AN INTEGRATIVE SYSTEM FOR PREDICTION OF NAC PROTEINS IN RICE USING DIFFERENT FEATURE EXTRACTION METHODS

Hemalatha N.[1,*], Rajesh M. K.[2] and Narayanan N. K.[3]

[1]AIMIT, St. Aloysius College, Mangalore, India
hemasree71@gmail.com
[2]Division of Crop Improvement, Central Plantation Crops Research Institute, Kasaragod 671124, India
mkraju_cpcri@yahoo.com
[3]School of Information Science and Technology, Kannur University, Kannur, India.
csirc@rediffmail.com

ABSTRACT

The NAC gene family encodes a large family of plant-specific transcription factors with diverse roles in various developmental processes and stress responses in plants. Creation of genome wide prediction tools for NAC proteins will have a significant impact on gene annotation in rice. In the present study, NACSVM, a tool for computational genome-scale prediction of NAC proteins in rice was developed integrating compositional and evolutionary information of NAC proteins. Initially, support vector machine (SVM)-based modules were developed using combinatorial presence of diverse protein features such as traditional amino acid, dipeptide (i+1), tripeptide (i+2), four-parts composition and PSSM and an overall accuracy of 79%, 93%, 93%, 79% and 100% respectively was achieved. Later, two hybrid modules were developed based on amino acid, dipeptide and tripeptide composition, through which an overall accuracy of 83% and 79% was achieved. NACSVM was also evaluated using position-specific iterated – basic local alignment search tool which resulted in a lower accuracy of 50%. In order to benchmark NACSVM , the tool was evaluated using independent data test and cross validation methods. The different statistical analyses carried out revealed that the proposed algorithm is an useful tool for annotating NAC proteins in genome of rice.

KEYWORDS

SVM, NAC, RBF, PSSM, ROC, AUC

1. INTRODUCTION

Rice (*Oryza sativa* L.) is the staple food for millions of people in South Asia, Southeast Asia and sub-Saharan Africa [1]. It is considered to be the model genome as well as experimental model for comprehending the biology of all cereals. Rice production is subjected to a number of abiotic and biotic stresses, as a result of which potential yields cannot be achieved. Drought is a major problem in traditionally irrigated areas, which adversely affects the growth and productivity of rice-based farming systems. Assimilation of water stress tolerance into high-yielding rice varieties has proven to be a very successful approach to developing varieties that can cope with these extreme situations.

DOI: 10.5121/ijsc.2013.4102

The method of molecular response of plants to abiotic stresses has been investigated by a complete study of genes unregulated under the specific stress conditions. One class of important transcription factors induced during abiotic stress tolerance is NAC. NAC (NAM, ATAF1, -2, and CUC2) are genes which encodes a polypeptide containing a plant-specific highly conserved N-terminal domain. They are represented by more or less 140 genes in rice [2]. They regulate by binding to specific cis-acting promoter elements. This binding activates or represses the transcriptional rates of their target genes [3, 4]. The identification and functional characterization of these transcription factors therefore assumes significance for the reconstruction of transcriptional regulatory networks.

Computational prediction techniques, in contrast to experimental ones, are faster and highly accurate compared to traditional experimental methods for high-throughput analysis of large-scale genome sequences. In this study, our efforts were directed towards development of a prediction system for NAC transcription factors in rice. A new prediction method called NACSVM was developed based on a powerful machine learning algorithm *viz.*, Support Vector Machine (SVM) for the prediction of NAC proteins in *indica* rice (*Oryza sativa* L. ssp. *indica).* The evolutionary and compositional features of NAC protein sequences were taken into consideration. A two-fold criteria *viz.* cross-validation and independent data test techniques were used to evaluate the performance of the developed model. Finally, a web-based server was also developed based on the best model obtained, where the users have the option to query their sequence/sequences for the prediction of stress-responsive NAC proteins in rice.

2. MATERIALS AND METHODS

2.1. Data Sets

Development of prediction methods has the major concern of selection of a dataset for the experimental results. The data set used in this study comprised of 95 NAC proteins of *indica* rice. Additionally, 95 NAC proteins from diverse plant families *viz.* Soybean, Arabidopsis, poplar, wheat cotton and maize which were taken from Uniprot Knowledgebase. In order to cross-check species-specific classifier's (indica rice) performance on some non-trained plants, the above 'All Plants' dataset was used. The 95 NAC proteins of indica rice, of which some where uncharacterized proteins, were established to be of NAC family through Prosite and Pfam databases. From the dataset consisting of 95 NAC proteins, 10 were randomly selected for creating the test set and remaining 85 proteins were used as positive dataset for creating the training set. Non-NAC protein sequences were used in the training as the negative data set. This was done to ensure redundancy of proteins in test set and training set and these were used for independent dataset test as well since in independent data test both training set and test set has to be entirely different. Similar procedure was applied for 'All plants' model also.

2.2. Support Vector Machine

Support vector machine (SVM) is a strong machine learning technique [5, 6]. The algorithm is conceptually simple and easy to implement. It learns by example to assign labels to objects [7]. Presently, SVMs are being employed in a variety of biological applications [8]. Being a mathematical entity, the SVM algorithm maximizes a particular mathematical function in respect of a given set of data. In the present study, we have utilized SVM^{light} [9], which is a freely downloadable package, to predict the NAC proteins. The user has the option to choose a number of parameters in this software, apart from in-built kernel functions *viz.*, linear, polynomial and radial basis function (RBF). Seven different feature extraction methods, based on various features of a protein sequence, were employed to achieve maximum accuracy and the same are highlighted below.

2.3. Composition-based classifiers

2.3.1. Simple amino-acid composition

While considering simple amino-acid composition, the fraction of each amino acid occurring in a protein sequence is manipulated and the compostion is of dimension 20. The fraction of all 20 natural amino acids was calculated using the following equation:

$$\text{Fraction of amino acid} = \frac{\text{Total number of amino acid i}}{\text{Total number of amino acids inprotein}} \qquad (1)$$

2.3.2. Traditional dipeptide composition

The traditional dipeptide composition, utilizing the sequence order effects, gives global information about each protein sequence. Using this composition, a fixed pattern length of dimension 400 (20x20) can be obtained. The traditional didpeptide compostion encompasses both the information of the amino-acid composition and local order of amino acids. The fraction of each dipeptide was calculated according to the equation:

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

2.3.3. Tripeptide composition

The tripeptide composition reflects both the total amino acid composition and also the sequence order effects [10, 11] and also provides global information about each protein sequence. This representation gives a fixed pattern length of 8000 (20x400) which encompasses the information of the amino-acid composition along with the local order of amino acids and fraction of each tripeptide was calculated using Equation 3,

$$\text{Fraction of tripep}\,(\mathrm{i} + \mathrm{n}) = \frac{\text{Total number of tripep(i+n)}}{\text{Total number of all possible tripeptides}} \qquad (3)$$

where n=2 and tripep (i + 2) is one of 8000 tripeptides.

2.4. Split amino acid composition technique

2.4.1 Four parts composition

This composition divides each protein sequence into four equal parts, mainly depending on length. For each divided part, occurrence of amino acid was calculated separately using Eq. (1) and then all the four were combined to obtain a fixed pattern length of 80 (20 x 4) which helps to collect more information about the protein sequence. This composition type has shown good results as revealed in some earlier studies [11, 12].

2.5. Hybrid techniques

For further evaluating the prediction accuracy of NACSVM various hybrid approaches by combining different features of a protein sequence was used.

2.5.1. Hybrid 1

Here, combining amino acid composition and dipeptide composition features of a protein sequence as calculated by using Equation (1) and (2), respectively, a hybrid module was developed, which generated a SVM input vector pattern of 420 consisting of 20 for amino acid and 400 for dipeptide composition.

2.5.2 Hybrid 2

In the second hybrid approach, another hybrid module was developed by combining amino acid composition and tripeptide composition (i + 2) as calculated using Eq. (1) and (3), respectively which generated a SVM input vector pattern of 8020-dimension consisting of 20 for amino acid and 8000 for tripeptide.

2.6. Position-specific scoring matrix (PSSM)

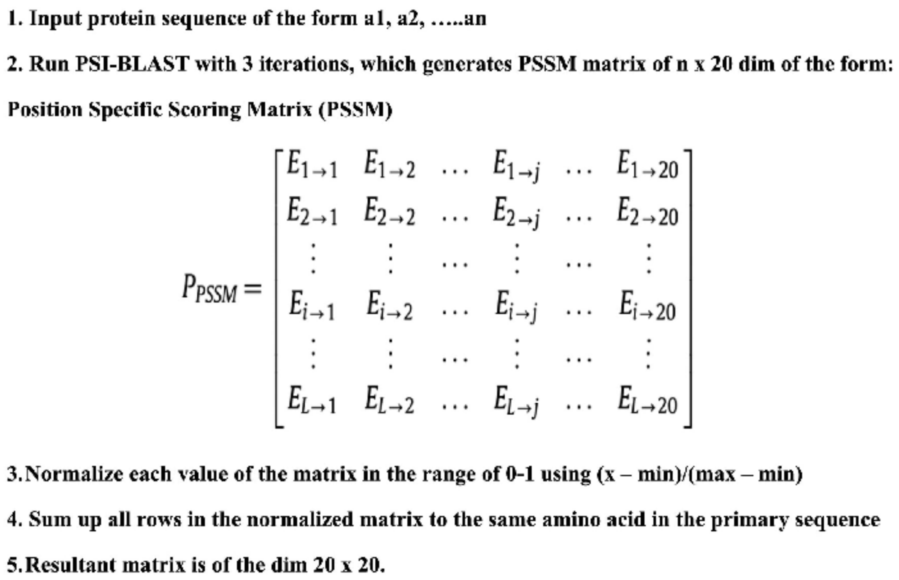
The evolutionary information of a protein sequence stored in the matrix known as PSSM matrix was used to create PSSM based module. This information is stored in a position-specific scoring table called profile, created from a group of sequences which are previously aligned by PSI-BLAST against the non-redundant (NR) database at GenBank. This method adopts PSSM matrix extracted from sequence profiles as input data to find distantly related proteins by sequence comparison [13]. The PSSM matrix gives the log-odds score for determining a particular matching amino acid in a target sequence as shown in Figure 1. A profile can be constructed from any number of known sequences by allowing more information to be used in the testing of the target sequence which makes this method to be different from other composition methods of sequence comparison. The PSSM provides a matrix of n x 20 dimension for a protein input sequence with n amino acid residues in the row and the occurrence of each type of 20 amino acids represented in 20 columns. The PSSM matrix of a protein sequence which was obtained from the profiles of PSI-BLAST was used to create a 20 x 20-dimensional input vector by calculating the sum of all rows in the PSSM matrix of the same amino acid which occurred in the primary sequence in the range of 0-1 by using the function $\frac{(X-minimum)}{(maximum-minimum)}$. Here X is considered as the individual score of each amino acid and maximum, minimum parameters in the equation are the corresponding maximum and minimum value of score in each row of the matrix (Fig 1).

2.7. Measurement of Performance of NACSVM

The effectiveness of a classifier is generally checked using either single independent dataset test, cross-validation test or jackknife test. The jackknife test is supposed to be the most rigorous and objective one amongst these three evaluation criteria's as illustrated by a comprehensive review [14]. However, since the method takes a longer time to train a predictor based on SVM, in this work, only 10-fold cross-validation and independent data set validation techniques were adopted for performance measurement of the classifier. For 10-fold cross-validation, the available dataset was divided randomly into ten equally sized sets. These ten equally divided sets were used for training and testing with each distinct set used for testing and the remaining nine sets for training with iteration being repeated ten times. In the independent dataset test it has to be specifically taken care that data used to test and those used for training are unique and the selection of data which are used for the testing can be chosen arbitrarily.

2.8. Evaluation Parameters

In this paper for evaluation criteria we have adopted the following five measures *viz.*, accuracy (Ac), sensitivity (Sn), specificity (Sp), precision (Pr) and Mathew's Correlation Coefficient (MCC). The correct ratio between both positive (+) and negative (-) data sets of NAC proteins is defined as Accuracy (Ac) given by Eq. 6. The sensitivity (Sn) and specificity (Sp) are defined as the correct prediction ratios of positive (+) sets of NAC proteins and negative data (-) sets of nonNAC proteins respectively (Eq. 4 and 5).



1. Input protein sequence of the form a1, a2, …..an

2. Run PSI-BLAST with 3 iterations, which generates PSSM matrix of n x 20 dim of the form:

Position Specific Scoring Matrix (PSSM)

$$P_{PSSM} = \begin{bmatrix} E_{1\to1} & E_{1\to2} & \dots & E_{1\to j} & \dots & E_{1\to20} \\ E_{2\to1} & E_{2\to2} & \dots & E_{2\to j} & \dots & E_{2\to20} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ E_{i\to1} & E_{i\to2} & \dots & E_{i\to j} & \dots & E_{i\to20} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ E_{L\to1} & E_{L\to2} & \dots & E_{L\to j} & \dots & E_{L\to20} \end{bmatrix}$$

3. Normalize each value of the matrix in the range of 0-1 using (x – min)/(max – min)

4. Sum up all rows in the normalized matrix to the same amino acid in the primary sequence

5. Resultant matrix is of the dim 20 x 20.

Figure 1. Flow chart of the algorithm to generate PSSM matrix of 20x20 dimensions from the input pattern of nx20 dimension matrix used as input pattern for various models of SVM

The proportion of the predicted positive cases that were obtained correct is defined as Precision (Eq.7). When the numbers of positive and negative data differ too much from each other, the Mathew correlation coefficient (MCC) becomes decisive to evaluate the prediction performance of the developed tool (Eq. 8). The MCC which is considered to be the most robust parameter of any class prediction method is a measure used in machine learning for judging the quality of binary (two-class) as well as multi-labeled classifications. The value of MCC ranges from -1 to 1, and a positive MCC value stands for better prediction performance. The data with positive hits by NACSVM which are the real positives were defined as true positives (TP), while the others were defined as false positives (FP).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \text{ x } 100 \quad (4)$$

$$\text{Specificity} = \frac{TN}{FP+TN} \text{ x } 100 \quad (5)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{ x } 100 \quad (6)$$

$$\text{Precision} = \frac{TP}{TP+FP} \text{ x } 100 \quad (7)$$

$$\text{MCC} = \frac{(TP\ X\ TN) - (FP\ X\ FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN}} \quad (8)$$

where TP and TN are truly or correctly predicted positive NAC protein and negative (non- NAC protein), respectively. FP and FN are falsely or wrongly predicted NAC and non-NAC proteins, respectively.

2.9. Sequence Similarity Search

PSI-BLAST is a method to explore the similarities between a protein query sequence and all the sequences in a protein database. In contrast to BLAST, this search method uses the position specific scoring matrix (PSSM) to score matches between query and database sequences. The advantage of using a profile to search a database is that it often detects close relationships between proteins that are structurally or functionally distant. Here we have used PSI-BLAST in place of normal BLAST because of former’s ability to detect remote homology of NAC proteins against Swiss Prot database and result was analysed.

2.10. ROC Curves

The performance of a binary classifier can be explained with ROC curve which is a graphical plot drawn by altering threshold values. The analysis of ROC curve helps to characterize the prediction for individual locations in the curve [15, 16]. It is a plot of sensitivity and specificity (or false positive rate = 1 - specificity) as x and y axes, respectively that shows the trade off between sensitivity and specificity. This curve is created by plotting the fraction of true positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold levels. The area under the curve (AUC) which is depicted in the ROC curve further validates the classifier accuracy.

3. RESULTS AND DISCUSSION

In the present study we have carried out 10-fold cross-validation and independent data test validation to evaluate the performance of NACSVM (Table 1 and Table 2) and to assess the accuracy of the classifier. From the tables, it is evident that cross validation tests have better results for different composition methods compared to independent data test. From these tables, it is also clear that PSSM method has 100% accuracy with respect to all the three kernels of SVM compared to all the other six composition methods. Based on these results, PSSM was selected to be the best model for NACSVM.

3.1. Statistical Tests of Various Classifiers

In the composition-based modules, three different classification methods were adopted for both cross validation and independent data tests. The amino-acid composition based module with polynomial kernel obtained a maximum accuracy of 96 % with MCC of 0.93. A module based on traditional dipeptide composition (i + 1) was generated which gave more information about frequency and local order of residues in the sequence. This module could attain a maximum accuracy of 83% with MCC of 0.72. Tripeptide composition based module developed could obtain more comprehensive information on the sequence order effects and this could achieve an accuracy of 95% with MCC of 0.9. The four parts composition method obtained maximum accuracy of 97% with MCC of 0.93. The entire validation test resulted on the base of 10-fold cross-validation. The detailed performance of amino acid, traditional dipeptide, tripeptide and four parts composition based modules with cross validation are given in Table 2.

The detailed performance results of amino acid, traditional dipeptide, tripeptide and four parts composition based modules with independent data test validation are reported in Table 1. It could be observed on analyzing independent test validation for composition based modules that amino-acid composition achieved an accuracy of 79% with MCC of 0.64, dipeptide composition achieved an accuracy of 93% with MCC of 0.86, tripeptide (i+ 2) composition based module achieved an accuracy of 93% with MCC of 0.87 and four parts composition method achieved maximum accuracy of 79% with MCC of 0.63 for all the three kernels. From Tables 1 and 2, it is clearly evident that compared to independent data test, best performance results are achieved for 10-fold cross validation for the various composition based modules.

Hybrid methods, which were combinations of various features of a protein sequence, were employed in addition to the above composition methods. Hybrid 1, combination of amino acid feature with dipeptide feature achieved an accuracy of 83% with MCC of 0.67 (all three kernels) for independent data test and 95% accuracy for RBF kernel with MCC of 0.91 in the case of 10-fold cross validation test. Hybrid 2, combination of amino acid and tripeptide composition also obtained a little less accuracy of 79% and MCC value of 0.64 for independent data test and 94% for RBF kernel with an MCC value of 0.89 for cross validation data test. Hybrid results shows that cross validation test results has an upper hand over independent test results (Table 1 and 2). The performance comparison of composition, evolutionary and hybrid based methods for cross validation are reported in the figure 2.

In the evolutionary information based PSSM module, a PSSM was constructed for each protein sequence from the generated sequence profiles by PSI-BLAST. The PSSM based module achieved high accuracy of 100% for all the three kernels of SVM with cross validation and independent data test. The results of PSSM-based module with three algorithms are reported in Tables 1 and 2. It is clear from the results that the prediction performance can be significantly improved on application of PSSM-based module, which offers important evolutionary information about proteins than any other composition methods. This shows that PSSM-based classifiers are statistically better compared to the other modules with the best overall accuracy achieved by PSSM- based SVM module alone and which was also statistically significant compared to other modules.

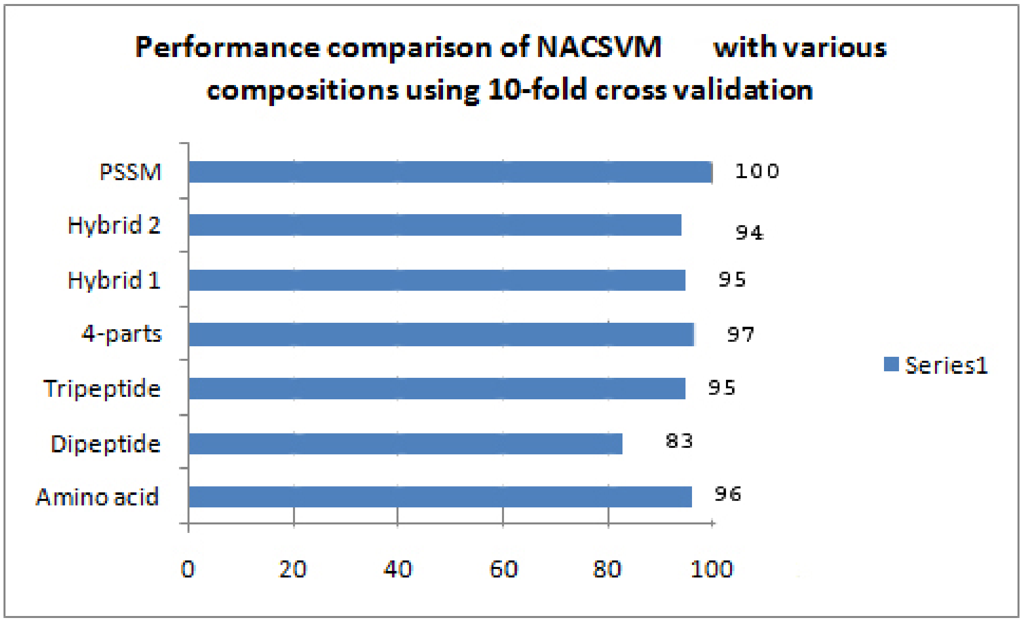


Figure 2. Comparison of performance validation of NACSVM with different composition methods

3.2. ROC Curves

The receiver operating characteristic or simply ROC curve is defined as a measure which represents the connection between sensitivity and (1-specificity) for a class. In this paper we have plotted the ROC curves of the various compositions based on the independent test performance. The ROC curve of PSSM composition module represents a perfect classifier since the curve represents an inverted ‘L’ which is a desirable characteristic of an ROC curve (Figure 3). With respect to different compositions the figure also depicts the comparison of the performance of various classifiers. Based on different threshold scores each point on the ROC curve was plotted. The area under the curve (AUC) of value one for PSSM composition and high confidence AUCs for all other compositions from the figure also depicted “excellent classification” [17]. The AUC shows the probability that when one positive and negative sample are drawn at random then in such cases the decision function assigns a higher value to the positive sample than to the negative.

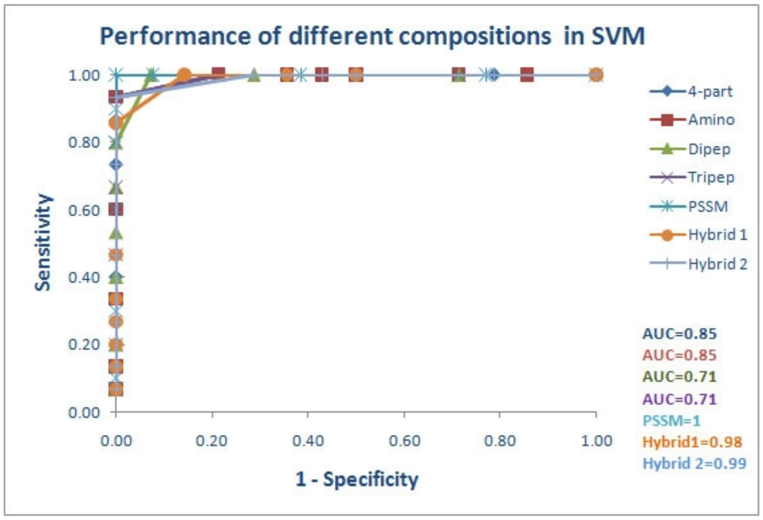


Figure 3. ROC curve for different compositions in SVM using independent test results

3.3 Sequence Similarity Search

The homology of a protein with other connected sequences provides a broad range of information about each functional encoded protein thus summarizing the evolutionary information about the proteins and this is carried out through sequence similarity search PSI-BLAST. To produce the homology of the given sequence with other related sequences in the database, a protein sequence was compared with a created database which provided a broad range of information about each functional encoded protein [18]. A 10-fold cross-validation was conducted with sequence similarity achieving no significant hits and an accuracy of only 53.5% was obtained (Table 3). This result suggests that similarity based search tools are not efficient and consistent as compared to different composition based modules based on computational methods.

Table 1**.** Validation of independent data test results of NAC proteins with SVM

Composition	Algorithm	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Amino acid	Linear	88	90	89	91	0.79
	Polynomial	98	95	96	96	0.93
	RBF	94	92	93	93	0.87
Dipeptide	Linear	99	63	81	77	0.64
	Polynomial	76	96	83	95	0.72
	RBF	79	84	83	96	0.63
Tripeptide	Linear	100	0	53	53	0.00
	Polynomial	100	0	53	53	0.00
	RBF	95	95	95	96	0.90
4- parts	Linear	90	100	95	100	0.90
	Polynomial	96	95	97	99	0.93
	RBF	91	94	94	97	0.87
Hybrid 1	Linear	93	94	93	94	0.88
	Polynomial	93	94	93	95	0.87
	RBF	94	96	95	96	0.91
Hybrid 2	Linear	93	94	93	94	0.87
	Polynomial	93	94	93	95	0.87
	RBF	94	94	94	95	0.89
PSSM	Linear	100	100	100	100	1
	Polynomial	100	100	100	100	1
	RBF	100	100	100	100	1

Table 2. Comparison of the prediction performance of three kernels of SVM with different composition techniques using 10-fold cross validation

Composition	Algorithm	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Amino acid	Linear	100	57	79	71	0.64
	Polynomial	100	57	79	71	0.64
	RBF	100	57	79	71	0.64
Dipeptide	Linear	93	93	93	93	0.86
	Polynomial	93	93	93	93	0.86
	RBF	93	93	93	93	0.86
Tripeptide	Linear	87	100	93	100	0.87
	Polynomial	87	100	93	100	0.87
	RBF	87	100	93	100	0.87
4- parts	Linear	100	57	79	70	0.63
	Polynomial	100	57	79	70	0.63
	RBF	100	57	79	70	0.63
Hybrid 1	Linear	93	71	83	78	0.67
	Polynomial	93	71	83	78	0.67
	RBF	93	71	83	78	0.67
Hybrid 2	Linear	100	57	79	71	0.64
	Polynomial	100	57	79	71	0.64
	RBF	100	57	79	71	0.64
PSSM	Linear	100	100	100	100	1
	Polynomial	100	100	100	100	1
	RBF	100	100	100	100	1

Table 3: Prediction result of NAC proteins with similarity search (10-fold cross validation)

Test	No. of sequences given	Correctly predicted	Accuracy (%)
1	20	10	50
2	20	10	50
3	20	10	50
4	20	8	40
5	20	10	50
6	20	11	55
7	20	10	50
8	20	10	50
9	15	11	73.3
10	15	10	66.7
Average			53.5

3.4 Comparison of NACSVM with 'All Plant' method

A species-specific predictor(s) is much more advantageous than 'All Plant' method and to prove this fact we trained a corresponding method using the same encoding method as used in NACSVM on a dataset derived from all the plant species. The performance of two methods on the rice independent dataset was later compared. A dataset consisting of six plants, namely Arabidopsis, soybean, wheat, poplar, maize and cotton were downloaded from Uniprot knowledgebase consisting of 95 sequences in total were used to train the 'All plant' method. To obtain fair result it was ensured that 'All Plant' training dataset were independent of rice sequences as we wanted to compare the performance of rice independent dataset on both the NACSVM and 'All Plant' method.

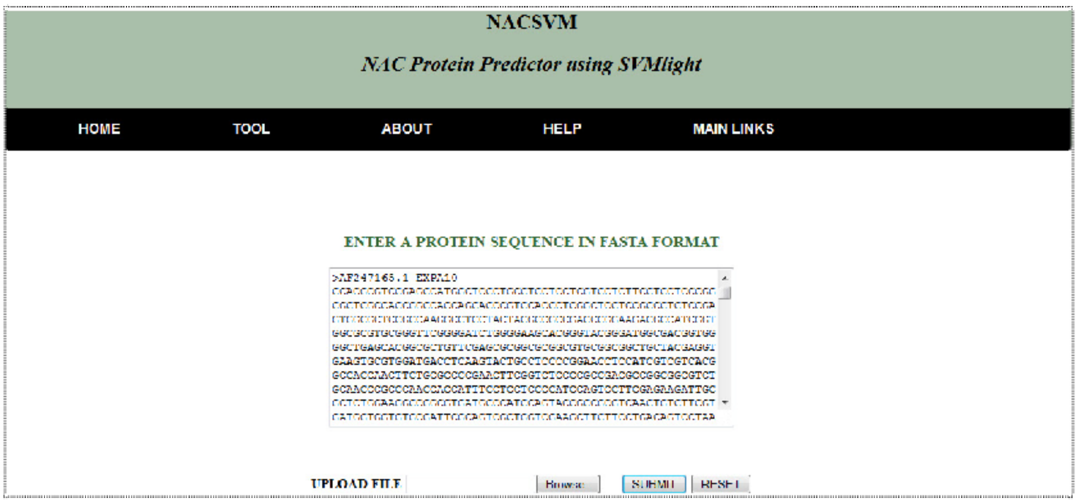
For model generation, the traditional amino acid composition based classifier was used for 'All Plant' dataset using the independent dataset approach and rice independent dataset was tested on the model files generated from the 'All Plant' classifier. Statistical parameters for "All plant" method were calculated similar to that in NACSVM and these were compared with the amino acid based model of NACSVM .The best classifier obtained in NACSVM was from PSSM matrix using evolutionary information of a protein sequence, but for comparison purpose we have used simpler composition method for 'All Plant' module. Results tabulated in Table 5 shows that species specific tools are much superior to 'All Plant' tool.

Table 4. Comparison of performance of NAC proteins with All-Plant tool and NACSVM

Method	Algorithm	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
All-Plant	Linear	100	50	76	68	0.58
	Polynomial	93	57	76	70	0.55
	RBF	100	50	76	68	0.58
NACSVM	Linear	88	90	89	91	0.79
	Polynomial	98	95	96	96	0.93
	RBF	94	92	93	93	0.87

3.8. Description of web server

‘NACSVM’ is a dynamic web server implemented on the World Wide Web using the best performing module. The tool was developed in Perl language and PHP, HTML were used to create web interface to assess user queries. It is a user friendly web server which allows users to submit/paste their protein sequences through the standard FASTA format or allows uploading of sequence through a file (Figure 4, Figure 5). The result of the prediction will be displayed in a user friendly format on the screen with some needed information (Figure 6). NACSVM uses PSSM-based module for prediction which achieved the overall best accuracy compared to other models. The overall architecture of the NACSVM web server is depicted in the Figure 5.



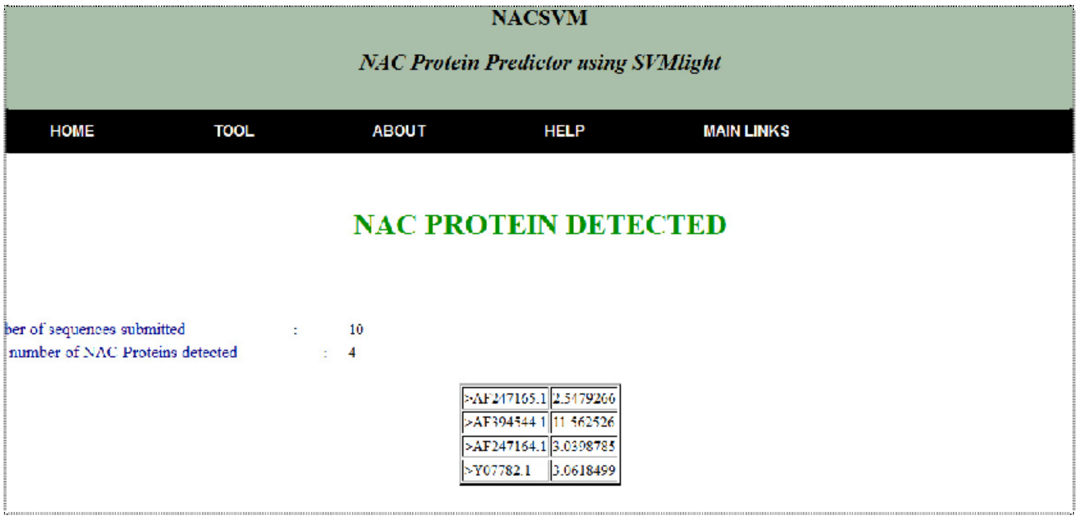
NACSVM

NAC Protein Predictor using SVMlight

ENTER A PROTEIN SEQUENCE IN FASTA FORMAT

```
>AF247165.1 EXPA10
CCACCCGTCCGACCCATGCCTCCCTGCCTCCTCCTCCTCCTCTTCCTCCTCCCCC
CGCTCCCCACCCCCCACCAGCACCCCGTCCACCCTCGCCTCCTCCCCCCTCTCCCA
[illegible]
GGCGCGTGCGGGTTCGGGGATCTGGGGAAGCACGGGTACGGGATGGCGACGGTGG
GGCTGAGCACGGCGCTGTTCGAGCGCGGCGCGGCGTGCGGCGGCTGCTACGAGGT
GAAGTGCGTGGATGACCTCAAGTACTGCCTCCCCGGAACCTCCATCGTCGTCACG
GCCACCAACTTCTGCGCCCCGAACTTCGGTCTCCCCGCCGACGCCGGCGGCGTCT
GCAACCCGCCCAACCACCATTTCCTCCTCCCCATCCAGTCCTTCGAGAAGATTGC
CCTCTCGAACGCCCCCGGTCATCCCCATCCACTACCCGCCCCTCAACTCTCTTCGT
CATCCTGCTCTCCCATTCCCACTCGCTCGTCCAACCTTCTTCCTGCACACTCCTAA
```

Figure 5. An overview of the submission form of NACSVM



NACSVM

NAC Protein Predictor using SVMlight

NAC PROTEIN DETECTED

ber of sequences submitted : 10
number of NAC Proteins detected : 4

>AF247165.1	2.5179266
>AF394544 1	11 562526
>AF247164.1	3.0398785
>Y07782.1	3.0618499

Figure 5. An overview of the output page displaying the scores of the predicted proteins.

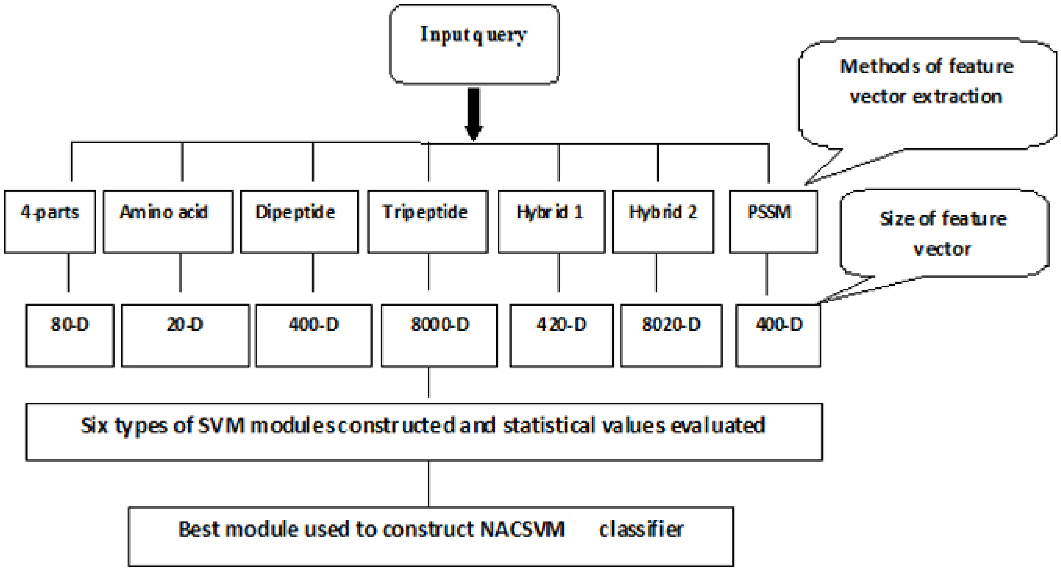

Input query
Methods of feature vector extraction
4-parts
Amino acid
Dipeptide
Tripeptide
Hybrid 1
Hybrid 2
PSSM
Size of feature vector
80-D
20-D
400-D
8000-D
420-D
8020-D
400-D
Six types of SVM modules constructed and statistical values evaluated
Best module used to construct NACSVM classifier

Figure 6. Schematic diagram showing the flow of steps in developing the web based tool

3. CONCLUSIONS

One of the most principal goals of genome sequencing projects is finding new protein-coding genes. Computational tools, in contrast to experimental techniques, provide faster and accurate access to predictions for any organism. There is a lack of accurate gene prediction programs with respect to various functionalities in rice and the availability of systems/tools that can predict location from sequence is essential to the full characterization of expressed proteins.

Identification of NAC proteins from sequence databases is difficult and less accurate due to poor sequence similarity. In this paper, we have presented a new method for NAC prediction based on SVM, a simple machine learning technique and for which performance was found to be highly satisfactory. Model was tested with different kernels of SVM to validate the accuracy. Validation tests shows high prediction accuracies which proves that NACSVM is a potentially useful tool for the prediction of NAC proteins.

REFERENCES

[1] IRRI, (2006) “Bringing Hope, Improving Lives : Strategic Plan 2007 – 2015”, Los Banos : IRR.

[2] Y. Fang, J. You, K. Xie, W. Xie & L. Xiong, (2008) “Systematic sequence analysis and identification of tissue-specific or stress-responsive genes of NAC transcription factor family in rice”, Molecular Genetics and Genomics, Vol. 280, No. 6, pp 547-563.

[3] J. L. Riechmann, J. Heard, G. Martin, L. Reuber, C. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J. Z. Zhang, D. Ghandehari, B. K. Sherman & G. Yu, (2000) “Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes”, Science , Vol. 290, No. 5499, pp 2105-2110.

[4] G.A. Wray, M.W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman & L.A. Romano, (2003) “The evolution of transcriptional regulation in eukaryotes”, Molecular Biology and Evolution, Vol. 20, No. 9, pp 1377-1419.

[5] C. Cortes & V. Vapnik, (1995) “Support vector networks”, Machine Learning, Vol. 20, No. 5, pp 273–297.

[6] V. Vapnik, (1995) The Nature of Statistical Learning Theory, Springer, New York.

[7] B. E. Boser, I. M. Guyon, & V. N. Vapnik, (1992) “A training algorithm for optimal margin classifiers”, in 5th Annual ACM Workshop on COLT, D. Haussler, Ed. Pittsburgh, PA: ACM Press, pp 144–152.

[8] W.S. Noble, (2004) “Support vector machine applications in computational biology”, in Kernel Methods in Computational Biology, B. Sch olkopf, K. Tsuda & J. P. Vert, Eds. Cambridge, MA: MIT Press, pp 71–92 .

[9] T. Joachims, (1999) “Making large-scale SVM learning practical”, in Advances in Kernel Methods : Support Vector Learning, B. Sch olkopf, C. Burges & A. Smola, Eds. Cambridge, MA: MIT Press, pp 41–56.

[10] A. Garg, M. Bhasin & G. P. S. Raghava, (2005) “Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search”, Journal of biological Chemistry, Vol. 280, pp 14427–14432.

[11] M. Wang, A. Li, D. Xie , Z. Fan & H. Feng, (2005) “Improving prediction of protein sub-cellular localization using evolutionary information and sequence-order information”, 27th Annual International Conference of the IEEE-EMBS, pp 4434– 4436.

[12] D. Xie, A. Li, M. Wang, Z. Fan & H. Feng, (2005) “LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST”, Nucleic Acids Research, Vol.33, pp 105–110.

[13] D. T. Jones, (1999) “Protein secondary structure prediction based on position-specific matrices”, Journal of Molecular Biology, Vol. 292, pp 195-202.

[14] K. C. Chou & C. T. Zhang, (1995) “Prediction of protein structural classes”, Critical Reviews in Biochemistry and Molecular Biology, Vol. 30, pp 275–349.

[15] J. A. Swets, (1988) “Measuring the accuracy of diagnostic systems”, Science, Vol. 240, pp 1285–1293.

[16] M. H. Zweig & G. Campbell, (1993) “Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine”, Clinical Chemistry, Vol. 39, pp 561–577.

[17] D.W. Hosmer & S. Lemeshow, (2000) “Applied Logistic Regression”, Ed. 2, John Wiley and Sons, New York, pp 156–164.

[18] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller & D.J. Lipman, (1997) “Gapped Blast and PSI-Blast: a new generation of protein database search programs”, Nucleic Acids Research, Vol. 25, pp 3389–3402.

AUTHORS

N. Hemalatha is doing part time research in the Department of Information Technology, School of Information Science and Technology, Kannu r University, India. She is presently working as Assistant Professor in the Computer Science Department of St. Aloysius College, Mangalore, India. Her primary area of interests and expertise are in the areas of machine learning and Bioinformatics.

M. K. Rajesh, PhD, is a Senior Scientist in the Biotechnology Section at Central Plantation Crop Research Institute, Kasaragod, India. His primary areas of scientific expertise include plant tissue culture, plant molecular biology and bioinformatics.

N. K. Narayanan, PhD, is a Senior Professor in the Department of Information Technology, School of Information Science and Technology, Kannur University, India. He earned a PhD in the area of signal processing from Cochin University of science & Technology in 1990. His current research interests include bioinformatics, image processing, pattern recognition, neural networks, and speech signal processing.