ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Development of Prediction Tool for Drought Tolerant Protein in Rice Using Machine Learning Algorithm

Annapoorna Shetty[1], Hemalatha N[1] , Mohammed Moideen Shihab [2], Brendon Victor Fernandes[2]

Assisant Professor, AIMIT, St. Aloysius College, Mangalore, India[1]

[2]Student, Special Interest Group, AIMIT, St. Aloysius College, Mangalore, India[2]

ABSTRACT: Machine learning deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions. Classification is the problem of identifying to which of a set of protein categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known based on positive and negative datasets. This paper primarily emphasizes on the development of prediction tool for drought tolerant protein NAC in rice using Support Vector Machine algorithm. In this paper, we have used seven feature extraction methods including amino acid features, dipeptide, hybrid methods and exchange group features. Using dipeptide features, we have obtained a precision rate of 86% for the NACPredictor tool. This is also further compared with sequence similarity search tool PSI-BLAST.

KEYWORDS: Machine Learning, Classification, SVM

I. INTRODUCTION

In India, drought has resulted in tens of millions of deaths over the course of the 18th, 19th, and 20th centuries. Indian agriculture heavily depends on the climate of India. A favorable southwest monsoon is critical in securing water for irrigating Indian crops. In some parts of India, the failure of the monsoons result in water shortages, resulting in below-average crop yields. This is particularly true of major drought-prone regions such as southern and eastern Maharashtra, northern Karnataka, Andhra Pradesh, Odisha, Gujarat, and Rajasthan.

NAC proteins constitute one of the largest families of plant-specific transcription factors, and the family is present in a wide range of land plants.NAC consists of three different genes namelyNAM (no apical meristem), ATAF (Arabidopsis transcription activation factor), and CUC (cup-shaped cotyledon).It has a conserved domain calledNAC domain (from first letters of each genes).NAC proteins are thought to be involved in a developmental processes like formation of shoot apical meristem (SAM), floralorgans, and lateral shoots, plant hormonal control/defense mechanismsand programmed cell death.NAC family proteins contain a large number of genes around 135 that code for extremely important in plant development.

Machine learning can be defined as construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions. The machine is given training as for humans and is tested with a dataset for its prediction accuracy based on the intelligence obtained from the training set.

The tremendous amount of information is generated based on whole genome sequencing project which is ongoing. In such a scenario, to study about the various newly sequenced genes and to annotate them, manually is cumbersome. Computationally developed tools play a major role in such situations which may help genes for prediction and annotation. In this work, we have carried out the development of a prediction tool for drought tolerant protein NAC using the machine learning algorithm Support Vector Machine.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

II. METHODS

2.1. Data retrieval

Selecting dataset for the development is a major task. In this paper, we have selected 95 NAC proteins and 86 non NAC proteins for training. The positive dataset of NAC proteins were taken from Uniprot knowledgebase. Proteins which were putative uncharacterized proteins were run through Prosite and Pfam to confirm their protein family. The set of 86 non NAC datasets were constructed from plants which were non NAC proteins from other plants. Finally from the training set of positive and negative dataset 10 proteins were removed for testing because independent data testing was carried out. In this testing training and testing sets has to be independent on one another.

2.2. Methods for feature extraction

1) Residue method: Amino acid present in a protein encapsulates information of each amino acid present in a protein sequence. In this method of feature extraction, a protein is represented by a feature vector of 20. In another method to encapsulate the global information of each protein sequence utilizing the sequence order, dipeptide methods was calculated where two amino acid sequence was taken from the sequence for calculation. This consists of 400 features.

2) Exchange group: In this method, we adopted a 8-letter exchange group based on aminoacid properties {e1, e2, e3, e4, e5, e6, e7, e8} to represent a protein sequence where e1 ϵ {G, A, L, V, I}, e2 ϵ {F,Y,W},e3 ϵ {S,T},e4 ϵ{D,E},e5 ϵ{N,Q},e6 ϵ{R,K,H},e7ϵ{C,M},e8 ϵ{P}. These exchange groups are effectively classes of amino acids that has similar properties. Here, the protein sequence are represented as exchange groups. In first method, count of amino acids based on exchange group is considered. In the second method 2-gram exchange group is considered where two from exchange group is considered.For each protein sequence, both single letter exchange group and 2-gram exchange group is considered. Here the former has a feature dimension of 8 and later has a dimension of 64 features.

3) Hybrid group: In this paper, we have considered 3 hybrids which are the combination of different methods. Hybrid1 was developed by combining amino acid and dipeptide features of amino acid. Hybrid 2 was the combination of amino acid and single letter exchange group and hybrid 3 was developed as a combination of single letter and 2-gram exchange group. For each hybrid method, input feature had a dimension of 420 (20 + 400), 28 (20 + 8) and 72 (8 + 64) respectively.

2.3. Machine learning algorithm

Support Vector Machine: In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis [1][2]. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. This is depicted in the Figure 1.

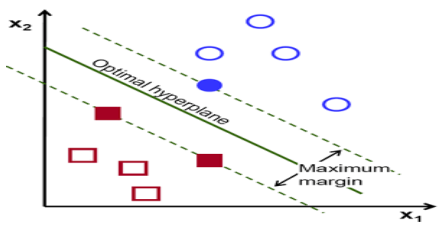

Figure 1: Optimal hyperplane in SVM

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

2.4. Performance Evaluation

For evaluation of performance of the prediction tool independent data test was carried out [4]. Jackknife validation is considered the most perfect test but because of large dataset we have carried out only independent data test validation. In this validation, training dataset and testing set was considered to be independent of one another. Hence the name. The tool was evaluated by different performance metrics namely sensitivity, specificity, prediction, F-measure and Mathews correlation coefficient (MCC).

a. Sensitivity and specificity

These statistical measures of the performance of a binary classification test, also known in statistics as classification function. Sensitivity (also called thetrue positive rate, or the recall rate in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition), and is complementary to the false negative rate. Specificity (sometimes called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate. A perfect predictor would be described as 100% sensitive (i.e. predicting all people from the sick group as sick) and 100% specific (i.e. not predicting anyone from the healthy group as sick); however, theoretically any predictor will possess a minimum error bound known as the Bayes error rate.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{Specificity} = \frac{TN}{FP+TN} \quad (2)$$

b. Mathews correlation coefficient

The **Mathews correlation coefficient** is used in machine learning as a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation. The statistic is also known as the phi coefficient.

$$MCC = \frac{(TPXTN)-(FPXFN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3)$$

c. Precision

In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

$$Pr = \frac{TP}{TP+FP} \quad (4)$$

d. F-measure

In statistical analysis of binary classification, the F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score: *p* is the number of correct results divided by the number of all returned results and *r* is the number of correct results divided by the number of results that should have been returned. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0.The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall:

$$\text{F-measure} = \frac{2\ X\ PR\ X\ SN}{PR+SN}. \quad (5)$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

III. MATERIALS & METHODS

3.1. Analysis of Independent data test

On analyzing the independent data test result from the table (Table 1), we have the composition of dipeptide feature having better MCC and F-measure values compared to other compositions. This feature method which has a feature dimension of 400 and has a precision of 86%, F-measure which combines precision and recall 71% and MCC of 0.52 for linear kernel. These values are much better compared to other compositions and also with different kernels.

Comparable to dipeptide feature, hybrid 1, double exchange group, amino acid and hybrid 2 features with linear kernel are also having fair results better but not as good as dipeptide. This is depicted in the figure (Figure 2).

3.2 Comparison of prediction tool with similarity search tool

To summarize the evolutionary information about theproteins, similarity search PSI-BLAST is carried out. To produce thehomology of the given sequence with other related sequences in the database, a protein sequence in the test set was compared with a created database which provided a broad range of information about eachfunctional encoded protein in the database[18]. A 10-fold cross-validation was conducted with PSI-BLAST which achieved no significant hits and an accuracy of only 53.5% (Table 2).This result suggests that similarity based search tools are not efficient and consistent as compared to the different composition based modules implemented based on computational methods.

Table 1: Independent data results for NACPredictor

Method	Kernel	Independent data test validation				
		Sensitivity	Specificity	Precision	F-Measure	MCC
Amino	Linear	0.70	0.80	0.78	0.74	0.50
	Poly	0.10	0.90	0.00	0.0	0.00
	RBF	1.0	0.0	0.50	0.67	0.0
Dipeptide	Linear	0.60	0.90	0.86	0.71	0.52
	Poly	0.30	0.90	0.00	0.00	0.25
	RBF	1.00	0.00	0.50	0.67	0.00
Monomer	Linear	0.40	0.70	0.57	0.47	0.10
	Poly	1.00	0.00	0.00	0.00	0.00
	RBF	1.00	0.00	0.50	0.67	0.00
2 gram	Linear	0.50	0.90	0.83	0.63	0.44
exchange	Poly	0.30	0.90	0.00	0.00	0.25
	RBF	1.00	0.00	0.50	0.67	0.00
Hybrid1	Linear	0.70	0.80	0.78	0.74	0.50
	Poly	0.20	0.80	0.00	0.00	0.00
	RBF	1.00	0.00	0.50	0.67	0.00
Hybrid2	Linear	0.60	0.80	0.75	0.67	0.41
	Poly	0.30	0.80	0.00	0.00	0.12
	RBF	1.00	0.00	0.50	0.67	0.00
Hybrid3	Linear	0.30	0.80	0.60	0.40	0.12
	Poly	1.00	0.10	0.00	0.00	0.23
	RBF	1.00	0.00	0.50	0.67	0.00

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Table 2: Prediction result of NAC proteins with similarity search

Test	No. of sequences	Correctlypredicted	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

3.3. Comparison of prediction tool with other prediction tools

Predicting the characteristics of an unknown gene is useful for gene annotation especially in the era of high genome sequencing. Hemalatha*et al.* in their paper have described a tool NACSVM using Support Vector Machine and PSSM composition [1]. They obtained an accuracy of 100% with respect to the tool. In this paper, we have attempted with some other features and have obtained 86% precision for the same drought tolerant gene NAC. Several extension of this approach is possible with more feature extraction methods. More hybrid methods can be tried for this tool which may give some better results.

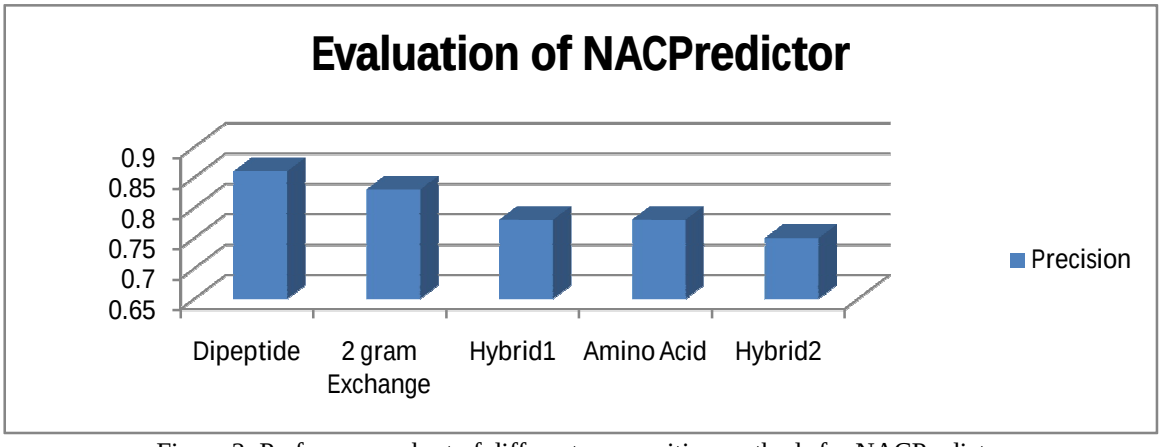

Figure 2: Performance chart of different composition methods for NACPredictor

3.4. Description of web based tool

NACPredictor is a dynamic web server implemented on the World Wide Web using the best performing algorithm. The tool was implemented using PHP and HTML scripting language. Tool is user friendly and allows the user to enter the queries either through standard FASTA format or allows uploading of sequence through a file. (Figure 3). The result of the user entered sequence will be displayed in a page with more description. The overall architecture of the tool is depicted in the Figure 4 .

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Figure 3: Web page for the NACPredictor tool

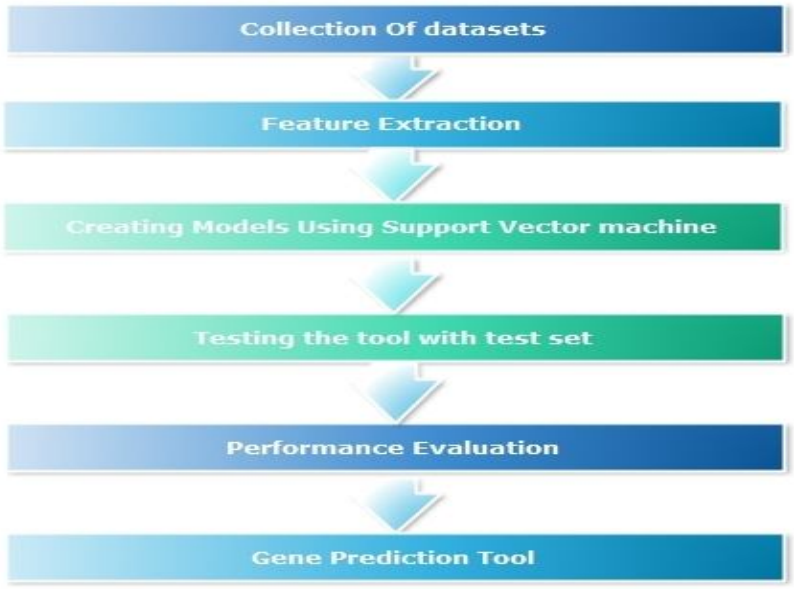

Figure 4: Architecture of the algorithm implemented for NACpredictor tool

IV. CONCLUSION

Computational tools as compared to techniques implemented experimentally provide faster and accurate prediction for any organism or plant. There is a lack of gene prediction programs with respect to rice functionalities and various strains. Because of the availability of rice genome, development of tools for various strains and functionalities of rice are something which is achievable. In this paper, we have attempted some new methods of feature extraction for the development of prediction tool and the performance of the same was found to be satisfactory.

REFERENCES

1. C. Cortes & V. Vapnik, (1995) "Support vector networks", Machine Learning, Vol. 20, No. 5, pp 273–297.
2. V. Vapnik, (1995) The Nature of Statistical Learning Theory, Springer, New York
3. N.Hemalatha, M.K. Rajesh,. N. K. Narayanan, "An integrative system for Prediction of NAC proteins in rice using different feature extraction methods," International Journal on Soft Computing (IJSCC), Vol.4(1), pp 9-21,2013

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

4. Hemalatha, N., M. K. Rajesh, and N. K. Narayanan. "NACPred: Computational Prediction of NAC Proteins in Rice Implemented Using SMO Algorithm." *Advances in Computing, Communication, and Control*. Springer Berlin Heidelberg, 2013. 266-275.
5. 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.