Nacpred: Computational Prediction of Nac Proteins in Rice Implemented Using Smo Algorithm

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

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

Abstract. The impact of abiotic stresses, such as drought, on plant growth and development severely hampers crop production worldwide. The development of stress-tolerant crops will greatly benefit agricultural systems in areas prone to abiotic stresses. Recent advances in molecular and genomic technologies have resulted in a greater understanding of the mechanisms underlying the genetic control of the abiotic stress response in plants. NAC (NAM, ATAF1/2 and CUC2) domain proteins are plant-specific transcriptional factors which has diversified roles in various plant developmental processes and stress responses. More than 100 NAC genes have been identified in rice. In the proposed method, NACPred, an attempt has been made in the direction of computational prediction of NAC proteins. The well-known sequential minimum optimization (SMO) algorithm, which is most commonly used algorithm for numerical solutions of the support vector learning problems, has been used for the development of various modules in this tool. Modules were first developed using amino acid, traditional dipeptide (i+1), tripeptide (i+2) and an overall accuracy of 76%, 90%, and 97% respectively was achieved. To gain further insight, a hybrid module (hybrid1 and hybrid2) was also developed based on amino acid composition and dipeptide composition, which achieved an overall accuracy of 90% and 97%. To evaluate the prediction performance of NACPred, cross validation, leave one out validation and independent data test validation were carried out. It was also compared with algorithms namely RBF and Random Forest. The different statistical analyses worked out revealed that the proposed algorithm is useful for rice genome annotation, specifically predicting NAC proteins.

Keywords: Rice, gene prediction, NAC, SMO.

1 Introduction

Rice (*Oryza sativa* L.), a source of staple food, has a major influence on human nutrition and food security. Billions of people world-wide depend on rice-based production systems for their main source of employment and development. Rice production

S. Unnikrishnan, S. Surve, and D. Bhoir (Eds.): ICAC3 2013, CCIS 361, pp. 266–275, 2013.

continuously faces the challenge of keeping pace with rapid rise in human population and declining natural resource base, two of the critical resources being land and water. In addition, abiotic stresses, such as drought, adversely affect the growth and productivity of rice-based farming systems.

The development of stress-tolerant crops will be of immense advantage in modern agriculture, especially in areas that are prone to such stresses. In recent years, several advances have been made towards identifying potential stress related genes which are capable of increasing the tolerance of plants to abiotic stress. NAC transcription factors have major functions in plant development as well as in abiotic stress responses. NAC (NAM, ATAF1/2 and CUC2) domain proteins comprise of one of the largest plant-specific transcriptional factors which is represented by approximately 140 genes in rice [1]. These transcription factors (TFs) regulate gene expression by binding to specific cis-acting promoter elements, thereby activating or repressing the transcriptional rates of their target genes [2, 3]. Thus, for the reconstruction of transcriptional regulatory networks, the identification and functional characterization of these transcription factors is essential [4].

Computational prediction methods, compared with the experimental methods are fast, automatic and more accurate especially for high-throughput analysis of large-scale genome sequences. Therefore, a fully automatic prediction system for NAC transcription factors in rice is a systematic attempt in this direction. The SMO module for the prediction of NAC proteins in genome of indica rice (*Oryza sativa* L. ssp. indica) was developed using various features of a protein sequence and the performance of these models was evaluated using cross-validation techniques.

2 Materials and Methods

2.1 Datasets

The selection of dataset is the most important concern during development of a prediction method. The data set used in the present study, consisted of 95 NAC proteins of indica rice taken from Uniprot Knowledgebase. These 95 proteins were screened strictly in order to develop a high quality data set for the prediction tool. Fifteen NAC proteins were randomly selected from the main dataset for the creation of test set and remaining 80 proteins were used for positive dataset / training set. Non-NAC protein sequences were used as the negative data set. For training and testing, independent datasets were used which means training set and test set were entirely different.

2.2 Performance Evaluation and Parameters

Three methods often used for examining the effectiveness of a predictor, in statistical prediction are single independent dataset test, cross-validation test and jackknife test. Out of these, the jackknife test is considered to be most rigorous and objective one, as illustrated by a comprehensive review [5]. However, since the size of the dataset in the present study was large and jackknife test method takes much longer time to train a predictor based on SMO, cross-validation (5-fold, 8-fold) and independent dataset

test were adopted for performance measurement. In n-fold cross validation, all the positive and negative datasets were combined and then divided equally into n parts, keeping the same distribution of positive and negative datasets in each part. Then n-1 parts were merged into a training data set with the one part left out taken as a test data set and the average accuracy of n-fold cross validation was used to estimate the performance. In the independent dataset test, although none of the data to be tested occurs in the training dataset used to train the predictor, the selection of data for the testing dataset could be quite arbitrary. In "leave-one-out" cross-validation (LOO), each sample in the dataset is separated out in turn as an independent test sample, and all the remaining samples are used as training data. This process is repeated until every sample is used as test sample one time with no repetition. All models were implemented in the WEKA software package [6].

2.3 The Machine Learning Algorithms

Sequential Minimal Optimization(SMO) is a support vector machine learning algorithm (SVM) that is conceptually simple, easy to implement, generally faster, and has better scaling properties for difficult SVM problems than the standard SVM training algorithm [7]. Training a support vector machine requires the solution of a very large quadratic programming (QP) optimization problem which is quite time consuming. In the case of SMO, it breaks this large QP problem into a series of smallest possible QP problems which are solved analytically and avoids using a time-consuming numerical QP optimization as an inner loop. The amount of memory required for SMO is linear in the training set size allowing SMO to handle very large training sets. Because matrix computation is avoided, SMO scales somewhere between linear and quadratic in the training set size for various test problems, where as SVM algorithm scales somewhere between linear and cubic in the training set size. SMO's computation time is dominated by SVM evaluation which makes SMO faster for linear SVMs and sparse data sets.

The Radial Basis Function (RBF) network is as a variant of artificial neural network [8]. An RBF is embedded in three layers, *viz*., the input layer, the hidden layer, and the output layer. The input layer broadcasts the coordinates of the input vector to each of the nodes in the hidden layer and contains one neuron in the input layer for each predictor variable. Each node in the hidden layer then produces an activation based on the associated radial basis function and this layer has a variable number of neurons based on the training process. Finally in the output layer each node computes a linear combination of the activations of the hidden nodes. The result of an RBF network to a given input stimulus is completely determined by the activation functions associated with the hidden nodes and the weights associated with the links between the hidden layer and the output layer.

Random Forests (RF) grows many classification (decision) trees. To classify a new object from an input vector, RF puts the input vector down into each of the trees in the forest. Each tree gives a classification, and it is said that the tree "votes" for that class. The forest chooses that classification which has the most votes (over all the trees in the forest).The forest error rate depends on the correlation between any two

trees in the forest and the strength of each individual tree in the forest. Decreasing forest error rate increases the strength of the individual trees.

2.4 Features and Modules

Amino-acid composition: Amino-acid composition is the fraction of each amino acid occurring in a protein sequence. This representation completely misses the order of amino acids. To calculate the fraction of all 20 natural amino acids following equation was used:

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

Traditional dipeptide composition: Traditional dipeptide composition gives information about each protein sequence giving fixed pattern length of 400 (20x20). This composition 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:

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

In addition, to observe the interaction of the ith residue with the 3rd residue in the sequence, tripeptide (i + 2) was generated using Equation 3,

$$\text{Fraction of tripep}(i + 2) = \frac{\text{Total number of (i+2) tripep}}{\text{Total number of all possible tripeptides}} \quad (3)$$

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

Hybrid SMO module(s): The prediction accuracy was further enhanced with various hybrid approaches by combining different features of a protein sequence.

Hybrid 1: In this approach, 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 a WEKA input vector pattern of 420 (20 for amino acid and 400 for dipeptide composition).

Hybrid 2: In the second approach we developed another hybrid module by combining amino acid composition and tripeptide composition as calculated using Eq. (1) and (3), respectively. The WEKA input vector pattern thus formed was 8020-dimensional [20 for amino acid and 8000 for tripeptide)].

2.5 Sequence Similarity Search

In this study, a query sequence was searched against the existing non-redundant database of NAC proteins (95 sequences used in training set) using PSI-BLAST

(Position-Specific Iterative Basic Local Alignment Search Tool). Here PSI-BLAST was used instead of normal BLAST because it has the capacity to detect remote homologies. Position-Specific Iterated BLAST (PSI-BLAST), a variant of BLAST is used for the discovery of weak but relevant protein sequence matches. This carries out an iterative search in which sequences found in one round was used to build score model for next round. Thus, PSI-BLAST provides a method of detecting distant relationships between proteins.

2.6 Evaluation Parameters

We adopted five frequently considered measurements for evaluation, *viz.*, accuracy (Ac), sensitivity (Sn), specificity (Sp), precision (Pr) and Mathew's Correlation Coefficient (MCC). Accuracy (Ac) defines the correct ratio between both positive (+) and negative (-) data sets. The sensitivity (Sn) and specificity (Sp) represent the correct prediction ratios of positive (+) and negative data (-) sets of NAC proteins respectively. Precision is the proportion of the predicted positive cases that were correct. However, when the number of positive data and negative data differ too much from each other, MCC should be included to evaluate the prediction performance of the developed tool. MCC is considered to be the most robust parameter of any class prediction method. The value of MCC ranges from -1 to 1, and a positive MCC value stands for better prediction performance. Among the data with positive hits by NACPred, the real positives are defined as true positives (TP), while the others are defined as false positives (FP).

$$\text{Sensitivity} = \frac{TP}{FN+TP} x\ 100 \tag{4}$$

$$\text{Specificity} = \frac{TN}{FP+TN} x\ 100 \tag{5}$$

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

$$\text{Precision} = \frac{TP}{TP+FP}\ x\ 100 \tag{7}$$

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

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.

2.7 ROC Curves

To compare the performance of different algorithms and performance of different composition methods on best algorithm in detail, ROC curves were used for intuitively

visualizing prediction performance. ROC curves plots the true positive rate (TPR) as function of the false positive rate (FPR) which is equal to 1-specificity. The area under the ROC curve is the average sensitivity over all possible specificity values which can be used as a measure of prediction performance at different thresholds. ROC curves of random predictors will be around the diagonal line from bottom left to top right of the graph with scores of about 0.5, while a perfect predictor will produce a curve along the left and top boundary of the square and will receive a score of one.

3 Results and Discussion

The prediction accuracy was assessed by two different validation techniques namely cross-validation and independent data set tests. In order to achieve maximum accuracy, five different feature extraction techniques, including three composition-based and two hybrid-based, were used and models were developed with three different algorithms namely SMO, RBF and RF. Performance accuracy of SMO algorithm was found to be the best compared to other algorithms. A graphical representation of the accuracy values of the different feature extraction methods using SMO is shown in Figure 1.

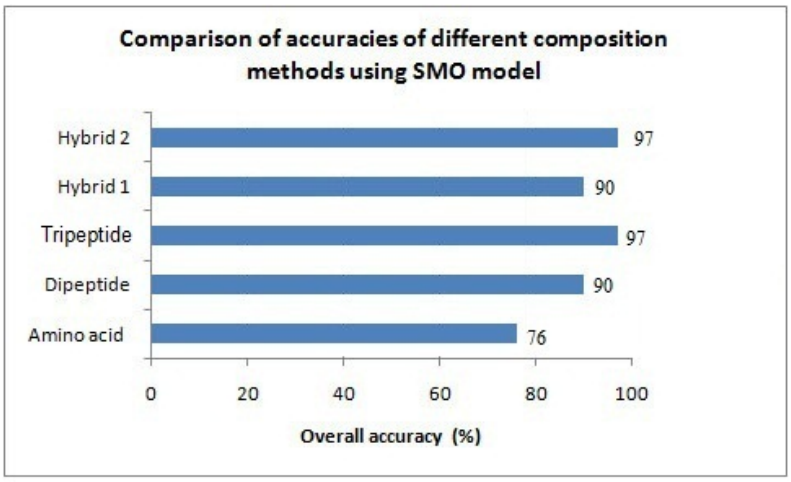

Fig. 1. Comparison of overall accuracy of various SMO modules constructed using five composition methods

3.1 Composition-Based Modules

The amino-acid composition-based module, with RBF algorithm, achieved an accuracy of 76% with different validation techniques applied in this study. The module

implemented based on traditional dipeptide composition (i+1) gave more information about frequency and local order of residues. This module could achieve a maximum accuracy of 90% with RBF algorithm. Tripeptide (i+2) composition-based module was also developed to obtain more comprehensive information on the sequence order effects. This could achieve an accuracy of 97% with sequential minimum optimization algorithm (SMO) with various validation techniques. It could be observed that traditional dipeptide composition-based modules achieved higher accuracy compared to other independent compositions (Tables 1 and 2). This may be because dipeptide composition uses the actual order of sequence while calculating the composition where as the tripeptide is based on the pseudo sequence order. The detailed performance of amino acid, traditional dipeptide and tripeptide based modules with different validation techniques are presented in Tables 1 and 2.

3.2 Sequence Similarity Search

PSI-BLAST was used to compare a protein sequence with a created database to generate the homology of the given sequence with other related sequences [9]. This provided a broad range of information about each functional encoded protein. A 10-fold cross-validation was conducted with no significant hits and an accuracy of only 50% was obtained. This result suggests that similarity-based search tools alone cannot be efficient and reliable as compared to different composition-based modules.

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

Approaches	Algorithm	Sn(%)	Sp(%)	Acc(%)	Pr(%)	MCC
Aminoacid	RBF	100	79	90	83	0.81
	SMO	93	57	76	70	0.55
	RF	93	64	79	74	0.61
Dipeptide	RBF	93	93	93	93	0.86
	SMO	93	86	90	88	0.79
	RF	87	93	90	93	0.80
Tripeptide	RBF	33	86	59	71	0.22
	SMO	100	93	97	94	0.93
	RF	93	57	76	70	0.55
Hybrid 1	RBF	100	86	93	88	0.87
	SMO	93	86	90	88	0.79
	RF	93	64	79	74	0.61
Hybrid 2	RBF	33	86	59	71	0.22
	SMO	100	93	97	94	0.93
	RF	93	57	76	70	0.55

Table 2. Comparison of the prediction performance of three machine learning algorithms with different compositions

Approach	Algorithm	5-fold cross validation					8-fold cross validation					Leave one out cross validation				
		Sn	Sp	Ac	Pr	MCC	Sn	Sp	Ac	Pr	MCC	Sn	Sp	Ac	Pr	MCC
Amino acid	RBF	100	79	90	83	0.81	100	79	90	83	0.81	100	79	90	83	0.81
	SMO	93	57	76	70	0.55	93	57	76	70	0.55	93	57	76	70	0.55
	RF	93	64	79	74	0.61	93	64	79	74	0.61	93	64	79	74	0.61
Dipeptide	RBF	93	93	93	93	0.86	93	93	93	93	0.86	93	93	93	93	0.86
	SMO	93	86	90	88	0.79	93	86	90	88	0.79	93	86	90	88	0.79
	RF	87	93	90	93	0.80	87	93	90	93	0.80	87	93	90	93	0.80
Tripeptide	RBF	33	86	59	71	0.22	33	86	59	71	0.22	33	86	59	71	0.22
	SMO	100	93	97	94	0.93	100	93	97	94	0.93	100	93	97	94	0.93
	RF	93	57	76	70	0.55	93	57	76	70	0.55	93	57	76	70	0.55
Hybrid1	RBF	100	86	93	88	0.87	100	86	93	88	0.87	100	86	93	88	0.87
	SMO	93	86	90	88	0.79	93	86	90	88	0.79	93	86	90	88	0.79
	RF	93	64	79	74	0.61	93	64	79	74	0.61	93	64	79	74	0.61
Hybrid 2	RBF	33	86	59	71	0.22	33	86	59	71	0.22	33	86	59	71	0.22
	SMO	100	93	97	94	0.93	100	93	97	94	0.93	100	93	97	94	0.93
	RF	93	57	76	70	0.55	93	57	76	70	0.55	93	57	76	70	0.55

3.3 Hybrid Approach

In addition to the different composition methods, hybrid methodologies were also developed and used by combining various features of a protein sequence. Firstly, hybrid 1 was developed by combining amino acid composition and dipeptide composition. This obtained an accuracy of 90% with SMO algorithm. Secondly, hybrid 2 was developed by combining amino acid and tripeptide composition which also had a higher accuracy of 97% with SMO algorithm. Comparison of both of these hybrid approaches revealed that hybrid 2 composition method achieves an accuracy rate equivalent to tripeptide (i+2) composition method (Fig. 1).

3.4 ROC Curves

A ROC curve is a measure which shows the relationship between sensitivity and specificity of a given class. To evaluate the best classifier obtained, we plotted ROC curves based on the results of independent data test and cross validation (results obtained were similar). Figure 2 shows the ROC curve for SMO algorithm for five different compositional methods and it can be observed from the figure that all the curves result in a straight horizontal line. This is a desirable property of ROC curves and such models have high probability of correct prediction, with a minor chance of negative prediction. This is also reflected by area under the curve values of all compositions of SMO models. Figure 3 shows the best results of each algorithm.

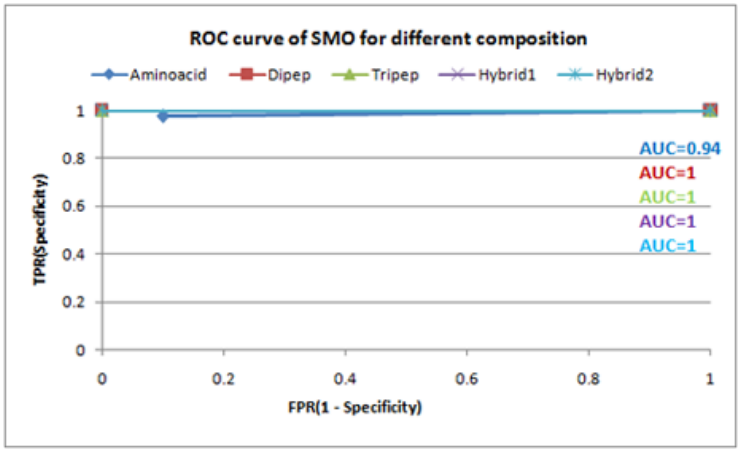

Fig. 2. ROC curve of SMO algorithm with various composition methods

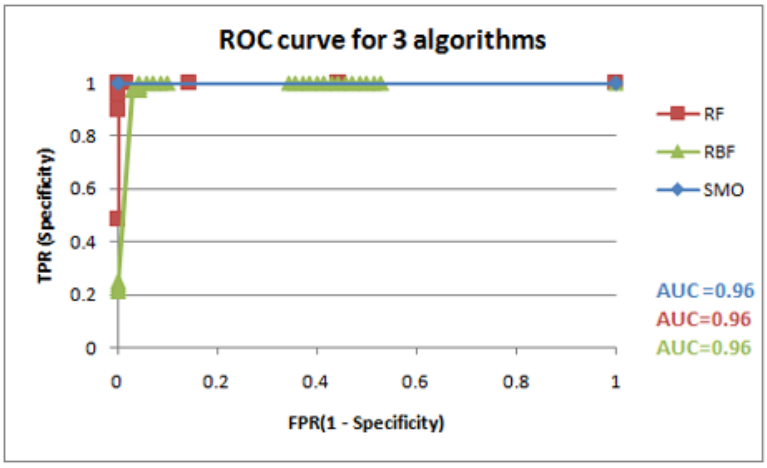

Fig. 3. ROC curve for the three algorithms for the best prediction results

4 Conclusions

Tools and resources are being developed to maximally construe the rice genome sequence. A major difficulty with rice annotation is the lack of accurate gene prediction programs. Rice has a substantial number of genes that are hypothetical in that they are predicted solely on the basis of gene prediction programs, making it vital that the quality of gene prediction programs for rice be improved further. Moreover rice, which is a model species, is the plant in which the function of most cereal genes will be discovered. Thus, the availability of systems/tools that can predict characteristics from sequence is essential to the full characterization of expressed proteins. Computational tools provide faster and accurate access to predictions for any organism and plants.

Identification of NAC proteins from sequence databases is difficult due to poor sequence similarity. In this work, we present a new method for NAC prediction based on SMO implemented in WEKA. The performance was found to be highly satisfactory. Comparison between different machine learning algorithms viz. RBF Network and Random forest was also carried out. Very high prediction accuracies for the validation tests show that NACPred is a potentially useful tool for the prediction of NAC proteins from genome of indica rice.

References

1. Fang, Y., You, J., Xie, K., Xie, W., Xiong, L.: Systematic Sequence Analysis and Identification of Tissue-specific or Stress-responsive Genes of NAC Transcription Factor Family in Rice. Mol. Genet. Genomics 280, 547–563 (2008)
2. Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K., Yu, G.: Arabidopsis Transcription Factors: Genome-wide Comparative Analysis Among Eukaryotes. Science 290, 2105–2110 (2000)
3. Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., Romano, L.A.: The Evolution of Transcriptional Regulation in Eukaryotes. Mol. Biol. Evol. 20, 1377–1419 (2003)
4. Riano-Pachon, D.M., Ruzicic, S., Dreyer, I., Mueller-Roeber, B.: PlnTFDB: An Integrative Plant Transcription Factor Database. BMC Bioinformatics 8, 42 (2007)
5. Chou, K.C., Zhang, C.T.: Prediction of Protein Structural Classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349 (1995)
6. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
7. Platt, J.C.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning, pp. 185–208. MIT Press, Cambridge (1999)
8. Mitchell, T.: Machine Learning. McGraw Hill, New York (1997)
9. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped Blast and PSI-Blast: A New Generation of Protein Database Search Programs. Nucleic Acids Res. 25, 3389–3402 (1997)