataset. In the past, it has been observed from several studies that SVM is more successful in handling problems related to protein classification and subcellular localizations prediction [14-16, 18-22]. Hence, in order to improve the performance, SVM modules based on different features of proteins such as composition of amino acids, physico-chemical properties and dipeptide composition of proteins were also constructed.

First, we constructed a SVM module based on amino acid composition for the prediction of secretory proteins and achieved an overall accuracy of 79.4%, which is significantly better than achieved using ANN (76.1%) for the same kind of input vector (Tab. 2). The amino acid composition provides information only about frequency but no information about the local order of residues. Therefore, in order to implement information about frequency as well as local order of residues in proteins, a SVM module based on dipeptide composition was also constructed. The dipeptide composition based SVM module (kernel = RBF, $\gamma$ = 100, C = 4) achieved an accuracy of 79.9%, comparable to amino acid composition based SVM modules. In addition, physico-chemical properties based SVM module was able to attain an accuracy of 77.4%.

Hence, the performance of SVM-based module was found to be much better than ANN-based modules. As shown in Fig. 3, the accuracies achieved for all the individual modules using SVM was greater than that achieved by ANN-based modules for both Alt-DS and Main-dataset types. Hence, SVM is more powerful for the prediction of secretory proteins as compared to ANN as shown in Fig. 3. Besides, it was also observed that the performance of ANN was better when Main-dataset was used in comparison to Alt-DS types, whereas in the case of SVM the performance is comparable for both types of datasets. Hence, the prediction accuracy of ANN is dependent on the size of the dataset, the more examples are present in the training dataset, the better would be the performance.




Click on the thumbnail to enlarge the picture

**Figure 3:** Comparing performance of ANN and SVM techniques using Main-dataset and Alt-DS1 type.
It can be seen that SVM outperforms ANN for all types of features based individual modules developed. Though SVM based modules achieved nearly same accuracies for both dataset types, ANN has produced some interesting results when using different dataset types. It seems that ANN technique is sensitive for the dataset size in comparison to the SVM, as significant improvement can be seen for ANN based modules using Main-dataset type which consisted of 6975 proteins.

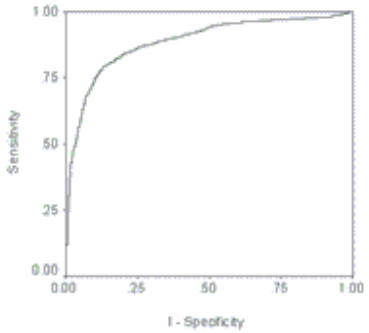
It is well known that hybrid approach-based SVM modules perform better than individual feature-based modules [14-16]. Therefore, in order to enhance the performance, we also constructed a hybrid using different features of a protein and output of PSI-BLAST. In the hybrid module, SVM was provided with an input vector of 423 dimensions, 20 of which for amino acid composition, 400 for dipeptide composition and 3 for PSI-BLAST output. The hybrid module (kernel = RBF, $\gamma$ = 3, C = 1) achieved an accurac of 83.2% (Tab. 2). It was proved that the hybrid module was able to encapsulate more comprehensive information, which successfully improved the prediction accuracy.

Further, the performance of different modules was also computed after removing redundancy from each test sets of Main-dataset. Doing so, maximum accuracies of 76.8%, 78.7% and 81.8 % were attained for amino acid composition, dipeptide composition and hybrid approach based SVM modules respectively. As shown in Tab. 2, the performance reduced slightly when tested on non-redundant test datasets.

### Comparison with existing methods

Since, identification of secretory proteins is important in computational biology and bioinformatics, a number of computational methods has been developed to predict the secretory or extra-cellular proteins in the past.

Although the present method has also been developed to predict secretory proteins, but concentrates on prediction irrespective of the presence or absence of N-terminal SP, it would be a little unfair to compare the overall performance of the present method with existing methods such as SignalP, pTarget and SecretomeP. Methods such as SignalP and pTarget recognize the signal peptide and hence, prediction is on the basis of identification of the N-terminal sequence. On the other hand, a NN-based method such as SecretomeP is specific for the prediction of secretory proteins that enter the non-classical pathway but is trained on secretory proteins that enter the classical and the non-classical pathway after cleaving the N-terminal sequence of classical secreted proteins. This method has been able to obtain a sensitivity of 40% with a low level of false positive prediction i. e. 5%. Following the same strategy of making 5 subsets non-redundant (but without cleaving N-terminal sequence of classical proteins), and using 33 physico-chemical properties as an input to ANN, we were able to achieve an overall accuracy of 73.1%. Further, sensitivity of 44.1% was achieved with 5% false positive prediction. In addition, the hybrid approach based SVM module was able to achieve high sensitivity of 60.4% with 5% false positive prediction (Fig. 4), which is significantly better than SecretomeP. The possible reason behind this high sensitivity with low level of false positive predictions is the consideration of protein features such as amino acids and dipeptide composition with similarity-search based results together in a hybrid module in comparison to only six properties used by SecretomeP. In addition, it was also observed that performance of amino acid and dipeptide composition based SVM modules was better than SecretomeP.



Click on the thumbnail to enlarge the picture

**Figure 4:** ROC plot for the SVM based hybrid module, where 1-specificity represents the false positive rate.

Recently, Klee and Sosa [23] carried out extensive analysis and comparison of different computational methods using an independent data set which was categorized into four main sets such as i)secretory and non-membrane; ii) secretory and membrane; iii) non-secretory and non-membrane and iv) non-secretory and membrane. In the present study the same dataset was also used to compare the performance of different predictors such as HSLpred, PSLpred, PredSL [24], SignalP, pTarget and SecretomeP. Here, the first three methods are multiple localization predictions, the next two are based on N-terminal SP identification and the last one is specifically to predict non-classically secreted proteins. It was observed that using an independent dataset the present method was able to achieve high sensitivity values such as 94% and 75% for predicting secretory proteins both for membrane and non membrane protein sets, respectively, which was found to be a better performance in comparison to HSLpred, PSLpred, PredSL, SignalP, pTarget and SecretomeP as shown in Tab. 3. However, SRTpred was not able to yield remarkable performance in classifying non-secretory proteins when compared with HSLpred, SignalP, pTarget and PredSL. Still, the overall performance of the present method was observed to be better than SecretomeP and comparable to other methods.

**Table 3:** Comparison of the prediction performance of existing computational methods using an independent test set.

| Parameter | SRTpred | HSLpred | PSLpred | PredSL | SecretomeP | SignalP | pTarget |
|---|---|---|---|---|---|---|---|
| **Without membrane proteins** | | | | | | | |
| True positives | 31 | 7 | 4 | 31 | 29 | 30 | 26 |
| False negatives | 2 | 26 | 29 | 2 | 4 | 3 | 7 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| True Negatives | 206 | 225 | 176 | 226 | 208 | 232 | 219 |
| False Positives | 31 | 12 | 61 | 11 | 29 | 5 | 18 |
| Sensitivity | 0.94 | 0.21 | 0.12 | 0.94 | 0.88 | 0.91 | 0.79 |
| Specificity | 0.87 | 0.95 | 0.74 | 0.95 | 0.88 | 0.98 | 0.92 |
| MCC | 0.63 | 0.21 | ---- | 0.81 | 0.60 | 0.87 | 0.63 |
| **With membrane proteins** | | | | | | | |
| True positives | 108 | 97 | 14 | 100 | 107 | 86 | 93 |
| False negatives | 36 | 47 | 130 | 44 | 37 | 58 | 51 |
| True Negatives | 221 | 239 | 188 | 240 | 218 | 246 | 235 |
| False Positives | 34 | 16 | 67 | 15 | 37 | 9 | 20 |
| Sensitivity | 0.75 | 0.67 | 0.10 | 0.69 | 0.74 | 0.60 | 0.65 |
| Specificity | 0.87 | 0.94 | 0.74 | 0.94 | 0.85 | 0.96 | 0.92 |
| MCC | 0.62 | 0.65 | ---- | 0.67 | 0.60 | 0.63 | 0.60 |

### SRTpred web server

Since the performance of SVM-based modules for the prediction of secretory proteins was found to be better in comparison to ANN-based modules, all the SVM modules constructed in the present study have been implemented on World Wide Web as dynamic web server 'SRTpred', freely available at http://www.imtech.res.in/raghava/srtpred/ using a CGI/Perl script. The server runs on a SUN server 420R under Solaris environment. It is a user-friendly web server and allows users to enter protein sequence in one of the standard formats such as FASTA, GenBank, EMBL, GCG, or plain format. Users can input their sequence by typing or pasting in the box or by using the file upload facility. The server also provides the options to select different approaches with threshold values ranging from −1.5 to 1.5. In the case of default prediction, the server uses the hybrid-based approach at the threshold value of 0.00. The displayed prediction results consist of classification of the respective query sequence as secretory or non-secretory protein along with the SVM predicted score.

## Acknowledgements

The authors are thankful to the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India for financial assistance. AG is thankful to CSIR for providing Senior Research Fellowship.

## References

1. Hughes, R. C. (1999). Secretion of the galectin family of mammalian carbohydrate-binding proteins. Biochim. Biophys. Acta **1473**, 172-185.
2. Cooper, D. N. W. (2002). Galectinomics, finding themes in complexity. Biochim. Biophys. Acta **1572**, 209-231.
3. Nickel, W. (2003). The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. Eur. J. Biochem. **270**, 2109-2119.
4. Lory, S. (1998). Secretion of proteins and assembly of bacterial surface organelles, shared pathways of extracellular protein targeting. Curr. Opin. Microbiol. **1**, 27-35.
5. Bendtsen, J. D., Nielsen, H., Krogh, A., von Heijne, G. and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. **340**, 783-795.

6. Guda, C. (2006). pTARGET: a web server for predicting protein subcellular localization. Nucleic Acids Res. **34**, W210-213.

7. Bendtsen, J. D., Jensen, L. J., Blom, N., von Heijne, G. and Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng. Des. Sel. **17**, 349-56.

8. Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28**, 45-48.

9. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**, 403-410.

10. Raghava, G. P. S. and Barton, G. J. (2006). Quantification of the variation in percentage identity for protein sequence alignments. BMC Bioinformatics **19**, 415.

11. Zell, A. and Mamier, G. (1997). Stuttgart Neural Network Simulator version 4.2, University of Stuttgart.

12. Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagating errors. Nature **323**, 533-536.

13. Joachims, T. (1999). Making large-scale SVM learning practical. *In:* Advances in Kernel Methods - Support Vector Learning. Edited by Scholkopf, B., Burges, C. and Smola, A.

14. Bhasin, M. and Raghava, G. P. S. (2004). ESLpred, SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res. **32**, 414-419.

15. Bhasin, M., Garg, A. and Raghava, G. P. S. (2005). PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics **21**, 2522-2524.

16. Garg, A., Bhasin, M. and Raghava, G. P. S. (2005). Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J. Biol. Chem. **280**, 14427-1432.

17. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta. **405**, 442-451.

18. Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. Jr. and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA **97**, 262-267.

19. Dönnes, P. and Elofsson, A. (2002). Predictions of MHC class I binding peptides, using SVMHC. BMC Bioinformatics **3**, 25.

20. Yu, C.-S., Lin, C.-J. and Hwang, J.-K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. **13**, 1402-1406.

21. Hua, S. and Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. Bioinformatics **17**, 721-728.

22. Park, K.-J. and Kanehisa, M. (2003). Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics **19**, 1656-1663.

23. Klee, E. W. and Sosa, C. P. (2007). Computational classification of classically secreted proteins. Drug Discov. Today **12**, 234-240.

24. Petsalaki, E. I., Bagos, P. G., Litou, Z. I. and Hamodrakas, S. J. (2006). PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. Genomics Proteomics Bioinformatics **4**, 48-55.