A machine learning approach for detecting MAP kinase in the genome of *Oryza sativa L. ssp. indica*

Hemalatha N., Rajesh M.K. and Narayanan N.K.

***Abstract*—Plant development and crop yield are highly influenced by temperature. High temperature negatively affects different stages of plant development in rice, mainly booting and flowering. Identifying candidate genes associated with high-temperature stress response may provide knowledge for the improvement of heat tolerance in rice. As the rice genome sequencing has already been undertaken, a major work challenge is annotating proteins and decoding their functionalities. MAP kinase (MAPK) proteins are involved in signaling various abiotic and biotic stresses, like temperature stress or drought, wounding and pathogen infection. Moreover, MAPKs have also been implicated in cell cycle and developmental processes. In this study, an attempt has been made in developing a MAP kinase prediction tool for rice, MapPred. The computational approach has been developed using Sequential Minimum Optimization (SMO) algorithm in Weka workbench, and a sensitivity of 100% was obtained using dipeptide method. MapPred was also tested with three plants, namely *Arabidopsis*, maize and tomato to prove that developed tool has higher accuracy with rice than other plants which further proves the higher prediction accuracy of species-specific tools. Prediction performance of MapPred was evaluated using cross validation, independent data test and leave one out validation. Our experimental results demonstrated that proposed algorithm based on dipeptide method could be very effective in the computational approach for predicting MAPK proteins in *Oryza sativa*subsp.*indica*.**

I. Introduction

Rice, *Oryza sativa L.*, is a major cereal grown across the world and has been used as a model plant for many years.Rice production may be adversely affected by the predicted 24 degree Celsius increment in temperature by the end of the 21st Century[1]. The problem in precisely predicting the impacts of climate change in agriculture further deteriorates the situation [2]. According to Nagai and Makino, the growth responses of rice to high temperature are still poorly understood [3].

Temperature is one of the main driving forces for crop development [4]. Temperature ranging from 27 to 32 degree Celsius is the optimum temperature for the normal development of rice [5]. Almost all the growth stages of rice are affected by high temperature, i.e. from emergence to ripening and harvesting. The severity of the possible damage to the crop is determined by the developmental stage at which the plant is exposed [6]. However, the most susceptible stages in rice to higher temperature are flowering (anthesis and fertilization) and to a lesser extent the preceding stage i.e. booting [7],[8]. Temperatures higher than the optimum suggested, induces floret sterility and decreases grain plumpness, starch content, and protein accumulation, thus decreasing both yield and quality [9],[10], [11].

Development of stress-tolerant crops will be of huge benefit in modern agriculture, especially in regions exposed to regular stress conditions. Taking this into consideration, there has been development in identifying potential stress tolerant genes in recent years.

Mitogen activated protein kinases (MAPK) are proteins which can be an important mediator in transmission of signal and connecting the perception of external stimuli to cellular responses [12]. These proteins are also involved in signaling various stresses like high temperature or drought. MAPK are universal signal transduction modules in eukaryotes such as plants.

High-throughput genome sequencing in the past few years and the completion of annotation of genomic sequences of many plants especially *Oryza sativa* has been a challenging area in the field of bioinformatics. Because of this, bioinformatics approaches for the development of gene/protein prediction tools have attracted a greater deal of research interest in this area [13],[14].

The present work introduces a machine learning approach, MapPred, for predicting high temperature responsive genes in indica rice *(Oryza sativa L. ssp. indica)*. High-throughput analyses of large-scale genome sequences rely on computational prediction methods which are much superior compared to old experimental methods on the basis of speed, accuracy and reliability. MapPred, which is developed in Weka workbench using SMO classifier, is compared with many other algorithms and validated using various validation techniques. Finally, a web based server has also been developed using SMO where the users are given the option of inputting the sequence and viewing the output in a user friendly manner.

II. Materials and methods

A. Datasets

The selection of dataset for the prediction method is considered the most important task during the development of a prediction tool. A set of 81 MAP kinase proteins of rice were extracted from the Uniprot knowledgebase. Proteins which were putative uncharacterized proteins were run through Prosite and Pfam databases to confirm their protein family. Proteins having uncertain annotation labels were not considered. Sequences selected for creating the model should be highly curated data and should not contain homologous genes from various organisms. CD-HIT was

Hemalatha N. is with Dept. of Computer Science, AIMIT, St. Aloysius college, Beeri, Mangalore, Rajesh M.K. is with Central Plantation Crops Research Institute,Kasargod, Kerala and Narayanan N.K. with Dept. of Information Science Technology,Kannur University (Corresponding author email: {hemasree71}@gmail.com).

used to remove sequences with high degree of similarity to other sequences to construct a non-redundant subset, in which the homologous sequences with sequence similarity greater than 95% were excluded [15]. Similarly, a set of negative samples, which were non-MAP kinase, was constructed from other plants. Finally, the training dataset consisted of 71 MAP kinase proteins in rice that formed the positive dataset and 70 non-MAP kinase proteins from other plants which constituted the negative dataset.

B. Features

1) Residue method: Amino acid method encapsulates the information of each amino acid in the protein sequence. In this method a protein is represented by a vector of dimension 20. In order to encapsulate the global information of each protein sequence utilizing the sequence order effects, dipeptide method was computed. This method consisted of 400 (20 x 20) dimensional vectors of each protein sequence. Tripeptide method containing information about the amino acid method, along with the local order of amino acids in the given protein sequence, was also computed. This provides a 8000 (20 x 400) dimensional feature vector for the given protein sequence. To calculate the fraction of each amino acid, dipeptide and tripeptide methods, following equations were used.

$$P\left(a_i\right) = \frac{Na_i}{\sum_{j=1}^{20} Na_j} \tag{1}$$

$$P\left(a_i a_j\right) = \frac{Na_i a_j}{\sum_{i=1}^{20}\sum_{j=1}^{20} Na_i a_j} \tag{2}$$

$$P\left(a_i a_j a_k\right) = \frac{Na_i a_j a_k}{\sum_{i=1}^{20}\sum_{j=1}^{20}\sum_{k=1}^{20} Na_i a_j a_k} \tag{3}$$

where $P(a_i)$ represents the fraction of each $(a_i)^{th}$ amino acid, $N(a_i)$ gives the total number of $(a_i)^{th}$ amino acid and denominator represents the total number of amino acids present in the given protein sequence.

2) Composition of terminal residues: In this method, protein sequence was divided into n overlapping fragments and calculates the amino acid composition from each fragment separately using (Eq.1). Thus the dimension of final input vector will have n x 20 dimensions, where n is the number of fragments. There are 2 types of terminal residue composition *viz*. N-terminal and C-terminal composition. N terminus is calculated on the basis that many proteins in the plant cell have sorting signals which depends on the presence of an N terminal targeting sequence. Also these signals are responsible for targeting proteins to various subcellular localizations in the cell. For both N terminal and C terminal residue lengths of 15, 30 were tried to find optimal length.

3) Evolutionary information based features: PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) is used as a strong measure of residue conservation for any particular location. When an input sequence has no alignment with the database, this tool returns 20-dimensional vectors representing probabilities of conservation against mutations to 20 different amino acids including the input sequence. A matrix that contains vector representations of this type for the given sequence is called position specific scoring matrix (PSSM). During the iterations of PSI-BLAST when a residue in the sequence is conserved then this may be due to some biological function and hence it represents the evolutionary information of a protein sequence. This input information is expressed in a profile (position specific scoring table) created from a group of sequences aligned by PSI- BLAST previously against nonredundant database. The PSSM matrix consists of L rows and 20 columns for a protein chain of L amino acid residues. Here 20 columns represent the occurrence of each type of 20 amino acids. For finding a particular amino acid, PSSM gives the log-odds score for the target sequence. Compared to other sequence similarity approaches this method allows more information to be used in testing of the target sequence by having any number of known sequences in profile construction. Later every element in PSSM is divided by the length of the given sequence and then reduced to a scale of 0-1 using the linear function $\frac{X-Minimum}{Maximum-Minimum}$. Here X stands for individual PSSM score of each amino acid, Minimum stands for minimum score in the PSSM and maximum is the maximum score in the PSSM. A 400 dimensional feature vector is generated by adding all rows corresponding to the same amino acid in the primary sequence. The conversion process of L x 20 sized PSSM to 20 x 20 dimensional feature vectors is diagrammatically represented in the Figure 1 .

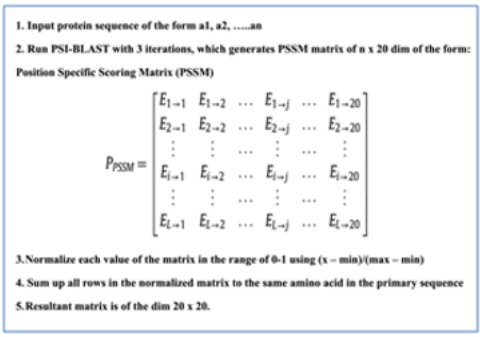

Fig. 1. Schematic diagram of the algorithm used to convert PSSM matrix from Lx20 vector to 400 D vector.

4) Hybrid-based Method: In this approach, different methods are combined to obtain new ones. Three hybrid approaches were used in this study. Hybrid-1 was developed by combining amino acid and dipeptide features of a protein sequence using (Eq.1) and (Eq.2), Hybrid-2 by combining amino acid and tripeptide features of a protein sequence using (Eq.1) and (Eq.3). Finally, amino acid calculated using (Eq.1) was combined with the evolutionary information stored in the matrix of a protein sequence called PSSM to generate Hybrid-3. For each method, input feature vector pattern had a dimension of 420 (20 + 400), 8020 (20 + 8000) and 420 (20 + 400) respectively.

C. The machine learning algorithms

1) Sequential Minimum Optimization (SMO): SMO is a support vector machine learning algorithm (SVM) that is theoretical wise simple, easily executable, usually faster, and has better scalable properties for complicated SVM problems compared to standard and simple SVM training algorithm. In the case of training, support vector machine

requires more time in solving of a very large quadratic programming (QP) optimization problem which is usually time consuming. Same QP problems, when given to SMO, is broken into a series of smaller QP problems which can be solved analytically faster than SVM and also avoids using time consuming numerical QP optimization as an inner loop. SMO can easily handle very large training set because the amount of memory required for SMO is linear in training set size. Another main difference between SMO and SVM is that, in SMO matrix calculation is avoided which makes it scalable between linear and quadratic in the training set size for various test problems whereas in SVM algorithm it scales between linear and cubic in the training set size. For sparse data sets, calculating speed for SMO is highly dominative when compared to solving linear SVMs.

2) J48 decision tree: A J48 decision tree is a predictive machine-learning model. This algorithm decides the value of the target (dependent variable) of a new sample based on various attribute values of the available data. In the decision tree, the internal nodes denotes the different attributes, the branches between the nodes represents the possible values that these attributes can have in the observed samples, while the terminal nodes denotes the final value (classification) of the target variable. Dependent variable (target variable) is the attribute that is to be predicted and is called dependent since its value depends upon, or is decided by, the values of all the other attributes. On the other hand, independent variables are those attributes which help in predicting the value of the dependent variable in the dataset.

3) ADtree: An alternating decision tree (ADTree) is a machine learning method which is used for classification. This method generalizes decision trees and has association with boosting. Boosting is a machine learning meta-algorithm which reduces biasedness in supervised learning. An alternating decision tree consists of two nodes *viz* decision node and prediction node. Decision nodes specify a predicate condition whereas prediction nodes contain a single number. An alternating decision tree always has prediction nodes as both roots and leaves. In ADTree, by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed, an instance is classified.

D. Similarity search

In similarity search method, PSI-BLAST was used for searching a query sequence against existing non-redundant database of classified proteins. In the present case, database contained 81 MAP kinase proteins. PSI-BLAST when compared to normal BLAST gives better result in detecting remote homologies [16]. This tool also detects close relationships between proteins that are functionally or structurally distant which makes it much more capable than BLAST.

E. Performance evaluation

For evaluation of performance of the prediction tool, jackknife or leave one out cross-validation is considered to be the most perfect test. But when dataset is large, jackknife validation test takes very long time and leave one out tests, along with other validation tests, are considered. Here we have conducted 10-fold and leave one out cross validation along with independent dataset test validation. In 10-fold cross validation, nine parts were used for training and remaining one for testing. This process was repeated ten times so that each of the ten set was used for testing at least once. In the case of independent dataset test, training and test set are created such that each data in both the sets are unique. The performances of the models are assessed by both threshold-dependent and threshold-independent metrics. Several measures used for threshold-dependent metrics are sensitivity (Sn) or recall, specificity (Sp), precision (Pr), F-measure (F) and Mathew correlation coefficient (MCC). Following are the brief description of each parameters; sensitivity (recall) measures the percentage of correctly predicted MAP kinase proteins, specificity measures the percentage of correctly predicted non-MAP kinase proteins, F-measures computes some average of the information retrieval precision, recall and Mathews correlation coefficient (MCC) measures the quality of prediction and balances the positive and negative data [17]. F-measure is computed using the harmonic mean and larger of this value indicates a higher classification accuracy. Perfect prediction has MCC value equal to 1, 0 for random prediction and -1 as worst prediction. These parameters are calculated using the following equations

$$Sensitivity = \left(\frac{TP}{TP+FN}\right) * 100 \tag{4}$$

$$Specificity = \left(\frac{TN}{FP+TN}\right) * 100 \tag{5}$$

$$F-measure = \left(\frac{2*Pr*Sn}{Pr+Sn}\right) * 100 \tag{6}$$

$$Precision = \left(\frac{TP}{FP+TP}\right) * 100 \tag{7}$$

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

where TP, TN are correctly predicted positive and negative proteins and FP, FN are wrongly predicted positive and negative proteins respectively.

III. Results and Discussion

A. Analysis of Independent data test

In statistical prediction, three validation approaches generally used to examine a predictor for its effectiveness in practical applications are independent dataset test, cross validation test and jackknife test. In the present study, because of large dataset and to reduce computational time, we adopted independent dataset test and cross validation test.

On applying independent dataset test on all nine different methods, it was observed that classifier sequential minimum optimization (SMO) obtained 100% sensitivity with a high confidence of MCC 1 and F-measure of 100 for dipep, tripep, hybrid-1, hybrid-2 and hybrid-3 methods (Table I). Here SMO classifier achieved an overall prediction accuracy of 100% with specificity also being 100%. Sensitivity and specificity are the two competing nonexclusive measures of quality used to measure the performance of classification methods. The MCC provides a balanced measure between

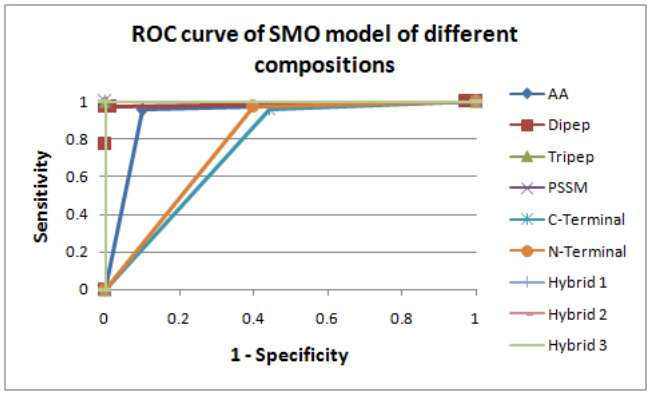

Fig. 2. ROC curve of SMO model of different methods for MapPred

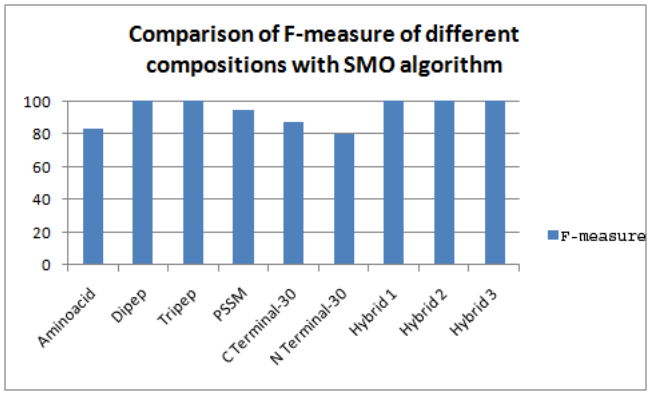

Fig. 3. Comparison of accuracy of nine methods with SMO classifier

TABLE I. INDEPENDENT DATATEST RESULTS OF MAPPRED WITH NINE METHODS

Method	Algorithm	Independent data test				
		Sn	Sp	Acc	F	MCC
Aminoacid	J48	100	100	100	86	1
	SMO	100	100	100	83	1
	Adtree	100	80	90	91	0.82
Dipep	J48	70	100	85	86	0.73
	SMO	100	100	100	100	1
	Adtree	90	80	85	86	0.70
Tripep	J48	100	100	100	100	1
	SMO	90	100	95	100	0.90
	Adtree	100	100	100	100	1.00
PSSM	J48	90	100	95	95	0.90
	SMO	100	90	95	95	0.90
	Adtree	100	90	95	95	0.90
C Terminal-30	J48	80	100	90	50	0.82
	SMO	80	100	90	87	0.82
	Adtree	100	100	100	100	1.00
N Terminal-30	J48	90	100	95	95	0.90
	SMO	90	100	95	80	0.90
	Adtree	100	80	90	91	0.82
Hybrid 1	J48	70	100	85	86	0.73
	SMO	100	100	100	100	1
	Adtree	90	80	85	86	0.70
Hybrid 2	J48	100	100	100	100	1
	SMO	90	100	95	100	0.90
	Adtree	100	100	100	100	1.00
Hybrid 3	J48	100	90	95	100	0.90
	SMO	100	90	95	100	0.90
	Adtree	100	100	100	100	1.00

TABLE II. CROSS VALIDATION RESULTS OF MAPPRED WITH NINE METHODS

Method	Algorithm	10-fold Cross validation				LOO Cross validation			
		Sn	Sp	F	MCC	Sn	Sp	F	MCC
Aminoacid	SMO	53	23	66	-0.14	53	23	66	-0.14
	J48	50	50	93	0.00	51	29	65	-0.09
	ADtree	51	20	95	-0.11	52	0	67	-0.16
Dipep	SMO	51	0	67	-0.14	52	0	67	-0.16
	J48	51	0	67	-0.12	51	0	67	-0.12
	ADtree	50	50	67	0.00	51	0	67	-0.12
Tripep	SMO	97	100	99	0.97	97	100	99	0.97
	J48	100	96	98	0.96	100	96	98	0.96
	ADtree	99	100	99	0.99	99	99	99	0.97
PSSM	SMO	100	96	98	0.96	97	100	98	0.97
	J48	99	100	99	0.99	100	96	99	0.96
	ADtree	100	99	99	0.99	99	100	99	0.99
C Terminal-30	SMO	64	9	66	-0.26	96	56	80	0.56
	J48	51	48	63	-0.01	87	87	87	0.74
	ADtree	54	33	64	-0.10	93	84	89	0.78
N Terminal-30	SMO	97	56	81	0.58	97	54	80	0.57
	J48	80	74	78	0.55	82	74	79	0.56
	ADtree	92	83	88	0.75	89	80	85	0.69
Hybrid 1	SMO	100	96	98	0.96	100	96	98	0.96
	J48	100	97	99	0.97	100	97	99	0.97
	ADtree	100	99	99	0.99	100	97	99	0.97
Hybrid 2	SMO	99	100	99	0.99	99	99	99	0.97
	J48	100	96	98	0.96	100	96	82	0.82
	ADtree	99	100	99	0.99	99	99	99	0.97
Hybrid 3	SMO	100	99	99	0.99	100	99	99	0.99
	J48	100	99	99	0.99	100	99	99	0.99
	ADtree	97	99	98	0.96	97	99	98	0.96

these two nonexclusive measures of quality. An ideal classification method should have an MCC equal to 1 and sensitivity, specificity and precision values either equal to or close to 100%. The F1-measure effectively references the True Positives to the Arithmetic Mean of Predicted Positives and Real Positives, being constructed as rate normalized to an idealized value. F1-measure expressed in this form is known in statistics as a Proportion of Specific Agreement as it is applied to a specific class and hence is applied to the Positive Class. To further select the best method for MapPred among the five, receiver operator curve (ROC) was plotted for all the methods of SMO classifier Figure 2 . From Figure 2, we can conclude that dipeptide, tripeptide, hybrid-2 and hybrid-3 methods has area under the curve (AUC) of 1. We have, dipep method having a feature vector of dimension 400 and others having features of dimension more than 400. Since dipeptide method has 100% precision with less number of features we arrive at the conclusion that dipeptide method with SMO classifier is the best for the prediction tool MapPred. Performance comparison of the nine methods with SMO classifier is represented in the Figure 3 .

B. Analysis of cross validation data test

Even after applying 10-fold cross validation and leave one out cross validation for all the nine different methods, the overall accuracy could not match with that of independent data test. As shown in the Table II, we obtained a maximum of 99% precision with MCC of 0.9 and F-measure of 99 for tripeptide, hybrid-2 and hybrid-3 method with SMO classifier which has large number of features compared to that obtained in independent data test. Sensitivity and specificity, the two competing non-exclusive measures of quality also showed 99%.

C. Comparison of MapPred with other machine learning methods

The proposed MapPred method was compared with two other classifiers namely J48 decision tree and ADtree classifier using all the nine features tested for SMO classifier. All models were tested with the same test dataset as used

TABLE III. PREDICTION RESULT OF MAP KINASE PROTEINS WITH SIMILARITY SEARCH (10 FOLD CROSS VALIDATION)

Test	No. of sequences given	Correctly predicted	Accuracy
1	18	10	55.5
2	18	11	61
3	18	10	55.5
4	18	11	61
5	18	10	55.5
6	18	10	55.5
7	18	10	55.5
8	18	10	55.5
9	18	10	55.5
10	20	10	50
			56

TABLE IV. PREDICTION RESULT OF MAPPRED WITH THREE PLANTS USING SMO CLASSIFIER

Plants	No. of sequences given	Correctly predicted	Accuracy
Arabidopsis	40	35	88
Maize	40	38	95
Tomato	23	22	96

for SMO. The optimal values obtained for these algorithms were comparable with respect to all statistical values to that of SMO classifier, however, the method used for these classifiers were feature vectors of higher dimension. For SMO classifier, we could obtain optimal classification for dipep method which had a feature vector of dimension 400.

D. Sequence similarity search

The sequence similarity search tool, PSI-BLAST was used to compare the protein query sequence against a database containing all the protein sequences. PSI-BLAST when compared to BLAST has the advantage of using a profile to search a database which often detects remote homology between proteins which are structurally or functionally distant. To conduct similarity search, we have used 10-fold cross validation which generated no much significant hits and an accuracy of only 56% Table III . This result indicated that sequence similarity based tools are not much efficient and consistent as compared to modules based on computational methods.

E. Performance on other plants

For further validating the MapPred tool, we cross checked the performance of MapPred on three other plants namely *Arabidopsis*, maize and tomato which contained a total of 174 Map kinase proteins. If there are any species specific features of protein in MapPred, the performance of this tool on other plants should have lower accuracy. To verify this, MapPred with SMO classifier was run with best method obtained for the three diverse plants. The results obtained for the three plants are tabulated Table IV . The table shows the accuracy of three plants using MapPred tool which are 88%, 95% and 96% respectively. However, when the same model was run on indica rice, the overall prediction accuracy was 100% during independent dataset test. This difference between the performances of MapPred with indica rice and other plants proves that there might be some species specific features of protein that has resulted in the better performance of *Oryza sativa* -specific classifer on its proteins and lower on other proteomes.

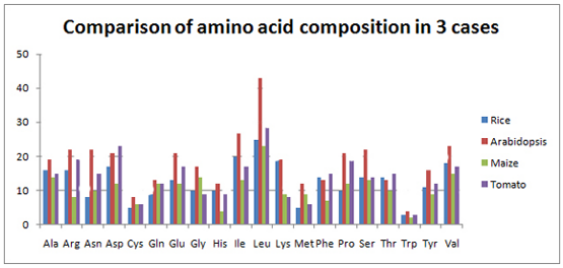

Fig. 4. Comparison of amino acids in rice and three plants

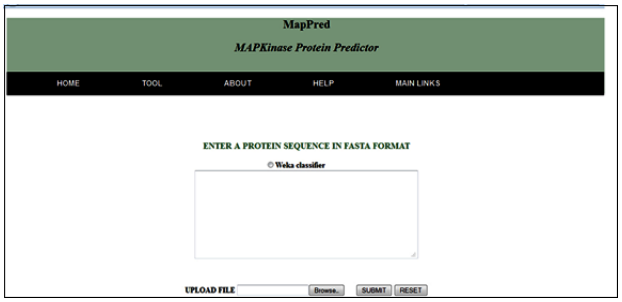

Fig. 5. Web page to show the user input screen of MapPred

F. Reason for prediction differences

The above tests show that a species specific predictor has always a higher accuracy with respect to the same species compared to other plants. To test the variation in prediction performance, we analyzed the variation in the amino acid method across the three plants used above and compared with the amino acid composition of the indica rice proteome. There are various studies relating the fact that amino acid composition differs across species [18],[19],[20],[21],[22],[23]. In our analysis, we also found significant differences in the amino acid methods among the three plants and rice (Figure 4). This change in the composition of amino acid across plants suggests that the differences in prediction performance can be correlated with them and hence developing a species specific predictor will be more realistic to achieve better accuracy.

G. Description of web server

Based on the work, a webserver MapPred has been developed for predicting MAP kinase rice proteins. The common gateway interface (CGI) script was written in programming language Perl and web interface in HTML and PHP to

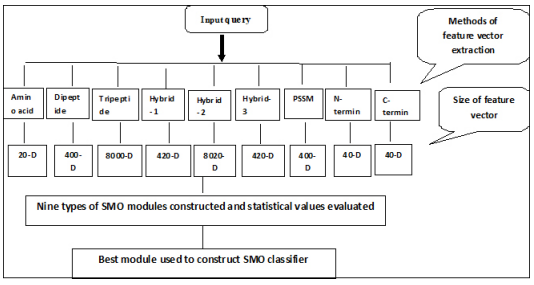

Fig. 6. Overall architecture of MapPred

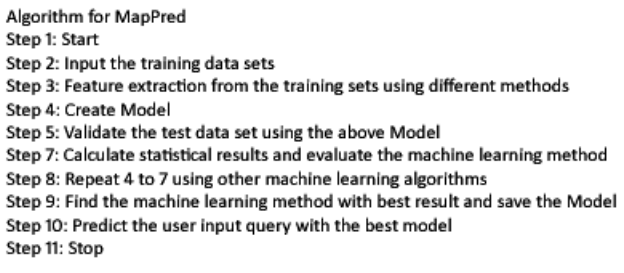

Fig. 7. Algorithm of MapPred

assess user specific queries. The web page allows the user to submit/paste their queries in the FASTA format or upload a sequence in the form of a file Figure 5 . The overall architecture and algorithm of MapPred has been depicted in the Figure 6 and Figure 7 .

IV. Conclusion

The identification of various functional proteins has been facilitated by the use of bioinformatics tools and web-available databases. Hence the prediction of different proteins can be an important step towards understanding the functional characteristic and interactions of these proteins. Here we have developed a highly accurate and species specific MAP kinase prediction method MapPred. The independent dataset method with dipep method showed an accuracy of 100% using MapPred. With the study of MapPred with other plants, we could also reveal the advantage of using species specific predictor (MapPred) with rice rather than with other plants and thus better suited for the respective species specific annotations. Hence we believe that the newly developed tool MapPred will contribute significantly to the future annotation projects and development of predictors.

References

[1] I. P. O. C. Change, "Report of the nineteenth session of the intergovernmental panel on climate change (ipcc) geneva, 17-20 (am only) april 2002," 2007.

[2] T. Watanabe and T. Kume, "A general adaptation strategy for climate change impacts on paddy cultivation: special reference to the japanese context," *Paddy and Water Environment*, vol. 7, no. 4, pp. 313–320, 2009.

[3] T. Nagai and A. Makino, "Differences between rice and wheat in temperature responses of photosynthesis and plant growth," *Plant and cell physiology*, vol. 50, no. 4, pp. 744–755, 2009.

[4] H. V. L. M. J. Kropff, R. B. Mathews and H. F. Berge, "The rice model oryza 1 and its testing," in *Modeling the Impact of Climate Change on Rice Production in Asia*, 1995, pp. 27–50.

[5] K. J. M. YIin, Xinuou, Goudriaan, and Jan, "Differential effects of day and night temperature on development to flowering in rice," *Annals of Botany*, vol. 77, no. 3, pp. 203–213, 1996.

[6] A. Wahid, S. Gelani, M. Ashraf, and M. Foolad, "Heat tolerance in plants: an overview," *Environmental and Experimental Botany*, vol. 61, no. 3, pp. 199–223, 2007.

[7] T. Satake and S. Yoshida, "High temperature-induced sterility in indica rices at flowering," in *Proceedings of the Crop Science Society of Japan*, vol. 47, 1978.

[8] T. Farrell, K. Fox, R. Williams, and S. Fukai, "Genotypic variation for cold tolerance during reproductive development in rice: screening with cold air and cold water," *Field Crops Research*, vol. 98, no. 2, pp. 178–194, 2006.

[9] S. Morita, J.-I. Yonemaru, and J.-i. Takanashi, "Grain growth and endosperm cell size under high night temperatures in rice (oryza sativa l.)," *Annals of Botany*, vol. 95, no. 4, pp. 695–701, 2005.

[10] C.-J. Lin, C.-Y. Li, S.-K. Lin, F.-H. Yang, J.-J. Huang, Y.-H. Liu, and H.-S. Lur, "Influence of high temperature during grain filling on the accumulation of storage proteins and grain quality in rice (oryza sativa l.)," *Journal of Agricultural and Food Chemistry*, vol. 58, no. 19, pp. 10 545–10 552, 2010.

[11] H. Jiang, W. Dian, and P. Wu, "Effect of high temperature on fine structure of amylopectin in rice endosperm by reducing the activity of the starch branching enzyme," *Phytochemistry*, vol. 63, no. 1, pp. 53–59, 2003.

[12] M. Wrzaczek and H. Hirt, "Plant map kinase pathways: how many and what for?" *Biology of the Cell*, vol. 93, no. 1-2, pp. 81–87, 2001.

[13] B. A. Vinatzer, J. Jelenska, and J. T. Greenberg, "Bioinformatics correctly identifies many type iii secretion substrates in the plant pathogen pseudomonas syringae and the biocontrol isolate p. fluorescens sbw25," *Molecular plant-microbe interactions*, vol. 18, no. 8, pp. 877–888, 2005.

[14] L. M. Schechter, M. Vencato, K. L. Jordan, S. E. Schneider, D. J. Schneider, and A. Collmer, "Multiple approaches to a complete inventory of pseudomonas syringae pv. tomato dc3000 type iii secretion system effector proteins," *Molecular Plant-Microbe Interactions*, vol. 19, no. 11, pp. 1180–1192, 2006.

[15] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[16] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[17] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[18] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of Molecular Biology*, vol. 238, no. 1, pp. 54–61, 1994.

[19] J. Lobry, "Influence of genomic g+ c content on average amino-acid composition of proteins from 59 bacterial species," *Gene*, vol. 205, no. 1, pp. 309–316, 1997.

[20] M. A. Andrade, S. I. ODonoghue, and B. Rost, "Adaptation of protein surfaces to subcellular location," *Journal of Molecular Biology*, vol. 276, no. 2, pp. 517–525, 1998.

[21] F. Tekaia and E. Yeramian, "Evolution of proteomes: fundamental signatures and global trends in amino acid compositions," *BMC Genomics*, vol. 7, no. 1, p. 307, 2006.

[22] F. Tekaia, E. Yeramian, and B. Dujon, "Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis," *Gene*, vol. 297, no. 1, pp. 51–60, 2002.

[23] N. S. Bogatyreva, A. V. Finkelstein, and O. V. Galzitskaya, "Trend of amino acid composition of proteins of different taxa," *Journal of Bioinformatics and Computational Biology*, vol. 4, no. 02, pp. 597–608, 2006.