DOI 10.1002/pmic.200700597

Research Article

# RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information

*Rakesh Kaundal*[1] *and Gajendra P. S. Raghava*[2]

[1] Bioinformatics Lab, Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, OK, USA
[2] Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India

The attainment of complete map-based sequence for rice (*Oryza sativa*) is clearly a major milestone for the research community. Identifying the localization of encoded proteins is the key to understanding their functional characteristics and facilitating their purification. Our proposed method, RSLpred, is an effort in this direction for genome-scale subcellular prediction of encoded rice proteins. First, the support vector machine (SVM)-based modules have been developed using traditional amino acid-, dipeptide- ($i+1$) and four parts-amino acid composition and achieved an overall accuracy of 81.43, 80.88 and 81.10%, respectively. Secondly, a similarity search-based module has been developed using position-specific iterated-basic local alignment search tool and achieved 68.35% accuracy. Another module developed using evolutionary information of a protein sequence extracted from position-specific scoring matrix achieved an accuracy of 87.10%. In this study, a large number of modules have been developed using various encoding schemes like higher-order dipeptide composition, N- and C-terminal, splitted amino acid composition and the hybrid information. In order to benchmark RSLpred, it was tested on an independent set of rice proteins where it outperformed widely used prediction methods such as TargetP, Wolf-PSORT, PA-SUB, Plant-Ploc and ESLpred. To assist the plant research community, an online web tool 'RSLpred' has been developed for subcellular prediction of query rice proteins, which is freely accessible at http://www.imtech.res.in/raghava/rslpred.

Received: June 25, 2007
Revised: October 28, 2008
Accepted: December 3, 2008

**Keywords:**
Amino acid composition / Position-specific scoring matrix / PSI-BLAST / Rice / Subcellular localization

## 1 Introduction

Rice (*Oryza sativa* L.) is the single most important agricultural resource that feeds more than half of the world's population. The recently sequenced rice genome [1] has opened new challenges ahead for the plant research community in terms of assigning a biological role to these sequences, the function of only a few thousand of which can be defined with great confidence based on sequence similarity with genes of known function. It is the first crop plant to be sequenced and will therefore have a great impact in agriculture. Gaining an understanding of the biological functions of novel genes is a more ambitious goal than obtaining just their sequences; however, the wealth of information on nucleotide sequences that is being generated through the International Rice Genome Sequencing Project (IRGSP) far outweighs what is currently available on the amino acid sequences of known proteins. To narrow this huge gap be-

**Correspondence**: Dr. Gajendra P. S. Raghava, Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A, Chandigarh, India
**E-mail**: raghava@imtech.res.in
**Fax**: +91-172-269-0632 or 269-0585

**Abbreviations**: **MCC**, Matthews correlation coefficient; **PSI-BLAST**, position-specific iterated-basic local alignment search tool; **PSSM**, position-specific scoring matrix; **RBF**, radial basis function; **SAAC**, splitted amino acid composition; **SVM**, support vector machine tween the enormous amount of raw sequence data of the rice genome and the experimental characterization of the corresponding proteins, scientists therefore have to find computational ways to efficiently analyze these data. Genome functional annotation including the assignment of a function for a potential gene(s) in the raw sequence(s) is now the hot topic in rice bioinformatics. Subcellular location is one of the key functional characteristics of potential gene products such as proteins as they must be localized correctly at the subcellular level to have normal biological function. Moreover, it also provides information on the involvement of a protein in specific metabolic pathways [2–3]. Compared with the experimental methods, computational prediction methods provide fast, automatic and accurate assignment of subcellular location to a protein, especially for high-throughput analysis of large-scale genome sequences. Therefore, a fully automatic and reliable prediction system for subcellular localization of rice proteins would be very useful.

In the past, various methods have been developed to predict subcellular localization based on different features of a protein sequence. The similarity search-based tools have been used traditionally for functional annotation of proteins where a sequence is searched against an experimentally annotated database and a function is assigned to the protein [4]. However, this approach fails when an unknown query protein does not have significant homology to proteins of known functions [5]. Another way to predict subcellular localization of proteins is to identify sequence motifs such as signal peptide or nuclear localization signal [6]. This approach has been limited by the observation that not all of the proteins residing in a compartment have universal motifs [7]. To overcome these limitations, many machine learning technique-based methods such as artificial neural networks and support vector machines (SVM) have been developed to predict the subcellular localization of proteins. Currently, some widely used tools for subcellular localization are NNPSL [5], SubLoc [8], LOCtree [9] and PA-SUB [10] for both prokaryotes and eukaryotes, PSORT I [2] and PSORTB [11] for prokaryotic organisms, ESLpred [7], iPSORT [12], Wolf PSORT (updated version of PSORT II) [13], Euk-Ploc [14], LOCSVMPSI [15] and TargetP [16] for eukaryotes, HSLpred [17] and Hum-Ploc [18] specifically for human proteins and Plant-Ploc [19] specifically for plant proteins, all having good accuracy (>70%). Out of these, TargetP, PA-SUB, LOCtree, Wolf PSORT and LOCSVMPSI have also one module trained on plant networks for prediction of subcellular location of plant proteins. Most of these methods can be classified into two classes: one is based on the N-terminal sorting signals [3] and the other is based on amino acid composition. N-terminal sorting signals have a clear biological implication [20]. However, in large genome analysis projects, genes are usually automatically assigned, and these assignments are often unreliable for the 5'-regions [5]. This can result in leader sequences being missed or only partially included, thereby causing problems for prediction algorithms depending on them. Therefore, most of the methods are based on the amino acid composition rather than the N-terminal sorting signals alone. In the present study, we have also used various other approaches of amino acid composition like the higher-order dipeptide composition approach as used in some of the earlier studies [17, 21] combined with the N-terminal and C-terminal compositions as well as the similarity search-based position-specific iterated-basic local alignment search tool (PSI-BLAST) [22] and the position-specific scoring matrix (PSSM) of a protein sequence generated from the profiles of PSI-BLAST.

Recent advances in the prediction of protein-targeting signals have stressed the need for organism-specific prediction tools [23]. Moreover, by assessing on an independent dataset of proteins, our group has also demonstrated that organism-specific prediction methods are more accurate as compared to the performance of methods developed for general eukaryotic organisms [17]. To the best of our knowledge, there is no method available for the prediction of subcellular localization of rice proteins. Secondly, the availability of vast rice genome data now demands a reliable and accurate method for subcellular localization prediction of its encoded proteins. RSLpred is a systematic attempt in this direction, which is a SVM-based prediction method for four major target locations *viz*. chloroplast, cytoplasm, mitochondria and nucleus. The SVM modules were developed using various features of a protein sequence and the performance of these models was evaluated using fivefold cross-validation technique.

In addition, by using an independent dataset of query rice proteins, we have also compared the performance of our organism-specific method (RSLpred) with other widely used methods (TargetP, PA-SUB, Wolf PSORT and Plant-Ploc) for prediction of subcellular localization of plant proteins as well as with the general methods for predicting eukaryotic proteins like ESLpred. It was observed that RSLpred could predict the subcellular localization of rice proteins with far better accuracy as compared to the available methods. Finally, a web-based server was developed based on all the 13 approaches followed in this study to provide service to the research community, where the users have the option to select any of these feature-based modules for the prediction of subcellular localization of query rice proteins.

## 2 Materials and methods

### 2.1 The dataset(s)

#### 2.1.1 Main data for training/testing

Due to scarcity of annotated rice protein sequences, whole of the UniProt Knowledgebase (release 5.0) was searched for the available sequences. We identified rice proteins with specific subcellular locations according to the annotation information in the CC (comments or notes) fields of UniProt Knowledgebase, of which subcellular location information was available for only 825 proteins. These 825 proteins of known subcellular localization information were divided into 11 groups according to their subcellular localization *viz.* chloroplast, cytoplasm, cytoskeleton, ER, extra-cellular/secreted, Golgi apparatus, mitochondria, nucleus, peroxisome, plasma membrane and vacuole. Proteins annotated with two or more subcellular locations were not included in the current dataset. For example, a protein entry annotated with 'SUBCELLULAR LOCATION: NUCLEAR AND CYTOPLASMIC' in the CC field was not included. All protein entries computationally selected were then manually examined. However, amount of data available for some groups was too small for statistical analysis to be performed (Table 1).

Because the number of rice mitochondrial sequences extracted from UniProt Knowledgebase was too small to allow reliable network training, 242 mitochondrial sequences from other related plant species (*Arabidopsis thaliana*) were included in the final dataset (Table 1). Previous studies have not been able to reveal significant species-correlated differences between mitochondrial proteins, as a cluster analysis of mTP [24] using self-organized maps [25] did not reveal any significant species-specific features of mitochondrial proteins. Even the use of non-plant mitochondrial proteins in the training of plant mitochondrial proteins and conversely, the use of plant mitochondrial proteins in the training of non-plant mitochondrial proteins is reasonable [16]. Therefore, our inclusion of mitochondrial proteins from one of the other related plant species in our final dataset of rice mitochondrial proteins for training purpose seems justified. Further, the sequence redundancy was reduced by using PROSET software [26] such that no two sequences had >90% sequence identity in the final dataset. Although the 90% cutoff is not sufficient to avoid homology bias, considering the current limited number of location-known rice proteins in the Uniprot Knowledgebase, it was taken as a compromise. In this context, we would like to mention that to avoid homology bias, a 25% sequence identity cutoff threshold is needed to guarantee that none of the proteins included in the benchmark datasets has greater than 25% sequence identity to any other in a same subcellular location, as done in constructing the benchmark datasets for Euk-Ploc [14] and Hum-Ploc [18]. The number of sequences for about eight subcellular locations was not sufficient for developing a prediction method (Table 1). Therefore, a prediction method was developed for only four major subcellular locations on a final dataset of total 910 proteins (84 chloroplast, 103 cytoplasmic, 247 mitochondrial and 476 nuclear). The complete list of Uniprot Knowledgebase ID of these protein sequences has been provided in the Supporting Information and is available for free download from RSLpred web server.

**Table 1.** Number of sequences within each subcellular location group

| Subcellular location | Number of sequences available | Final dataset |
|---|---|---|
| Chloroplast | 140 | 84 |
| Cytoplasm | 150 | 103 |
| Mitochondria | 262[a)] | 247 |
| Nucleus | 477 | 476 |
| Extra-cellular/secreted | 10 | - |
| Golgi apparatus | 0 | - |
| Cytoskeleton | 0 | - |
| Endoplasmic reticulum | 18 | - |
| Peroxisome | 6 | - |
| Plasma membrane | 2 | - |
| Vacuole | 2 | - |
| Total | 1067 | 910 |

a) Due to scarcity of rice mitochondrial proteins, 242 *Arabidopsis thaliana* mitochondrial proteins were included.

### 2.1.2 Independent datasets for validation of RSLpred

Techniques such as cross-validation and bootstrapping are routinely used for evaluating the performance of any method, but the best way of testing the performance of a newly developed method is to test it on an independent dataset that contains the patterns used neither during training nor during testing of the method. To validate the performance of our method (RSLpred) for predicting subcellular localization of rice proteins, an independent dataset of rice proteins, which was not used in training/testing of RSLpred, was compared for its prediction performance with the best available methods especially trained on plant networks like TargetP, PA-SUB, Wolf PSORT and the recently developed all-plant method, Plant-Ploc, and for general eukaryotic organisms like ESLpred. For this, an independent rice data was again derived from the latest release 7.0 of the UniProt Knowledgebase and divided into the four subcellular classes under study. Further, the sequence redundancy was reduced by using PROSET software [26] such that no two sequences had >90% sequence identity in the independent dataset, as was also done while generating final training/testing dataset. This was done to ensure that no sequences in the independent dataset had >90% redundancy with any of the sequences in the training/testing dataset. We named it 'independent dataset-I' that contained 508 additional rice proteins (58 chloroplast, 165 cytoplasmic, 58 mitochondrial and 227 nuclear), which were not used in the training and testing of the RSLpred method.

Furthermore, to make this test dataset completely independent from the original training data of 910 proteins, we further reduced its redundancy to 25% cutoff level. This left us with 345 sequences in the testing set, which we named as 'independent dataset-II', consisting of 38 chloroplast, 109 cytoplasmic, 50 mitochondrial and 148 nuclear proteins. Both these datasets were made to run on our best classifier and on some other widely used general prediction tools available worldwide.

## 2.2 Support vector machine

SVM, an excellent machine learning technique introduced by Vapnik and co-workers [27, 28], is a universal approximator based on the statistical and optimization theory and has been applied in many classification and regression problems. It has been successfully used for classification of microarray data [29], disease forecasting [30] and protein secondary structure prediction [31] as well as for the subcellular localization of proteins in eukaryotic and prokaryotic organisms [7–9, 15, 17]. Further details about the SVM can be obtained from Vapnik's monographs [27, 28, 32]. In the present study, we have used SVM_light [33], a freely downloadable package of SVM (http://svmlight.joachims.org/old/svm_light_v4.00.html), to predict the subcellular localization of proteins. This software enables the users to define a number of parameters besides allowing a choice of inbuilt kernel function including linear, polynomial and radial basis function (RBF). As the prediction of subcellular localization of proteins is a multi-class classification problem, therefore we constructed *N* SVM for *N*-class classification. In the present study, the class number was equal to four for rice protein sequences. The *i*th SVM was trained with all the samples in the *i*th class as positive label and negative label for proteins of remaining subcellular locations. In this way, four SVM were constructed for subcellular localization of protein to chloroplast, cytoplasm, mitochondria and nucleus. This type of SVM is known as one *versus rest* (1-v-r SVM) [8]. An unknown sample was classified into a particular class that corresponded to the SVM with highest output score. To achieve maximum accuracy, we have attempted fifteen different approaches based on various features of a protein sequence, which are hereby discussed in brief.

## 2.3 Features and modules

### 2.3.1 Composition based

#### 2.3.1.1 Amino-acid composition

Amino-acid composition is the fraction of each amino acid in a protein sequence. The fraction of all the natural 20 amino acids was calculated using the following equation:

$$\text{Fraction of amino acid } i = \\ = \frac{\text{Total number of amino acid } i}{\text{Total number of amino acids in protein}} \tag{1}$$

where *i* can be any amino acid.

#### 2.3.1.2 Traditional dipeptide composition

To encapsulate the global information about each protein sequence, dipeptide composition was used. This representation, which gives a fixed pattern length of 400 (20 × 20), encompasses the information of the amino-acid composition along with the local order of amino acids. The fraction of each dipeptide was calculated according to the equation:

$$\text{Fraction of dep } (i\ +1) = \\ = \frac{\text{Total number of dep } (i\ +1)}{\text{Total number of all possible dipeptides}} \tag{2}$$

where dep ($i$ + 1) is one of 400 dipeptides.

#### 2.3.1.3 Higher-order dipeptide composition

This approach not only reflects the total amino acid composition, but also incorporates, to a considerable degree, the sequence-order effects [17, 21]. This representation, which also gives a fixed pattern length of 400 (20 × 20), encompasses the information of amino-acid composition along with the pseudo order of amino acids. Here, various higher-order dipeptides (can also be called as *pseudo dipeps*) such as $i$ + 2, $i$ + 3, $i$ + 4 and $i$ + 5 were generated in order to observe the interaction of the $i^{\text{th}}$ residue with the $3^{\text{rd}}$, $4^{\text{th}}$, $5^{\text{th}}$ and $6^{\text{th}}$ residue in the sequence, respectively using Eq. 3:

$$\text{Fraction of } (i\ +\text{n}) \text{ pseudo dep} = \\ = \frac{\text{Total number of } (i\ +\text{n}) \text{ pseudo dep}}{\text{Total number of all possible dipeptides}} \tag{3}$$

#### 2.3.1.4 Cumulative higher order dipeptide composition

In addition, the frequencies of all the ($i$ + n) dipeps were combined and divided by the sum of all possible dipeptides of each ($i$ + n) to again form a combined fixed length pattern of 400 for use in SVM. We designated this as cumulative higher-order dipeptide composition, which was calculated according to the equation:

$$\text{Cumulative fraction of } (i\ +\text{n}) \text{ dep} = \\ = \frac{\text{f}_{(\text{i}+1)} + \text{f}_{(\text{i}+2)} + \text{f}_{(\text{i}+3)} + \text{f}_{(\text{i}+4)} + \text{f}_{(\text{i}+5)}}{(N-1) + (N-2) + (N-3) + (N-4) + (N-5)} \tag{4}$$

where *N* is the sequence length.

#### 2.3.1.5 Four parts composition

Here, each protein sequence was divided into four equal parts. This type of approach has comparatively shown some good results as evident from some earlier studies [15, 21]. The occurrence frequency of each amino acid was calculated separately using Eq. (1) for each part and then a combined fixed pattern length of 80 (20 × 4) was formed in order to gather more information about the protein sequence.

### 2.3.2 Similarity search-based

A module RiPSI-BLAST was designed in which a query sequence was searched against the existing non-redundant database of classified proteins (910 entries of training set) using PSI-BLAST. In the present study, PSI-BLAST was used instead of normal standard BLAST because it has the capability to detect remote homologies [22]. It carries out an iterative search in which sequences found in one round were 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. This module could predict any of the four localizations (chloroplast, cytoplasmic, mitochondrial and nuclear) depending upon the similarity of the query protein to the proteins in the dataset. The module would return “unknown subcellular localization” if no significant similarity was found.

### 2.3.3 Position-specific scoring matrix

PSSM-based SVM was another module constructed by combining the evolutionary information stored in the matrix of a protein sequence called as PSSM that is a method for detecting distantly related proteins by sequence comparison. The idea of adopting PSSM extracted from sequence profiles generated by PSI-BLAST as input information was first proposed by David Jones [34]. This information is expressed in a position-specific scoring table (profile), which is created from a group of sequences previously aligned by PSI-BLAST against the non-redundant database at the GenBank. The PSSM gives the log-odds score for finding a particular matching amino acid in a target sequence (Fig. 1). It differs from other methods of sequence comparison in common use because any number of known sequences can be used to construct the profile, allowing more information to be used in the testing of the target sequence. The PSSM of a protein sequence extracted from the profiles of PSI-BLAST was used to generate a 400-dimensional input vector to the SVM by summing up all rows in the PSSM corresponding to the same amino acid in the primary sequence (Fig. 1). After that, every element in this input vector was divided by the length of the sequence and then scaled to the range of 0–1 by using the standard linear function:

$$\frac{(X - \text{minimum})}{(\text{maximum} - \text{minimum})}$$

where *X* is the individual PSSM score of each amino acid.

### 2.3.4 Hybrid SVM module(s)

To further enhance the prediction accuracy, we also adopted various hybrid approaches by combining different features of a protein sequence.

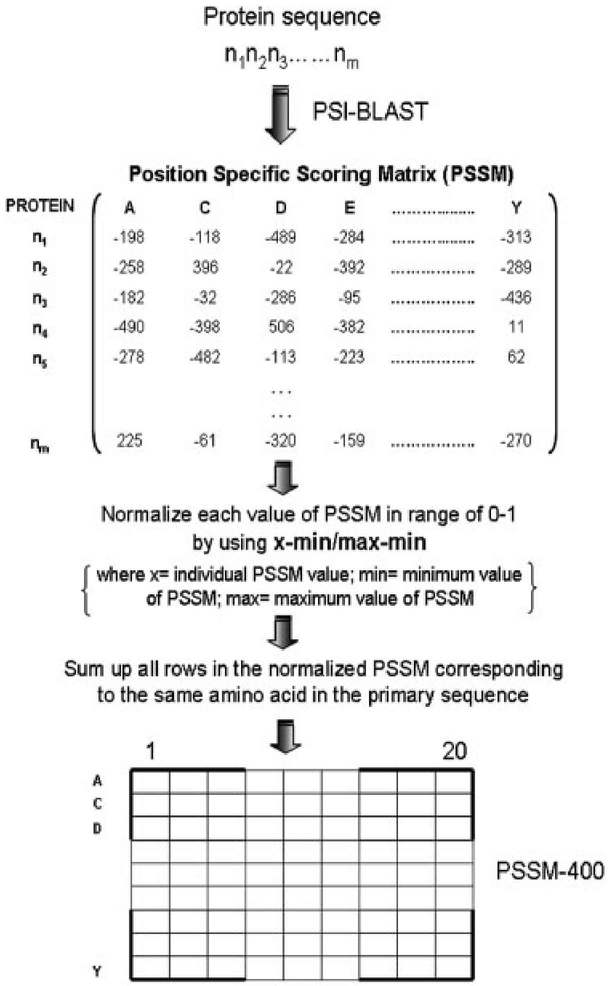

Protein sequence
$n_1n_2n_3\ldots\ldots n_m$
PSI-BLAST
Position Specific Scoring Matrix (PSSM)
PROTEIN A C D E ................ Y
$n_1$ -198 -118 -489 -284 ................ -313
$n_2$ -258 396 -22 -392 ................ -289
$n_3$ -182 -32 -286 -95 ................ -436
$n_4$ -490 -398 506 -382 ................ 11
$n_5$ -278 -482 -113 -223 ................ 62
...
...
$n_m$ 225 -61 -320 -159 ................ -270
Normalize each value of PSSM in range of 0-1
by using x-min/max-min
where x= individual PSSM value; min= minimum value of PSSM; max= maximum value of PSSM
Sum up all rows in the normalized PSSM corresponding to the same amino acid in the primary sequence
1
20
A
C
D
Y
PSSM-400

**Figure 1.** Schematic representation of algorithm used to convert 21*N dimensional PSSM matrix into PSSM-400 input pattern for SVM.

#### 2.3.4.1 Hybrid approach-I

In the first step, we developed a hybrid module by combining amino acid composition and dipeptide composition features of a protein sequence as calculated by using Eqs. (1) and (2), respectively. This module was provided with an SVM input vector pattern of 420 (20 for amino acid and 400 for dipeptide composition).

#### 2.3.4.2 Hybrid approach-II

Secondly, we developed another hybrid module by combining amino acid composition as calculated using Eq. (1) combined with the evolutionary information stored in the matrix of a protein sequence called PSSM. The SVM input vector pattern thus formed was 420-dimensional (20 for amino acid and 400 for PSSM).

#### 2.3.4.3 Hybrid approach-III

Finally, we attempted another hybrid module by combining amino acid composition as calculated according to Eq. (1), traditional dipeptide composition as calculated using Eq. (2) and the PSSM matrix as generated from profiles of PSI-BLAST. Here, the SVM input vector pattern increased to 820 (20 for amino acid, 400 for dipeptide composition and 400 for PSSM).

### 2.3.5 Terminal-based SVM modules

#### 2.3.5.1 N-terminal composition

Many proteins in the plant cell (*e.g.* chloroplast, mitochondria, ER, secretory and some peroxisome targeted) have sorting signals that relies on the presence of an N-terminal targeting sequence, which is recognized by a translocation machinery. These signals are responsible for targeting proteins to various subcellular localizations in the cell (chloroplast and mitochondria in the present study). Therefore, we also attempted an SVM module-based on the N-terminal amino-acid composition of each protein giving a 20-dimension input vector pattern. The SVM module was developed at various levels of N-terminal residue length (10, 15, 20, 25 and 30 amino acids) in order to achieve maximum accuracy.

#### 2.3.5.2 C-terminal composition

We also developed an SVM module based on the C-terminal amino acid composition of each protein, which gives a 20-dimension input vector pattern to the SVM. Here, we also altered the C-terminal residue length (10, 15, 20, 25 and 30 amino acids) in order to achieve maximum accuracy.

#### 2.3.5.3 N-Centre-C-terminal (three parts) composition

This type of approach is also called splitted amino-acid composition (SAAC) where the amino acid composition of N-terminal (25 residues), the C-terminal (25 residues) and the remaining centre portion of a protein sequence was calculated separately by using Eq. (1) and was combined to form a 60-dimension (20 × 3) input vector to SVM. The rationale behind using this type of approach is the fact that percentage composition of whole sequence does not give adequate weight to the compositional bias, which is known to be present in protein terminus.

#### 2.3.5.4 N-terminal + remaining part composition

We also attempted an SVM module based on the division of protein sequence into two parts *viz*. N terminus (25 residues) and the remaining part of the sequence. The amino-acid composition was calculated separately for both the parts so that it gave 40 (2 × 20) SVM vector pattern.

#### 2.3.5.5 C-terminal + remaining part composition

Similarly, another SVM module based on the division of protein into two parts *viz*. C terminus (25 residues) and the remaining part of the sequence was developed. The amino-acid composition was calculated separately for both the parts to form a 40 (2 × 20) SVM vector pattern.

## 2.4 Measurements for performance of RSLpred

In statistical prediction, the single independent dataset test, sub-sampling test and jackknife test are the three methods often used for cross-validation. Of these three, the jackknife test is deemed as the most rigorous and objective one, as illustrated by a comprehensive review [35]. Therefore, jackknife test has been increasingly used in literature [36–41] for examining the accuracy of various prediction methods. However, jackknife test method takes much longer time to train a predictor based on SVM, and hence, here the sub-sampling (fivefold) cross-validation was adopted for performance measurement, as done by most SVM-based methods. In this technique, the relevant dataset was partitioned randomly into five equally sized sets. The training and testing was carried out five times, each time using one distinct set for testing and the remaining four sets for training. For evaluating the performance of various modules developed, the accuracy (ACC) and Matthews correlation coefficient (MCC) were calculated as described by Hua and Sun [8], using Eqs. (5) and (6). The overall accuracy and MCC of RSLpred was calculated by using Eqs. (7) and (8), respectively:

$$\mathrm{Accuracy}(x) = \frac{p(x)}{Exp(x)} \tag{5}$$

$$\mathrm{MCC}(x) = \\ = \frac{p(x)n(x) - u(x)o(x)}{\sqrt{(p(x)+u(x))(p(x)+o(x))(n(x)+u(x))(n(x)+o(x))}} \tag{6}$$

$$\text{Overall accuracy} = \frac{\sum_x p(x)}{N} \tag{7}$$

$$\text{Overall MCC} = \frac{\sum_x MCC(x) * Exp(x)}{N} \tag{8}$$

where $x$ can be any subcellular location (chloroplast, cytoplasmic, mitochondrial or nuclear), Exp($x$) is the number of sequences observed in location $x$, p($x$) is the number of correctly predicted sequences of location $x$, n($x$) is the number of correctly predicted sequences not of location $x$, u($x$) is the number of underpredicted sequences, o($x$) is the number of overpredicted sequences and $N$ is the total number of sequences in the dataset.

### 2.5 Reliability index to the prediction

The reliability of prediction is an important factor that can provide users more information as well as confidence about the quality of prediction. We adopted the simple strategy of Hua and Sun [8] for assigning the reliability index (RI) to indicate the level of certainty in the prediction of a submitted sequence. The RI was assigned according to the difference (Δ) between the highest and second highest SVM output scores. This was calculated for the PSSM-based SVM module using Eq. (9):

$$\mathrm{RI} = \begin{cases} \mathrm{INT}(\Delta * 5/3 + 1 & \text{if } 0 \leq \Delta < 4, \\ 5 & \text{if } \Delta \geq 4. \end{cases} \tag{9}$$

### 2.6 Confusion matrix and its evaluation

In the field of artificial intelligence, a confusion matrix is a visualization tool typically used in supervised learning (in unsupervised learning, it is typically called a matching matrix). Confusion matrix is a table with the 'true' class in rows and the 'predicted' class in columns. The diagonal elements represent correctly classified instances while the cross-diagonal elements represent misclassified instances. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (*i.e.* commonly mislabeling one as another). For this, we generated an additional information from the performance of RSLpred on an independent dataset of rice proteins called as 'confusion matrix'.

Performance of such systems is commonly evaluated using the data in the matrix through the standard four statistics: specificity, precision, sensitivity and recall (the last two are identical) as followed in [10] also. Given a confusion matrix and a set of instances for class *k*, the standard definitions of these statistics are as follows.

The specificity (percentage of negative labeled instances that were predicted as negative) for each instance $\mathrm{I}_k$ is $\mathrm{S}_k$ defined by:

$$\mathrm{S}_k = \frac{TN}{TN + FP} \tag{10}$$

The precision (percentage of positive predictions those are correct) for each instance $\mathrm{I}_k$ is $\mathrm{P}_k$ defined by:

$$\mathrm{P}_k = \frac{TP}{TP + FP} \tag{11}$$

The sensitivity or recall (percentage of positive labeled instances that were predicted as positive) for each instance $\mathrm{I}_k$ is $\mathrm{R}_k$ defined by:

$$\mathrm{R}_k = \frac{TP}{TP + FN} \tag{12}$$

Here, the true positives (TP) is the number of instances correctly predicted as $\mathrm{I}_k$, which were actually labeled $\mathrm{I}_k$; the false positives (FP) is the number of instances incorrectly predicted as $\mathrm{I}_k$ that were actually not labeled as $\mathrm{I}_k$; the true negatives (TN) is the number of instances correctly predicted as not $\mathrm{I}_k$, that were actually not labeled $\mathrm{I}_k$; and the false negatives (FN) is the number of instances incorrectly predicted as not $\mathrm{I}_k$ that were actually labeled $\mathrm{I}_k$. An overall version of each statistic was computed as a weighted average.

### 2.7 Annotation of rice proteome

The completion of the rice genome draft has brought unprecedented opportunities for genomic studies of the world's most important food crop. Now, a standardized annotation is necessary so that the information from the genome sequence can be fully utilized in understanding the biology of rice and other cereal crops. Thus, to facilitate the application of subcellular localization information and to provide a foundation for functional and evolutionary studies of other important cereal crops, we performed whole rice proteome subcellular predictions. For this, complete rice proteome was downloaded from two different web sources; EBI (www.ebi.ac.uk/integr8) and TIGR (www.tigr.org). As the number of protein entries varied from both the sources (EBI = 30 952 proteins and TIGR = 62 827 proteins), subcellular predictions were done on both these proteomic datasets. As it was nearly impossible to generate PSSM matrices for such huge data, we used the faster and traditional amino acid composition-based classifier for performing the predictions for each of the four subcellular localizations.

## 3 Results and discussion

World agriculture gets a major boost with the completion of the rice genome sequence [1]. The rice genome sequencing has identified about 37 544 genes many of which are represented by two or more copies. The accurate, map-based sequence has already led to the identification of genes responsible for agronomically important traits such as genes that affect growth habit to promote yield and photoperiod genes to extend the range of elite cultivars. Moreover, about 98% of the genes known in cereals are found in the rice genome, confirming the potential of the model monocot for discovering gene function in other cereal crops [42]. This could probably provide the key in improving yield to feed an expanding world population at a time of increasing restraints on agriculture. Amazingly, genes are simple, consisting of four types of nucleotides (adenine, guanine, cytosine and thymine) and are translated into far more complex proteins that are made up of 20 different types of amino acids. Among many other things, these proteins control all type of crop development and physiology besides providing resistance to various crop pests and diseases. In order to perform its appropriate functions, each protein must be translocated to its correct intra- or extracellular compartments. Hence, the subcellular localization is a key step characteristic of each encoded functional protein.

Since 1991, various diverse algorithms have been developed to predict the subcellular localization of proteins, based on amino-acid compositions [43], *k* nearest neighbors [13, 44], neural networks [5], covariant discriminant algorithm [45], Markov chains [46], SVM [8, 47] and combination of several methods [11]. In general, machine-learning techniques such as artificial neural networks and SVM are considered as elegant approaches for the prediction of subcellular localization of proteins. Further, previous studies have shown that SVM performs better as compared to the artificial neural networks [8, 9, 30]. In the present study, we have developed various support vector machine-based modules, which were evaluated through a fivefold cross-validation technique. In the training of SVM, we used the method of one versus the others, or one versus the rest. For example, an SVM for the chloroplast protein group was trained with the chloroplast protein sequences used as positive samples and proteins in the other three subcellular location groups used as negative samples, because SVM basically train classifiers between only two different samples. Thus, we build 60 different SVM classifiers corresponding to four subcellular localizations under 15 different types of approaches followed as discussed above. For each of these 15 different approaches, a query protein was tested against the 4 SVM classifiers and assigned to the subcellular location that corresponds to the highest output value. The SVM training was carried out by the optimization of various kernel function parameters and the value of regularization parameter *C*. It was observed that the RBF kernel performs better than the linear and polynomial kernels in the case of amino acid composition-based SVM module. Thus, for all of the SVM modules developed further in the present study, only RBF kernel was used.

### 3.1 Composition-based SVM modules

The amino-acid composition-based SVM module was able to achieve an overall accuracy of 81.43% (kernel = RBF, $\gamma$ = 200, *C* = 3, j = 5) for all of the four subcellular localizations. An SVM module based on the traditional dipeptide composition ($i + 1$) to implement more information about frequency as well as the local order of residues was also constructed. This module could achieve a maximum overall accuracy of 80.88% with the RBF kernel ($\gamma$ = 225, *C* = 2, j = 2). In order to have more comprehensive information on the sequence-order effects, we also developed various higher-order dipeptide composition-based modules ($i$+2, $i$+3, $i$+4, $i$+5) including a cumulative higher-order amino-acid pairs module. The individual overall accuracies of all the higher-order dipeptide modules could not exceed the accuracy achieved by actual dipeptide ($i$+1) composition-based module (Table 2). This may be because ($i$+1) module uses the actual order of sequence while calculating the dipep composition, whereas the higher-order dipep modules are based on the pseudo sequence order effects. However, when the cumulative higher-order dipeptide composition-based module was developed, it achieved an overall accuracy of 82.97% ($\gamma$ = 300, *C* = 2, j = 7) which revealed an increase in overall accuracy of about 2% over the ($i$+1) actual dipeptide composition-based module. Further, each protein sequence was divided into four equal parts and amino acid composition was calculated individually for each part and this combined SVM module was able to achieve an overall accuracy of 81.10% (kernel = RBF, $\gamma$ = 15, *C* = 3, j = 1) for all of the four target locations. The detailed performance of amino acid-, traditional dipeptide-, higher-order dipeptide-, cumulative higher-order dipeptide- and four parts composition-based SVM modules in assigning different subcellular localizations has been presented in Table 2.

### 3.2 Sequence similarity search

To encapsulate evolutionary information about the proteins, the homology of a protein with other related sequences provides a broad range of information about each functional encoded protein. Hence, the similarity search-based module RiPSI-BLAST was also constructed. During 5-fold cross-validation, no significant hits were obtained for 262 out of 910 proteins in the data set. Thus, it proves that the performance of this similarity search-based module is poorer in comparison with amino acid-, dipeptide- as well as the four parts composition-based modules. RiPSI-BLAST was able to predict chloroplast, cytoplasmic, mitochondrial and nuclear subcellular localizations with 34.52, 65.05, 35.63 and 92.02% accuracy, respectively with an overall accuracy of 68.35% (Table 2). Therefore, it suggests that the similarity search-based tools alone cannot annotate the protein sequences as efficiently and reliably as the composition-based modules.

### 3.3 PSSM-based SVM module

Further, from the generated sequence profiles by PSI-BLAST, a PSSM was constructed for each protein sequence, which was used as 400-dimension input information to the SVM. The PSSM-based SVM module has achieved striking higher overall accuracy of 87.10% (kernel = RBF, $\gamma$ = 45, *C* = 2, j = 6) for all of the four subcellular localizations, which was significantly best of all the approaches attempted by us. Moreover, the *p*-value comparison of all the modules based on the statistical significance (0.05 level) also revealed that PSSM-based approach was statistically best over the rest modules (Table 3). This demonstrates that combining machine learning technology; prediction performance can be significantly improved with the usage of PSI-BLAST profiles that offers important evolutionary information about the protein subcellular localizations. The individual accuracies obtained with PSSM-based SVM for all four types of subcellular localization are presented in Table 2.

**Table 2.** Detailed performance of PSI-BLAST and various SVM modules developed using different features of a protein sequence

| Approaches used[a)] | Chloroplast | | Cytoplasm | | Mitochondria | | Nuclear | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ACC[1] (%) | MCC[2] | ACC (%) | MCC | ACC (%) | MCC | ACC (%) | MCC | ACC (%) | MCC |
| AA-based $^{A}$ | 42.86 | 0.49 | 58.25 | 0.59 | 78.54 | 0.67 | 94.75 | 0.83 | 81.43 | 0.73 |
| DIPEP (i+1)-based $^{B}$ | 30.95 | 0.48 | 48.54 | 0.55 | 86.24 | 0.68 | 93.91 | 0.81 | 80.88 | 0.72 |
| Higher-order Dipeps | | | | | | | | | | |
| (i + 2) | 29.76 | 0.44 | 55.34 | 0.61 | 83.40 | 0.65 | 93.70 | 0.82 | 80.66 | 0.72 |
| (i + 3) | 30.95 | 0.24 | 59.22 | 0.44 | 79.76 | 0.52 | 61.98 | 0.54 | 63.63 | 0.49 |
| (i + 4) | 22.62 | 0.31 | 63.11 | 0.59 | 74.10 | 0.61 | 95.17 | 0.82 | 79.12 | 0.69 |
| (i + 5) | 25.00 | 0.25 | 53.40 | 0.56 | 72.88 | 0.57 | 92.65 | 0.79 | 76.59 | 0.65 |
| Cumulative higher-order | 41.67 | 0.54 | 60.19 | 0.61 | 83.40 | 0.70 | 94.96 | 0.85 | 82.97 | 0.75 |
| Four parts-based | 41.67 | 0.55 | 54.37 | 0.61 | 75.30 | 0.66 | 96.85 | 0.79 | 81.10 | 0.71 |
| RiPSI-BLAST | 34.52 | – | 65.05 | – | 35.63 | – | 92.02 | – | 68.35 | – |
| **PSSM**-based $^{C}$ | **55.63** | **0.58** | **74.63** | **0.77** | **83.81** | **0.74** | **97.06** | **0.90** | **87.10** | **0.81** |
| Hybrid - I $^{A+B}$ | 52.38 | 0.48 | 57.28 | 0.62 | 78.54 | 0.69 | 95.38 | 0.86 | 82.53 | 0.75 |
| Hybrid - II $^{A+C}$ | 52.38 | 0.55 | 64.08 | 0.65 | 82.59 | 0.74 | 96.22 | 0.88 | 84.84 | 0.78 |
| Hybrid - III $^{A+B+C}$ | 50.00 | 0.52 | 62.14 | 0.64 | 82.59 | 0.72 | 96.43 | 0.89 | 84.51 | 0.78 |
| NT - 25 | 26.19 | 0.30 | 27.19 | 0.33 | 68.83 | 0.59 | 89.29 | 0.58 | 70.88 | 0.53 |
| CT - 25 | 20.24 | 0.27 | 22.33 | 0.29 | 50.61 | 0.34 | 88.03 | 0.51 | 64.18 | 0.42 |
| SAAC (N-Centre-C) | 40.48 | 0.44 | 49.52 | 0.55 | 82.59 | 0.69 | 91.81 | 0.77 | 79.78 | 0.70 |
| NT-25 + Remaining | 38.10 | 0.47 | 46.60 | 0.49 | 80.97 | 0.70 | 92.65 | 0.75 | 79.23 | 0.68 |
| CT-25 + Remaining | 22.62 | 0.35 | 49.52 | 0.53 | 76.52 | 0.58 | 91.81 | 0.75 | 76.48 | 0.64 |

a) AA-based, amino acid composition used as input; DIPEP-based, dipeptide composition used as input; Four parts-based, whole protein is divided into four equal parts, amino acid composition calculated separately and all parts combined to form 80-dimension input vector; PSI-BLAST, similarity-search against non-redundant database of rice proteins; PSSM-based, a 400-dimension position-specific scoring table generated from PSI-BLAST profiles and used as input vector; Hybrid-I, AA and DIPEP are combined to form 40-dimension input vector; Hybrid-II, AA and PSSM are combined to form 420-dimension input vector to SVM; Hybrid-III, AA, DIPEP and PSSM information is combined to generate 820-dimension input vector; NT-25, amino acid composition of N-terminal 25 residues used as input; CT-25, amino acid composition of C-terminal 25 residues used as input; SAAC, whole protein is divided into three parts, N-terminal 25 amino acids, C-terminal 25 amino acids and remaining sequence. Amino acid composition of all three fragments determined and together used as input (vector of 60-dimensions); NT-25+Remaining, whole protein is divided into two parts, N-terminal 25 amino acids and remaining sequence and combined vector of 40-dimensions is used as input; CT-25+Remaining, whole protein is divided into two parts, C-terminal 25 amino acids and remaining sequence and combined vector of 40-dimensions is used as input; [1] ACC, accuracy in %; [2] MCC, Matthews correlation coefficient.

### 3.4 Hybrid approach(es)

In addition, methodologies such as 'hybrids' were also devised to acquire more comprehensive information of the proteins by combining various features of a protein sequence. In the first step, we developed a hybrid module by combining amino acid composition and dipeptide composition. Best results were obtained with RBF kernel ($\gamma$ = 100, $C$ =6, j = 2) with an overall accuracy of 82.53%, which was about 1% higher than the amino-acid composition-based SVM method, though statistically nonsignificant as revealed by *p*-value (Table 3). Secondly, we developed another hybrid module by combining amino acid composition and PSSM-based matrix information. Best results were obtained with RBF kernel ($\gamma$ = 45, $C$ = 2, j = 4) with an overall accuracy of 84.84%, which was about 2% superior over the hybrid approach-I-based SVM method, but again not statistically significant. Further, we also attempted hybrid approach-III by combining amino-acid composition, dipeptide composition and PSSM matrix. This SVM module was able to achieve an overall accuracy of 84.51% (kernel = RBF, $\gamma$ = 35, $C$ = 4, j = 2), which was almost at par with the hybrid approach-II. The individual accuracy obtained for all four types of subcellular localization are shown in Table 2. However, the overall accuracy of all the hybrid approaches attempted could not exceed the overall accuracy obtained with PSSM-based SVM module. This indicates that PSSM matrices generated from profiles of PSI-BLAST, alone contains more important information about the subcellular locations of proteins as compared to the amino acid compositions.

### 3.5 Terminal-based SVM modules

It is a well-reported fact that many proteins contain signal sequences at their N-terminal region that is recognized by translocation machinery and is responsible for targeting proteins to various subcellular localizations in the cell (chloroplast and mitochondria in the present study). There

**Table 3.** The *p*-values for determining the statistical significance of one module over the other

| Modules developed | AA | Dipep | Four parts | Cumulative higher-order Dipeps | PSSM | AA + Dipep | AA + PSSM | AA + Dipep + PSSM |
|---|---|---|---|---|---|---|---|---|
| AA | – | 0.67252 | 0.83104 | 0.29003 | 0.000195[a] | 0.48623 | 0.00869[a] | 0.00556[a] |
| Dipep | | – | 0.90019 | 0.21426 | 0.000625[a] | 0.35555 | 0.01249[a] | 0.01103[a] |
| Four parts | | | – | 0.31761 | 0.00284[a] | 0.46663 | 0.03632[a] | 0.03877[a] |
| Cumulative higher-order Dipeps | | | | – | 0.01182[a] | 0.81119 | 0.20671 | 0.24887 |
| PSSM | | | | | – | 0.01295[a] | 0.02897[a] | 0.0043[a] |
| AA + Dipep | | | | | | – | 0.1658 | 0.19664 |
| AA + PSSM | | | | | | | – | 0.6984 |
| AA+Dipep+PSSM | | | | | | | | – |

[a] Significant at 0.05 level of significance.

fore, we also attempted various SVM modules based on the N-terminal as well as the C-terminal amino-acid composition of each protein and compared their respective accuracies. The SVM modules were developed at various levels of N-terminal and C-terminal residue length (10, 15, 20, 25 and 30 amino acids) in order to achieve maximum accuracy. Best results were obtained at 25-residue length with RBF kernel for both the terminal regions. At first, the N-terminal-based SVM module alone was able to achieve an overall accuracy of 70.88% ($\gamma = 30$, $C = 1$, j = 3). which was significantly higher compared to an overall accuracy of 64.18% ($\gamma = 40$, $C = 1$, j = 3) for C-terminal based SVM module alone. Further, when remaining part of the protein sequence was also taken into account, the overall accuracy increased drastically. In case of N-terminal + remaining part composition-based SVM module, a 40-dimension vector was used as input information to SVM instead of 20-dimension vector for N-terminal composition alone. The overall accuracy increased to 79.23% (kernel = RBF, $\gamma = 18$, $C = 2$, j = 3) in this case. Similarly, the overall accuracy of C-terminal + remaining part composition-based SVM module also increased to 76.48% (kernel = RBF, $\gamma = 15$, $C = 2$, j = 2) as compared to the C-terminal composition-based SVM module alone. However, best overall accuracy of about 80% (kernel = RBF, $\gamma = 6$, $C = 2$, j = 2) was achieved when amino acid compositions of both the N-terminal (25 residues) and C-terminal (25 residues) as well as the remaining centre part of the protein sequence were combined to provide 60-dimension input vector to the SVM (Table 2). This demonstrates that the SAAC-based approach has greater advantage over the simple N-terminal or C-terminal based amino-acid composition approach(es) as it showed the best overall accuracy and MCC among all the terminal-based modules attempted and greater advantage by providing more weight to proteins that have signal sequences at either the N- or C-terminal region. An overall comparison of accuracies of PSI-BLAST and all the SVM modules developed in this study based on various features of a protein sequence is presented in Fig. 2.

### 3.6 Is module X statistically better than module Y?

As in the present study, we have made a comprehensive and systematic attempt to develop various modules ranging from simple amino acid composition to similarity-search based including using the evolutionary information (PSSM) of a protein sequence, but how do we declare the superiority of one module over the other? For this, we calculated the *p*-values at 0.05 level of significance between each of the two modules based on their performance in fivefold cross-validations. The PSI-BLAST and terminal based modules were left out as they achieved low accuracy level as compared to the other remaining classifiers. The *p*-values as presented in Table 3 revealed that the PSSM-based classifiers were statistically better over all the modules developed in this study with the best overall accuracy achieved by PSSM-based SVM module alone, which was also statistically significant over the other modules.

### 3.7 Reliability index

When machine-learning approaches are followed for protein subcellular localization, it is important to know about the prediction reliability. To evaluate this, the reliability index (RI) assignment was carried out for the overall best method, PSSM-based SVM module. The RI is a measure of confidence in the prediction as it indicates the effectiveness of an approach in the prediction of subcellular localization of proteins. Ideally, the accuracy and probability of correct prediction should increase with the increase in RI values, which is demonstrated in this study as well (Fig. 3). The expected prediction accuracy with RI equal to a given value and the fraction of sequences predicted at each $\geq$ RI value was calculated. Here, we have computed the average prediction accuracy of proteins having a RI value greater than or equal to *n*, where *n* = 1, 2. . ...5. For example, the expected accuracy for a sequence with RI = 2 is 91.80% with 80.44% of sequences having RI $\geq$2. In other words, RSLpred has been

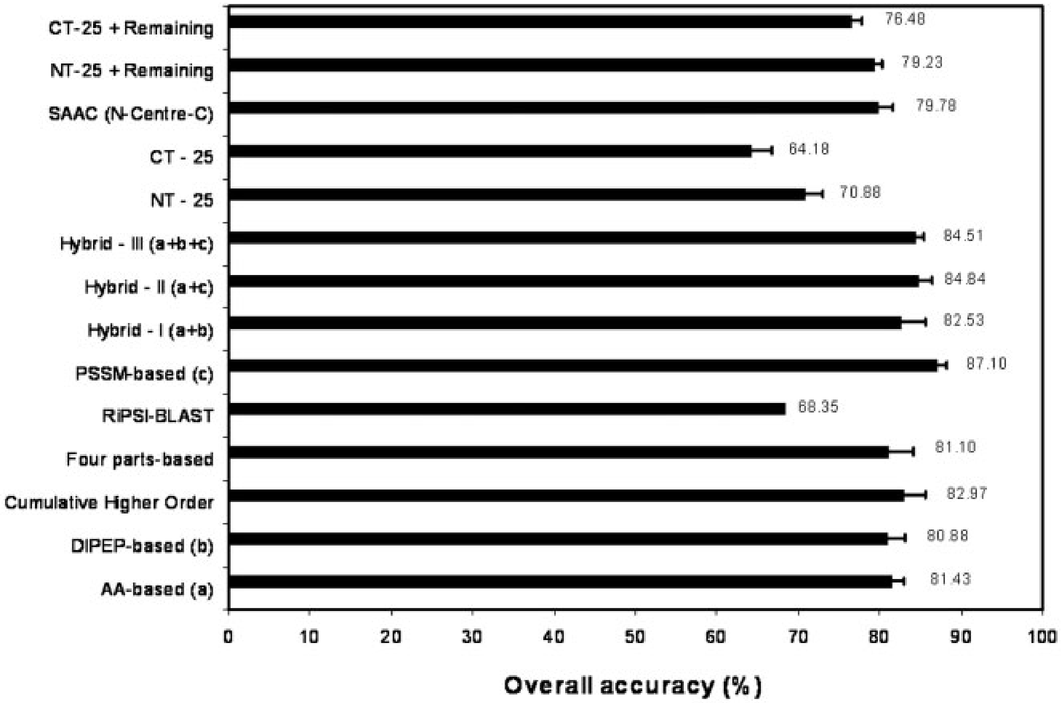


**Figure 2.** Comparison of overall performances of PSI-BLAST and various SVM modules constructed on the basis of different features of a protein sequence.

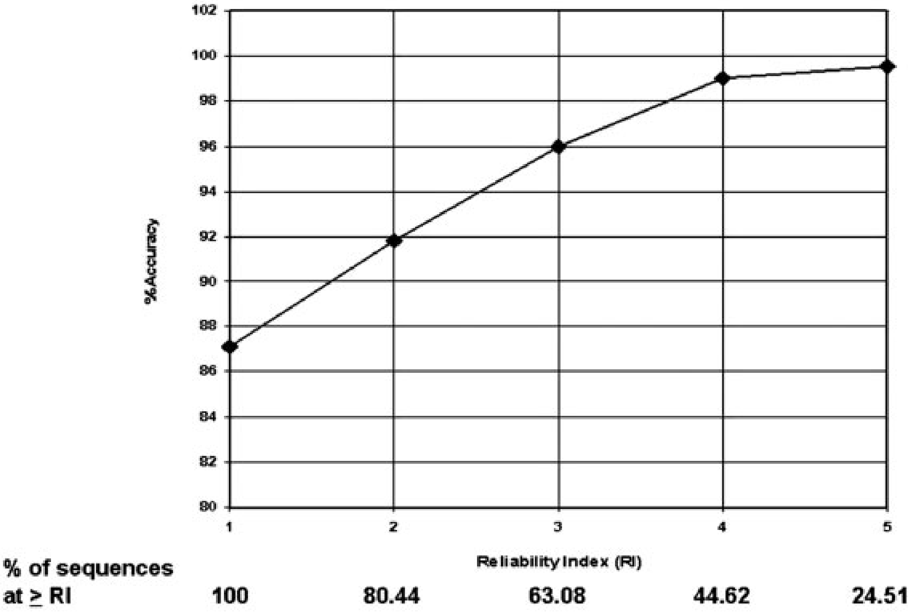


**Figure 3.** Expected prediction accuracy with a reliability index equal to a given value. The fraction of sequences that are predicted with RI ≥ *n*; *n* = 1, 2, 3, 4, 5 are also given.

able to predict 80.44% of sequences with an average prediction accuracy of 91.80% at RI ≥2. This demonstrates that a user can predict a large number of sequences with significantly higher accuracy for RI ≥2. Another calculation showed that RSLpred was able to predict 63.08% of sequences with an accuracy of 95.99% for RI ≥3.

### 3.8 Performance on independent datasets

In order to assess the unbiased performance of any developed method, one needs to evaluate it on an independent dataset. The main objective of the present study was to develop a method for predicting subcellular localization of rice proteins. Since the present method has been trained on specific organism's proteins, it should be more accurate and better for a particular organism as compared to the general methods for predicting plant proteins such as TargetP [16], PA-SUB [10], Wolf PSORT (newer version of PSORT II) [13] and the recently developed all-plant method, Plant-Ploc [19] as well as the general methods for eukaryotic proteins like ESLpred [7]. It has already been reported that ESLpred performs better than NNPSL [5] and SubLoc [8] prediction systems [7]; therefore, we also compared the performance of our method with ESLpred method on these two independent datasets (90 and 25% reduced) of rice proteins, which were not used in the training of original RSLpred method.

It was observed that RSLpred was able to correctly predict 46, 145, 21 and 207 proteins out of 58, 165, 58 and 227 (chloroplast, cytoplasmic, mitochondrial and nuclear) proteins, respectively, from the independent dataset-I, using the PSSM-based module. An overall accuracy of 82.48% has been achieved, whereas the Wolf PSORT method was able to correctly predict only 32, 56, 9 and 165 (chloroplast, cytoplasmic, mitochondrial and nuclear) proteins, respectively, with an achievable overall accuracy of 51.58% on this dataset-I (Table 4a). The recently developed all-plant method, Plant-Ploc, showed very low prediction performance as it correctly predicted only 46, 31, 4 and 94 (chloroplast, cytoplasmic, mitochondrial and nuclear) proteins, respectively with an overall accuracy of 34.45%. Similarly, TargetP, PA-SUB and ESLpred methods also performed poorly on this independent dataset-I with an overall accuracy of 31.04%, 52.76% and 53.11%, respectively.

Furthermore, when repeating the above same procedure on independent dataset-II, it was observed that even at 25% redundancy-reduced sequences, RSLpred was able to predict 30, 94, 18 and 140 proteins correctly out of 38, 109, 50 and 148 (chloroplast, cytoplasmic, mitochondrial and nuclear) proteins, respectively, showing an overall accuracy of 81.74% as compared to the low performance shown by other general methods available. We observed that the performance of our method as well as all other methods did not drop significantly by using a much diverse independent dataset-II (25% cutoff) in comparison to the 90% cutoff dataset-I. The accuracy level was found to be almost same on both these datasets.

The results on these two test datasets demonstrate that there must be some species-specific features of protein sorting in rice, thereby indicating that genome-specific prediction methods for subcellular localization of proteins are far much better than the general methods. The detailed comparison of all these web tools on each of the four subcellular localizations under study is presented in Table 4 (a, b). Hence, there is an urgent need for developing organism-specific methods for more reliable and accurate prediction of subcellular localization of proteins, which can ultimately accelerate the annotation of huge genomic data for those organisms.

### 3.9 Comparison with newly developed 'All Plant' method

Though the above comparison indicates towards the advantages of developing a species-specific predictor(s), one can think of another interesting question that whether the inclusion of non-rice proteins in the training set would make RSLpred perform better or worse on the rice independent proteins, which still remains unclear so far; and so probably needs to be further elaborated or strengthened. The only way to confidently answer this question is to train a correspond-

**Table 4.** Performance of RSLpred in comparison to other methods on (a) independent dataset-I (90% cutoff) and (b) independent dataset-II (25% cutoff) of rice proteins

| Subcellular location | Number of sequences | RSLpred % accuracy | TargetP % accuracy | Plant-Ploc % accuracy | Wolf-PSORT % accuracy | PA-SUB % accuracy | ESLpred % accuracy |
|---|---|---|---|---|---|---|---|
| **Dataset I** | | | | | | | |
| Chloroplast | 58 | 79.31 (46)[a] | 36.21 (21) | 79.31 (46) | 55.17 (32) | 44.83 (26) | [b] |
| Cytoplasm | 165 | 87.88 (145) | [b] | 18.79 (31) | 33.94 (56) | 24.24 (40) | 37.58 (62) |
| Mitochondria | 58 | 36.21 (21) | 25.86 (15) | 06.90 (04) | 15.52 (9) | 25.86 (15) | 22.41 (13) |
| Nuclear | 227 | 91.19 (207) | [b] | 41.41 (94) | 72.69 (165) | 82.38 (187) | 72.25 (164) |
| % Overall accuracy | 508 | 82.48 (419) | 31.04 (36) | 34.45 (175) | 51.58 (262) | 52.76 (268) | 53.11 (239) |
| **Dataset II** | | | | | | | |
| Chloroplast | 38 | 78.95 (30) | 36.84 (14) | 78.95 (30) | 55.26 (21) | 47.37 (18) | [b] |
| Cytoplasm | 109 | 86.24 (94) | [b] | 20.18 (22) | 36.70 (40) | 27.52 (30) | 36.70 (40) |
| Mitochondria | 50 | 36.00 (18) | 26.00 (13) | 02.00 (01) | 14.00 (7) | 26.00 (13) | 20.00 (10) |
| Nuclear | 148 | 94.59 (140) | [b] | 40.54 (60) | 70.95 (105) | 79.05 (117) | 74.32 (110) |
| % Overall accuracy | 345 | 81.74 (282) | 30.68 (27) | 32.75 (113) | 50.14 (173) | 51.59 (178) | 52.12 (160) |

[a] Values in parentheses represent number of correctly predicted sequences.

[b] Prediction not available.

ing method (using the same encoding method and location definitions used in RSLpred) on a dataset derived from all the plant species, and then compare the performance of two methods on the rice independent dataset.

To proceed, we downloaded all the plant proteins having subcellular localization information available from the latest UniProtKB/Swiss-Prot release 55.5 (total 17,308 sequences) and extracted the protein sequences for each of the four subcellular classes under study as in RSLpred (total 13 384 sequences, see Table 5a for distribution). As in RSLpred training data, we used 90% cutoff for reducing the sequence redundancy among the rice data, similar approach was also followed while proceeding for developing an 'All Plant' classifier. Therefore, using the same PROSET software [26], we also reduced the redundancy of 'All Plant' sequence dataset to 90% sequence identity level. As we wanted to compare the performance of rice independent dataset on both the RSLpred and this 'All Plant' method; for fair comparison, we further removed all the rice independent sequences from this 'All Plant' training dataset (Table 5a). In this way, we made sure that both the RSLpred as well as the 'All Plant' classifier had not been trained from any of the sequences in the rice independent dataset.

Finally, the traditional amino acid (AA) composition-based classifier for 'All Plant' dataset was developed following the same fivefold cross-validation approach and evaluation parameters as used in RSLpred development. In the end, we ran all of the rice independent dataset-I on the model files generated from this 'All Plant' classifier and compared its performance with the AA-based classifier of RSLpred (Note: our best classifier in RSLpred was obtained from PSSM matrix using evolutionary information of a protein sequence, but to avoid any confusion and compare similar encoding schemes, we have used RSLpred's simple amino acid composition-based classifier to compare with 'All Plant' amino acid composition-based classifier). Results are presented in Table 5b and are discussed herewith.

Technically, due to the larger training dataset and more number of rice sequences present in the 'All Plant' dataset, the AA-based classifier of 'All Plant' should have performed better than the AA-model files of RSLpred; however surprisingly, RSLpred even outperformed the 'All Plant' method significantly (Table 5b). This indicates that there might be some differences in the sorting signals and mechanisms between species, which enables a higher performance of a method designed for a specific species (rice in this case). To support this, some previous methods of multivariate analysis used to study the amino acid residue composition have also lead to the identification of species-specific compositional patterns [48]. Furthermore, it has been shown in the past that not only amino acid composition but also oligopeptide frequencies (dipeptides, tripeptides etc.) reflect independent segregation between species and there are several identified distinct factors that shape the landscape of species-specific proteomic composition [49]; thereby suggesting that all these general methods for predicting subcellular localization might be skipping these species-specific compositional patterns in their training process and learning only from the common patterns whereas, the similar encoding scheme(s) followed in an organism-specific prediction system is able to learn more efficiently from these species-specific compositional patterns, giving us a more efficient prediction model(s) and which, ultimately leads to the higher prediction accuracy of individual genome annotation. This also suggests that methods relying on amino acid composition should also take into account their species-specific background.

**Table 5a.** Number of protein sequences within each subcellular location group for all the plants in Swiss-Prot database

| Subcellular location | Sequences available | 90% cut-off | Final dataset (rice independent dataset reduced) |
|---|---|---|---|
| Chloroplast | 9456 | 4101 | 4043 |
| Cytoplasm | 1170 | 940 | 775 |
| Mitochondria | 820 | 685 | 627 |
| Nucleus | 1938 | 1789 | 1562 |
| Total | 13 384 | 7515 | 7007 |

**Table 5b.** Performance comparison of rice independent dataset-I on RSLpred and newly developed 'All Plant' method

| Subcellular location | Number of sequences | RSLpred % accuracy[a)] | All Plant % accuracy[a)] |
|---|---|---|---|
| Chloroplast | 58 | 60.34 (35)[b)] | 43.10 (25)[b)] |
| Cytoplasm | 165 | 42.42 (70) | 27.27 (45) |
| Mitochondria | 58 | 36.21 (21) | 12.07 (07) |
| Nuclear | 227 | 90.31 (205) | 78.41 (178) |
| **% Overall accuracy** | 508 | **65.16** (331) | **50.20** (255) |

a) Using amino acid composition-based classifier.

b) Values in parentheses represent number of correctly predicted sequences.

## 3.10 Confusion matrix

Though RSLpred performed much better on both the independent datasets (I, II) of rice proteins than the other web tools available as well as the above-mentioned 'All Plant' method, one leads to wonder about the false positive classifications. For example, if chloroplast proteins are only classified with 55% accuracy, are the chloroplast proteins that are incorrectly classified equally likely to be scored in any of the other three categories? Alternatively, are false positive predictions always classified in a single false category? Is there any biological basis for something systematic with the false positives? For this, two additional 'confusion matrix' Tables (6a, 6b) were generated from the performance of RSLpred on these independent datasets of rice proteins as presented in

**Table 6.** Confusion matrix for predictions on (a) independent dataset-I (90% cutoff) and (b) independent dataset-II (25% cutoff) of rice proteins[a)]

| | Chloroplast | Cytoplasm | Mitochondria | Nuclear | Actual Sum | Sensitivity (%) |
|---|---|---|---|---|---|---|
| **(a) Dataset I** | | | | | | |
| Chloroplast (58) | **46** | 2 | 7 | 3 | 58 | 79.31 |
| Cytoplasm (165) | 2 | **145** | 3 | 15 | 165 | 87.88 |
| Mitochondria (58) | 2 | 5 | **21** | 30 | 58 | 36.21 |
| Nuclear (227) | 2 | 11 | 7 | **207** | 227 | 91.19 |
| Predicted Sum | 52 | 163 | 38 | 255 | **508** | **$\bar{R}$ = 82.48** |
| Precision (%) | 88.46 | 88.96 | 55.26 | 81.18 | | **$\bar{P}$ = 81.58** |
| Specificity (%) | 98.67 | 94.75 | 96.22 | 82.92 | | **$\bar{S}$ = 90.08** |
| **(b) Dataset II** | | | | | | |
| Chloroplast (38) | **30** | 1 | 6 | 1 | 38 | 78.95 |
| Cytoplasm (109) | 1 | **94** | 2 | 12 | 109 | 86.24 |
| Mitochondria (50) | 2 | 3 | **18** | 27 | 50 | 36.00 |
| Nuclear (148) | 1 | 4 | 3 | **140** | 148 | 94.59 |
| Predicted Sum | 34 | 102 | 29 | 180 | **345** | **$\bar{R}$ = 81.74** |
| Precision (%) | 88.24 | 92.16 | 62.07 | 77.78 | | **$\bar{P}$ = 81.20** |
| Specificity (%) | 98.70 | 96.61 | 96.27 | 79.70 | | **$\bar{S}$ = 89.54** |

a) 'Actual Sum' and 'Predicted Sum' are the sums of the actual and predicted instances for each class, respectively. $\bar{R}$, $\bar{P}$ and $\bar{S}$ denote overall sensitivity (recall), precision and specificity, respectively.

Table 4 (a, b). Both datasets showed the same pattern of distribution for the false predictions. Though showing good recall (sensitivity) rate for chloroplast (79.31% for dataset-I and 78.95% for dataset-II), it was observed that most of the false positive chloroplast proteins were predicted towards mitochondrial false category. The biological basis for the same may be the fact that chloroplasts are similar to mitochondria in many ways, except since they are involved in photosynthesis, they only occur in eukaryotic autotrophs. Both these organelles likely resulted from a symbiotic relationship so they encode some of their own proteins [50]. Moreover, both have their signal sequence located at the N terminus of the protein. Similarly, most of the cytoplasmic false positives were found to be wrongly targeted to nucleus and *vice versa* (Table 6a, b); however, the highest recall rate was achieved for these two categories (cytoplasm = 87.88%, nucleus = 91.19% for dataset-I; and cytoplasm = 86.24%, nucleus = 94.59% for dataset-II). The nuclear proteins are also translated in the cytoplasm and are transported back into the nucleus to do their jobs. In addition, the nuclear localization signal (NLS) can be located anywhere in the protein sequence, which is not the case in chloroplast and mitochondrial proteins. Moreover, the NLS signal is not cleaved here and the transport mechanism is somewhat more complicated than the chaperone binding of chloroplast and mitochondrial proteins. That is why some of the cytoplasmic proteins get confused with the nuclear targeting proteins and vice versa.

Secondly, the results presented in Tables 4 (a, b) and 6 (a, b) revealed that even in the fine print, mitochondrial proteins showed poor prediction (sensitivity = 36.21% on dataset-I and 36.00% on dataset-II) as compared to the other cellular locations under study. This was probably due to the two main reasons. First, it has been proved in some of the earlier studies that mitochondrial proteins are notoriously difficult to predict [51, 52]. Secondly, the number of rice-specific mitochondrial proteins was less in the original training dataset. This probably could also have affected the prediction accuracy of mitochondrial proteins when independent sets of rice mitochondrial proteins were validated/tested. However, compared to the other methods available (TargetP, PA-SUB, Wolf-PSORT, Plant-Ploc and ESLpred) as shown in Table 4 (a, b), our method outperformed significantly on both the 90% cutoff as well as 25% cutoff independent datasets. It means that even with the less number of rice mitochondrial entries in the training set, the module could correctly predict 21 sequences out of the 58 available in the independent set-I and 18 correctly out of 50 in dataset-II, as compared to the lower accuracy of other methods.

Furthermore, when RSLpred model files were run on other plant species to crosscheck its performance (see Supporting Information Table 2), the mitochondrial proteins of *Arabidopsis thaliana* were predicted more correctly as compared to the other classes. This probably reflects the high proportion of *A. thaliana* data of the mitochondrial category in the RSLpred training set and suggests that a significant number of proteins of a particular class are needed to train an efficient and reliable classifier. Henceforth, with the future increase in the number of rice specific mitochondrial protein sequences in the Swiss-Prot database, the said module will be upgraded at regular intervals to achieve better and improved accuracy level of the said classifier.

## 3.11 Annotation of rice proteome

The subcellular predictions on both the proteomic datasets (EBI and TIGR) for chloroplast, cytoplasm, mitochondria and nuclear proteins are presented in Table 7. For highly re-

**Table 7.** Performance of RSLpred server on complete rice proteomes retrieved from EBI and TIGR. The predictions were done using the traditional AA-based module

| Subcellular location | Predictions at threshold | | | | |
|---|---|---|---|---|---|
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| **EBI proteome (30 952 sequences)** | | | | | |
| Chloroplast | 1728 | 1460 | 1237 | 993 | 792 |
| Cytoplasm | 2536 | 2162 | 1765 | 1452 | 1150 |
| Mitochondria | 5161 | 4544 | 3894 | 3270 | 2726 |
| Nuclear | 10 366 | 9223 | 8173 | 7244 | 6310 |
| Total | 19 791 | 17 389 | 15 069 | 12 959 | 10 978 |
| **TIGR proteome (62 827 sequences)** | | | | | |
| Chloroplast | 2728 | 2250 | 1856 | 1461 | 1142 |
| Cytoplasm | 3916 | 3279 | 2693 | 2169 | 1738 |
| Mitochondria | 10 383 | 9134 | 7892 | 6678 | 5561 |
| Nuclear | 24 232 | 21 644 | 19 130 | 16 803 | 14 572 |
| Total | 41 259 | 36 307 | 31 571 | 27 111 | 23 013 |

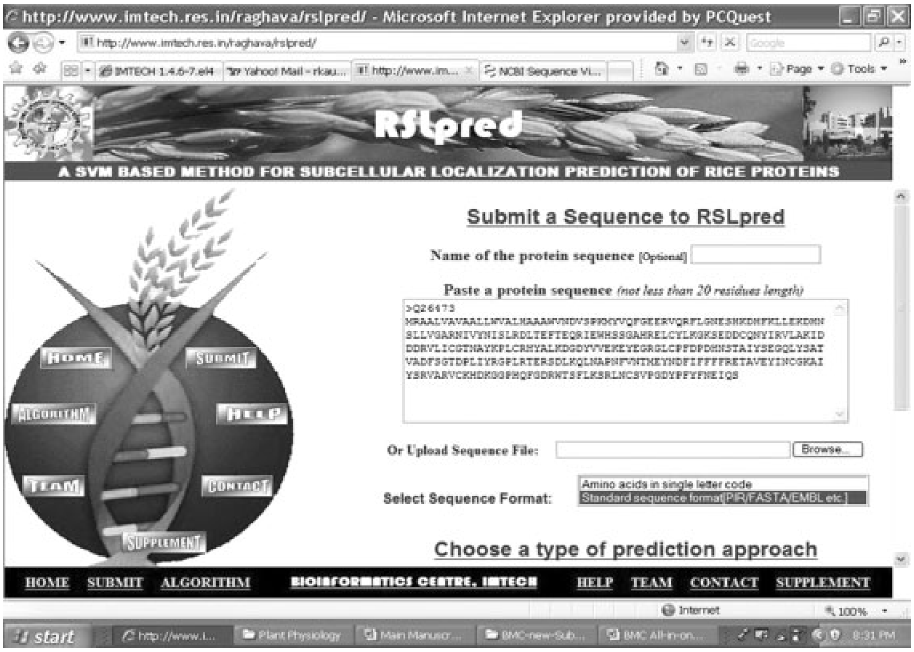
liable and accurate predictions, we put various levels of threshold values (0.1, 0.2, 0.3, 0.4 and 0.5) on the final score for each cellular class. For example, if the maximum score of a query protein was found for the chloroplast category; in the next step, we checked whether this score was more than the threshold value or not. Only then, we declared the query protein as predicted to be chloroplast. Therefore, one can say that higher the threshold value, more reliable are the predictions. From the results presented in Table 7, it was observed that on EBI proteome set; about 792 proteins are predicted to be chloroplast at >0.5 threshold value (high confidence), which stands to about 2.56% of the total proteome. Similarly, 1150 (3.72%) cytoplasmic, 2726 (8.81%) mitochondrial and 6310 (20.39%) nuclear proteins are predicted to be present in the rice proteome with high reliability and confidence level. On TIGR proteome set, about 1142 (1.82%) chloroplast, 1738 (2.77%) cytoplasmic, 5561 (8.85%) mitochondrial and 14 572 (23.20%) nuclear proteins are predicted to be correctly localized to the said cellular compartments with high confidence (>0.5 threshold). The complete list of Uniprot Knowledgebase ID/TIGR Locus identifiers of the predicted protein sequences has been provided on the RSLpred web server. The authors believe that the above information generated from whole rice proteome analysis will not only facilitate the application of subcellular localization information in whole proteome annotation but also provides a foundation for functional and evolutionary studies of other important cereal crops as well.

### 3.12 Description of web server

The best performing modules from the present investigation have been implemented on the World Wide Web as a dynamic web server 'RSLpred', which is freely available and can be assessed by at http://www.imtech.res.in/raghava/rslpred/. All the CGI scripts of RSLpred were written in PERL and the interface was designed using HTML to assess user queries. It is a user-friendly web server and allows users to submit their protein sequence in one of the standard formats such as FASTA, GenBank, EMBL, GCG or plain format (Fig. 4). Users can choose to type or paste the sequence in the box, or upload the sequence through a file. The server provides options to select various approaches for the prediction of subcellular localization of a query protein sequence. Due to PSI-BLAST searches and generation of profiles in the form of a PSSM table, the prediction of subcellular localization of the query sequence through PSSM-based approach may take some time to serve the query. For rest of the approaches, the prediction results will be displayed in a user-friendly format on the screen within a few seconds (Fig. 5). As the PSSM-based classifier is a little slower, we suggest that for larger analysis/predictions, the users should opt for a faster classifier like the traditional amino acid composition-based module. For more flexibility, we have provided four other good performing modules, *e.g.* if the user wishes to use terminal-based information of the input query sequence for prediction purpose, he/she may opt for splitted amino-acid composition module, which is based on N-Centre-C-terminal composition of the protein sequence. If the user wishes to utilize the sequence order effects of the query sequence, the dipeptide composition-based module may be used for prediction. However, if the users still want to use PSSM classifier for larger predictions, they may directly contact the authors for assistance. In case of the default prediction, RSLpred uses the PSSM-based module for prediction. An overall architecture of the 'RSLpred' web server is shown in Fig. 6.

## 4 Concluding remarks

An era of biological revolution has begun during which a tremendous amount of information on plant genetics will be accumulated over the next ten years. Rice, at a compact


http://www.imtech.res.in/raghava/rslpred/ - Microsoft Internet Explorer provided by PCQuest
http://www.imtech.res.in/raghava/rslpred/
IMTECH 1.4.6-7.el4
Yahoo! Mail - rkau...
http://www.im...
NCBI Sequence Vi...
Page
Tools
RSLpred
A SVM BASED METHOD FOR SUBCELLULAR LOCALIZATION PREDICTION OF RICE PROTEINS
HOME
SUBMIT
ALGORITHM
HELP
TEAM
CONTACT
SUPPLEMENT
BIOINFORMATICS CENTRE, IMTECH
Internet
100%
start
http://www.i...
Plant Physiology
Main Manuscr...
BMC-new-Sub...
BMC All-in-on...
8:31 PM

Submit a Sequence to RSLpred

Name of the protein sequence [Optional]

Paste a protein sequence *(not less than 20 residues length)*

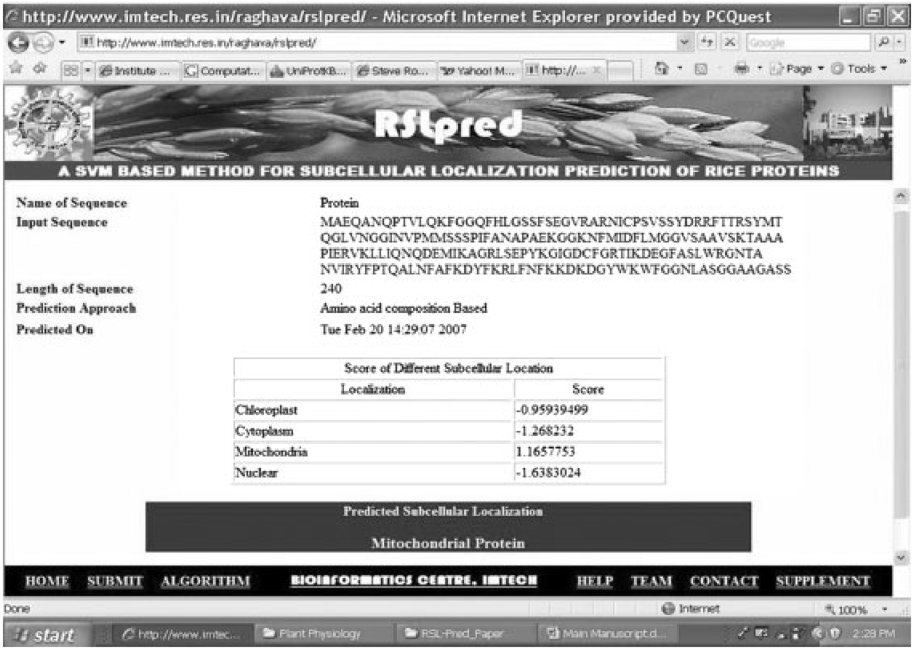
```
>Q26473
MRAALVAVAALLWVALHAAAWVNDVSPKMYVQFGEERVQRFLGNESHKDHFKLLEKDMN
SLLVGARNIVYNISLRDLTEFTEQRIEWHSSGAHRELCYLKGKSEDDCQNYIRVLAKID
DDRVLICGTNAYKPLCRHYALKDGDYVVEKEYEGRGLCPFDPDHNSTAIYSEGQLYSAT
VADFSGTDPLIYRGPLRTERSDLKQLNAPNFVNTMEYNDFIFFFFRETAVEYINCGKAI
YSRVARVCKHDKGGPHQFGDRWTSFLKSRLNCSVPGDYPFYFNEIQS
```

Or Upload Sequence File: Browse...

Select Sequence Format: Amino acids in single letter code / Standard sequence format[PIR/FASTA/EMBL etc.]

Choose a type of prediction approach

**Figure 4.** An overview of submission form for online subcellular localization prediction of rice proteins with 'RSLpred' web server.


http://www.imtech.res.in/raghava/rslpred/ - Microsoft Internet Explorer provided by PCQuest
http://www.imtech.res.in/raghava/rslpred/
Institute ...
Computat...
UniProtKB...
Steve Ro...
Yahoo! M...
http://...
Page
Tools
RSLpred
A SVM BASED METHOD FOR SUBCELLULAR LOCALIZATION PREDICTION OF RICE PROTEINS
HOME
SUBMIT
ALGORITHM
BIOINFORMATICS CENTRE, IMTECH
HELP
TEAM
CONTACT
SUPPLEMENT
Done
Internet
100%
start
http://www.imtec...
Plant Physiology
RSL-Pred_Paper
Main Manuscript.d...
2:28 PM

| | |
|---|---|
| **Name of Sequence** | Protein |
| **Input Sequence** | MAEQANQPTVLQKFGGQFHLGSSFSEGVRARNICPSVSSYDRRFTTRSYMT<br>QGLVNGGINVPMMSSSPIFANAPAEKGGKNFMIDFLMGGVSAAVSKTAAA<br>PIERVKLLIQNQDEMIKAGRLSEPYKGIGDCFGRTIKDEGFASLWRGNTA<br>NVIRYFPTQALNFAFKDYFKRLFNFKKDKDGYWKWFGGNLASGGAAGASS |
| **Length of Sequence** | 240 |
| **Prediction Approach** | Amino acid composition Based |
| **Predicted On** | Tue Feb 20 14:29:07 2007 |

| Score of Different Subcellular Location | |
|---|---|
| Localization | Score |
| Chloroplast | -0.95939499 |
| Cytoplasm | -1.268232 |
| Mitochondria | 1.1657753 |
| Nuclear | -1.6383024 |

Predicted Subcellular Localization

Mitochondrial Protein

**Figure 5.** A snapshot of output page displaying all the four scores for each subcellular location including the final prediction.

430 Mb, is only one-sixth the size of the human and maize genomes and provides the sequencing template for all the grasses. This includes every significant grain crop (including sorghum, maize, barley, oat, and wheat), most of which have enormous genomes that are not feasible to sequence at current costs. Understanding the biological functions of about 37 544 genes identified through International Rice Genome Sequencing Project, identifying the function and regulation of each encoded protein and whole genome functional annotation (assigning functions to potential gene(s) in the raw sequences) is now the hot topic in rice bioinformatics. Moreover, rice, as a model species, is the plant in which the function of most cereal genes will be discovered. Thus, the availability of systems that can predict location from sequence is essential to the full characterization of expressed proteins. Experimentally determining the subcellular loca-

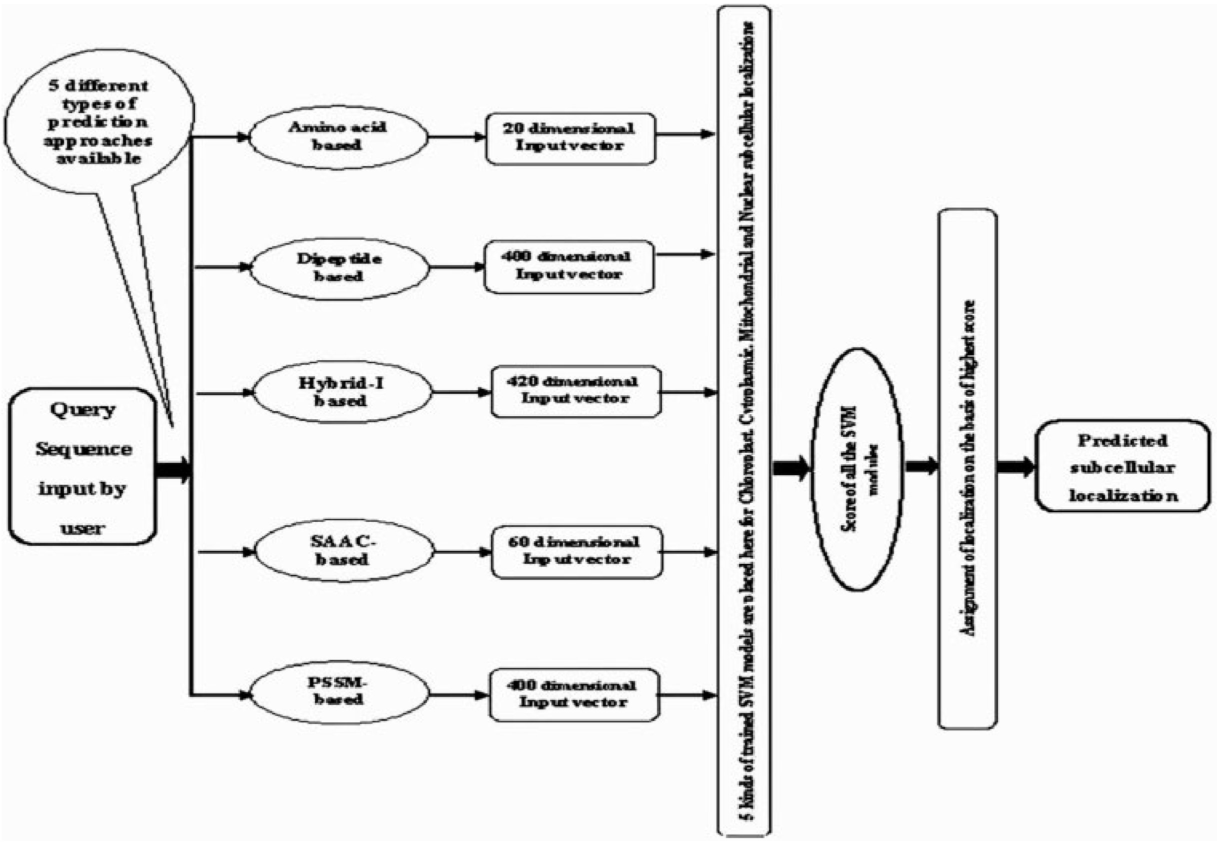

5 different types of prediction approaches available
Query Sequence input by user
Amino acid based
Dipeptide based
Hybrid-I based
SAAC-based
PSSM-based
20 dimensional Input vector
400 dimensional Input vector
420 dimensional Input vector
60 dimensional Input vector
400 dimensional Input vector
5 Kinds of trained SVM models are placed here for Chloroplast, Cytoplasmic, Mitochondrial and Nuclear subcellular localizations
Score of all the SVM modules
Assignment of localization on the basis of highest score
Predicted subcellular localization

**Figure 6.** Overall architecture of the RSLpred web server.

tion is a laborious and time-consuming task. Computational tools provide faster and accurate access to localization predictions for any organism.

A number of computational prediction methods are available, but all these methods have limitations in terms of their accuracy and breadth of coverage when organism-specific predictions are made. From the present investigation, it was concluded that the available prediction methods for subcellular localization of plant proteins are less accurate in predicting rice-specific destination of proteins. In this direction, a novel method for subcellular localization of rice proteins is presented, which will assist in assigning the subcellular location or function of rice proteins more reliably with higher accuracy. We also suggest that evolutionary information stored in the position-specific scoring matrices of a protein sequence combined with a suitable machine learning technique provides more comprehensive and reliable information about the subcellular localizations of proteins as compared to the amino acid composition-based methods. The authors believe that the prediction method presented here would act as a useful tool for the functional annotation of the recently piled-up rice genomic data and ultimately, for providing crucial insights into genome evolution, speciation, domestication and the development of improved strains of rice as well.

*The assistance rendered by Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India for financial assistance is duly acknowledged. The authors are grateful to Mr. Mamoon Rashid and Mrs. Aarti Garg of this centre for assisting in web server development. This report has IMTECH communication no. 030/2006.*

*The authors have declared no conflict of interest.*

## 5 References

[1] International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature* 2005, *436*, 793–800.

[2] Nakai, K., Kanehisa, M., Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins* 1991, *11*, 95–110.

[3] Nakai, K., Kanehisa, M., A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 1992, *14*, 897–911.

[4] Nair, R., Rost, B., Inferring subcellular localization through automated lexical analysis. *Bioinformatics* 2002, *18*, S78–S86.

[5] Reinhardt, A., Hubbard, T., Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* 1998, *26*, 2230–2236.

[6] Fujiwara, Y., Asogawa, M., Prediction of subcellular localizations using amino acid composition and order. *Genome Informatics* 2001, *12*, 103–112.

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

[8] Hua, S., Sun, Z., Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 2001, *17*, 721–728.

[9] Nair, R., Rost, B., Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.* 2005, *348*, 85–100.

[10] Lu, Z., Szafron, D., Greiner, R., Lu, P. *et al.*, Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* 2004, *20*, 547–556.

[11] Gardy, J. L., Spencer, C., Wang, K., Ester, M. *et al.*, PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.* 2003, *31*, 3613–3617.

[12] Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S., Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 2002, *18*, 298–305.

[13] Horton, P., Park, K. J., Obayashi, T., Nakai, K., Protein subcellular localization prediction with WoLF PSORT. *Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference APBC06, Taipei, Taiwan*, 2006, 39–48.

[14] Chou, K. C., Shen, H. B., Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic *K*-nearest neighbor classifiers. *J. Proteome Res.* 2006, *5*, 1888–1897.

[15] Xie, D., Li, A., Wang, M., Fan, Z., Feng, H., LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.* 2005, *33*, 105–110.

[16] Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 2000, *300*, 1005–1016.

[17] Garg, A., Bhasin, M., Raghava, G. P. S., Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.* 2005, *280*, 14427–14432.

[18] Chou, K. C., Shen, H. B., Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* 2006, *347*, 150–157.

[19] Chou, K. C., Shen, H. B., Large-scale plant protein subcellular location prediction. *J. Cell. Biochem.* 2007, *100*, 665–678.

[20] Chou, K. C., Prediction of protein signal sequences. *Curr. Protein Pep. Sci.* 2002, *3*, 615–622.

[21] Wang, M., Li, A., Xie, D., Fan, Z., Feng, H., Improving prediction of protein subcellular localization using evolutionary information and sequence-order information. *27th Annual International Conference of the IEEE-EMBS* 2005, 4434–4436.

[22] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., Basic local alignment search tool. *J. Mol. Biol.* 1990, *215*, 403–410.

[23] Schneider, G., Fechner, U., Advances in the prediction of protein targeting signals. *Proteomics* 2004, *4*, 1571–1580.

[24] Schneider, G., Sjöling, S., Wallin, E., Wrede, P. *et al.*, Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *Proteins: Struct. Funct. Genet.* 1998, *30*, 49–60.

[25] Kohonen, T., Self-Organized formation of topologically correct feature maps. *Biol. Cybern.* 1982, *43*, 59–69.

[26] Brendel, V., PROSET - a fast procedure to create non-redundant sets of protein sequences. *Math. Comput. Modeling* 1992, *16*, 37–43.

[27] Cortes, C., Vapnik, V., Support vector networks. *Machine Learning* 1995, *20*, 273–293.

[28] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, New York 1995.

[29] Brown, M. P. S., Grundy, W. N., Lion, D., Cristianini, N. *et al.*, Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 2000, *97*, 262–297.

[30] Kaundal, R., Kapoor, A. S., Raghava, G. P. S., Machine learning techniques in disease forecasting: a case study on rice blast prediction. *BMC Bioinformatics* 2006, *7*, 485.

[31] Ward, J. J., McGuffin, L. J., Buxton, B. F., Jones, D. T., Secondary structure prediction with support vector machines. *Bioinformatics* 2003, *19*, 1650–1655.

[32] Vapnik, V., *Statistical Learning Theory*, Wiley, New York 1998.

[33] Joachims, T., in: Schölkopf, B., Burges, C., Smola, A. (Eds.), *Advances in Kernel Methods - Support Vector Learning*, MIT-Press, Massachusetts 1999, pp. 41–56.

[34] Jones, D. T., Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 1999, *292*, 195–202.

[35] Chou, K. C., Zhang, C. T., Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* 1995, *30*, 275–349.

[36] Zhang, S. W., Pan, Q., Zhang, H. C., Shao, Z. C., Shi, J. Y., Prediction of protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. *Amino Acids* 2006, *30*, 461–468.

[37] Chen, C., Zhou, X., Tian, Y., Zou, X., Cai, P., Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Anal. Biochem.* 2006, *357*, 116–121.

[38] Mondal, R., Bhavna, R., Babu, M., Ramakumar, S., Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J. Theor. Biol.* 2006, *243*, 252–260.

[39] Zhou, G. P., Doctor, K., Subcellular location prediction of apoptosis proteins. *Proteins: Structure, Function and Genetics* 2003, *50*, 44–48.

[40] Guo, J., Lin, Y., Liu, X., GNBSL: A new integrative system to predict the subcellular location for Gram-negative bacteria proteins. *Proteomics* 2006, *6*, 5099–5105.

[41] Cao, Y., Liu, S., Zhang, S., Qin, J. *et al.*, Prediction of protein structural class with Rough Sets. *BMC Bioinformatics* 2006, *7*, 20.

[42] Yu, J., Hu, S., Wang, J., Wong, G. K. S. *et al.*, A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 2002, *296***,** 79–92.

[43] Chou, K. C., A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins* 1995, *21*, 319–344.

[44] Horton, P., Nakai, K., Better prediction of protein cellular localization sites with the *k* nearest neighbors classifier. *Proceedings of the 5th ISMB*, AAAI Press 1997, 298–305.

[45] Chou, K. C., Elrod, D. W., Protein subcellular location prediction. *Protein Eng.* 1999, *12*, 107–118.

[46] Yuan, Z., Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.* 1999, *451*, 23–26.

[47] Chou, K. C., Cai, Y. D., Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* 2002, *277*, 45765–45769.

[48] Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J., Gentles, A.J., Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 333–338.

[49] Pe'er, I., Felder, C.E., Man, O., Silman, I. *et al.*, Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. *Proteins* 2004, *54*, 20–40.

[50] Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., *Molecular Cell Biology* (fifth edition), W. H. Freeman and company, New Jersey 2004.

[51] Peng, W. M., Rajapakse, J. C., Multi-class protein subcellular localization prediction using support vector machines. *Proceedings of the IEEE Symposium* 2005, *online* 2006-02-21 09:22:31.0.

[52] Sarda, D., Chua, G. H., Li, K. B., Tang, F., Krishnan, A., pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics* 2005, *6*, 152.