Original Article

Journal of Phytology 2014, 6: 33-39
http://scienceflora.org/journals/index.php/jp/

$CnMAPK_{Pred}$: A machine learning approach for predicting mitogen-activated protein kinase in *Cocos nucifera*

V. Akhil,[1] C. U. Rahul,[1] V. Amal[1], N. Hemalatha[2], M. K. Rajesh[1]*

[1]ICAR-Central Plantation Crops Research Institute, Kasaragod, Kerala, India, [2] St. Aloysius Institute of Management & Information Technology, St. Aloysius College, Mangalore, Karnataka, India

Received: 25.09.2014
Revised: 20.11.2014
Accepted: 21.11.2014
Published: 28.11.2014

***Address for correspondence:**
M. K. Rajesh, ICAR-Central Plantation Crops Research Institute, Kasaragod - 671 124, Kerala, India.
E-mail: rajesh.mk@icar.gov.in

ABSTRACT

Mitogen-activated protein kinases (MAPKs) consist of a large group of ubiquitous proline-directed kinases that are specific to the amino acids threonine, serine, and tyrosine. Many cellular functions are handled by diverse and conserved MAPK pathways. The MAPKs sequenced so far share approximately 40% amino acid sequences. In this study, a tool has been developed for the prediction of coconut MAPKS using machine learning approaches. Of the different algorithms tested, Naïve Bayes algorithm gave the best results. To check the predictability and performance of the developed algorithm, cross-validation, and independent data set validations were carried out. The results revealed that the proposed algorithm could be very effective in computationally predicting MAPKs in coconut. This tool is freely available at http://210.212.229.52/cnmapk.

KEY WORDS: Coconut, mitogen-activated protein kinases, Naïve Bayes algorithm, Waikato Environment for Knowledge Analysis

INTRODUCTION

Mitogen-activated protein kinases (MAPKs) are a highly conserved family of ubiquitous proline-directed, protein-serine/threonine kinases. They display a high level of conservation among eukaryotes by virtue of their homologous domains of length 250-300 amino acids but display distinct roles in different organisms (Taj *et al.*, 2010). In plants, they control various functions from growth and development to hormonal and stress responses (Garcia *et al.*, 2012). The presence of MAPKS in prokaryotes gives reason enough to believe that the MAPK progenitor gene could have been present even before the evolution of eukaryotes from prokaryotes (Hanks *et al.*, 1995). MAPK cascades typically exist in three kinase architectures with the MAPK, MAPK activator, and mitogen activated protein kinase kinase (MAPKK) which phosphorylates the MAPK in the cascade (Schaeffer *et al.*, 1999). There may be common kinases among different cascades, but processes such as cross-inhibition, scaffolding, feedback control, and spatiotemporal constraints maintain the signaling specificity (Cristina *et al.*, 2010). In plants, MAPK studies have demonstrated that the activation of the kinase usually takes place due to external impulses or factors such as temperature changes and wounding (Sinha *et al.*, 2011). A study on the *Arabidopsis thaliana* genome has identified a total of 20 MAPKs, 10 MAPKKs, and in excess of 80 MAP3Ks (MAPK kinase) (Pitzschke *et al.*, 2009) with a total of 24 MAPK pathways (Wrzaczek *et al.*, 2001). MAP3Ks is the most structurally diverse constituent of a cascade with different motifs such as zinc fingers and G-protein binding sites (Wrzaczek *et al.*, 2001).

Handling huge amounts of sequence data is highly challenging. In the current scenario, data is generated at a much higher rate than tools can work, interpret, and simplify this data. Machine learning is the study relating to the creating of algorithms that expedite the processes of classification, pattern identification, and prediction on the basis of models obtained from existing data (Tarca *et al.*, 2007). Thus, machine learning techniques generally proceed in two main phases where the first phase involves the usage of data to let the program learn from the latent structure and relationships in this data. Once the system has been trained, in the second phase, this algorithm can now be applied to the data which needs to be analyzed (Sommer

and Gerlich, 2013). One of the earliest algorithms developed (named Perceptron) was for the identification of the initiation sites for translation in *Escherichia coli* (Stormo *et al.*, 1982). However, the use of machine learning tools to characterize large data sets in plants is a rare occurrence (Ma *et al.*, 2014). One tool developed to understand the functional associations between genes present in the same pathways is the Seed Co-Prediction Network based on microarray data solely from A. *thaliana* (Bassel *et al.,* 2011). In this study, we attempt to develop a new prediction method based on machine learning approach using Waikato Environment for Knowledge Analysis (WEKA). The following characteristics of a protein were considered while developing the models of WEKA: Amino acid composition, dipeptide method, tripeptide method, hybrid-1(amino acids + dipeptide features), and hybrid-2 (amino acids and tripeptide features). The performance of the models was validated, and a web server has been developed on the best model to predict MAPK protein and assist in automated genome annotations.

MATERIALS AND METHODS

Datasets

From the assembled and annotated data of leaf transcriptome of Chowghat Green Dwarf cultivar of coconut (SRX436961) generated in our laboratory, we extracted all MAPK proteins for this study. These proteins were used to construct datasets. For the development of positive and negative, datasets, 60 MAPK proteins were used as a positive data set, and negative dataset was created by using 40 non-MAPK proteins.

WEKA

WEKA is a popular machine learning environment developed in Java. WEKA is the open source software available at http://www.cs.waikato.ac.nz/ml/weka, which was developed at the University of Waikato, New Zealand. In this study, we compared algorithms such as Naïve Bayes, Kstar, Simple Logistic, sequential minimal optimization (SMO), and voting feature intervals (VFI) for the prediction accuracy of MAPK.

Features

Residue based method

In feature extraction, residue-based methods include amino acid, dipeptide, and tripeptide methods. Amino acids are basic constituents of proteins. Amino acid method is essential to find out the features of each amino acid in a protein sequence. In this method, amino acids are represented by a vector dimension of 20.

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

Where $P(a_i)$ represents a portion of each $(a_i)^{th}$ amino acid, denominator signifies the sum of amino acids present in the protein sequences and $N(a_i)$ gives the overall number of $(a_i)^{th}$ amino acids (Kaundal *et al.,* 2010).

The dipeptide method is used to extract the global information of each protein sequence. In this method, a dimension of each protein was fixed at a length of 400 (20*20).

$$P(a_ia_j)=\frac{Na_ia_j}{\sum_{i=1}^{20}\sum_{j=1}^{20} Na_ia_j}$$

Where $P(a_ia_j)$ is the fraction of each a_ia_j dipeptide, $N\ a_ia_j$ is the total number of a_ia_j dipeptides, and the denominator stands for the total number of dipeptides in the protein sequences provided (Kaundal *et al.*, 2010).

Tripeptide methods provide a vector dimension of 8000 (20*400)

$$P(a_ia_ja_k)=\frac{Na_ia_ja_k}{\sum_{i=1}^{20}\sum_{j=1}^{20}\sum_{k=1}^{20} Na_ia_ja_k}$$

Where $P(a_ia_ja_k)$ is the fraction of each $a_ia_ja_k$ tripeptide, $N\ a_ia_ja_k$ is the total number of $a_ia_ja_k$ tripeptide, and the denominator is the total number of tripeptides present in the provided protein sequences (Kaundal *et al.*, 2010).

Hybrid-based methods

Hybrid methods are combinations of different methods for the development of a new one. In this study, a combination of dipeptide features and amino acids of a protein sequence were utilized for the development of Hybrid-1. It has the vector dimension of 420(20*400). Hybrid-2 was developed by combining amino acids and tripeptide features of a protein sequence and has a dimension of 8020(20*800).

Machine Learning Algorithms

Naïve Bayes

In machine learning methods, Naïve Bayes is a probabilistic statistics method based on Bayesian theory which is fast, accurate, and easy to perform. Like the support vector

machine, this algorithm is widely used in text classification methods (Rennie *et al.*, 2003). In prediction tasks, this classifier is used to build a model based on training documents and then used to classify new documents (Cleary *et al.*, 1995).

KStar

This classifier is an instance-based algorithm, which predicts results based on training instances. KStar classifier is based on assumptions of classifiers such as K*, IB1, and PEBLS. This classifier uses entropy-based distance function instead of other instance-based learning methods (John *et al.*, 1995).

Simple logistic

A simple logistic algorithm generates a model that has lower parameters when compared to logistic regression. Sometimes, the accuracy of the generated model keeps increasing significantly (Landwehr *et al.,* 2005). This algorithm is very accurate, compact and shows some complexity while creating logistic regression models in the tree (Sumner *et al.,* 2005).

SMO

SMO algorithm is extremely easy to handle and implement. Compared to other algorithms, SMO is fast and has greater scaling properties (Platt *et al.*, 1998). SMO can handle large amounts of training sets because the memory quantity for training set sizes is linear. One of the advantages of this algorithm is that it splits large quadratic programs into a series of small quadratic programs. This algorithm utilizes only those possible quadratic programs, which utilize less memory while proceeding (Keerthi *et al.*, 2001; Hastie *et al.*, 1998).

VFI

VFI algorithm, per feature dimension, considers a set of feature intervals. The class of the training set is represented by a label, and the dataset is assigned as a vector of the feature value during training. The feature interval of each feature is formed by the VFI algorithm. This algorithm archives with more accuracy and is much faster (Demiröz and Güvenir, 1997).

Similarity Search

BLAST-Like Alignment Tool (BLAT) is a new alignment tool which is more accurate and faster than the other similarity searching tools/search engines. BLAT is available without charge for any type of usage and can be downloaded from http://hgdownload.cse.ucsc.edu/admin/exe/. In this study, BLAT was used for a protein similarity search against a non-redundant protein database, where the non-redundant database was created from MAPK proteins.

Evaluation Procedures

Model performance built by this method was checked using ten-fold cross-validation, independent dataset test, and leave-one-out cross-validation.

Ten-fold Cross-Validation

Cross-validation is one of the most common methods to evaluate a model. In ten-fold classification method, the holdout method is repeated a total of ten times after splitting the data into a total of ten subsets. Each time, nine subsets are merged to form a training set and one of the subsets is utilized as a test set. The average of the total number of errors across each trial is calculated. In this method, 90% of the complete data is used for training in each test.

Leave-one-out Cross-Validation (LOOCV)

It is a method, where the overall performance of the model is estimated. Before choosing the model, we need to ensure that the particular model fits all the training sets. In LOOCV, the model is repeatedly refit leaving out a single observation and then used to derive a prediction for the left-out observation. This cross-validation gives estimates of the prediction error that are more variable than other forms of cross-validation such as the 10-fold classification. One of the advantages of this method is that it has high variance, and the value of the model varies due to the usage of random sample data.

Independent Dataset Test

Testing data used for the independent dataset test is entirely different from the data for training. The dataset used for testing is selected from different databases randomly.

Parameters for Performance Evaluations

To evaluate the performance, we use several parameters described below:

i. Sensitivity

Sensitivity or Recall is the percentage of correctly predicted true positives (TPs). In this study, it is the proportion of MAPK proteins correctly predicted as MAPK proteins.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100$$

TP and true negative (TN) are true MAPK proteins and non-MAPK proteins, respectively. False positive

(FP) and false negative (FN) are incorrectly predicted MAPK proteins and non-MAPK proteins, respectively.

ii. Specificity
Specificity is the proportion of TNs that are correctly predicted as negative.

$$\text{Specificity} = \frac{\text{TN}}{\text{FP}+\text{TN}} \times 100$$

iii. Accuracy
Accuracy is the percentage of correctly predicted proteins.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100$$

iv. F-measure
F-measure or F-score is calculated based on precision and recall, where precision is the fraction of elements correctly classified as positive out of all the elements the algorithm classified as positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

Recall is the fraction of elements correctly allocated as positive out of all the positive elements.

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

The F-measure calculation is as follows:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$$

Mathew's Correlation Coefficient (MCC)

MCC is used in data mining for the measurement of quality between two algorithms and gives a balance between positive and negative classes. MCC equal to one is regarded as a perfect prediction, whereas zero is for a completely random prediction and −1 indicates perfect negative correlation.

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

Receiver Operating Characteristic (ROC) Curve

Prediction methods can have the same or a different MCC from each other. Each training sets can also be chosen so as to obtain a higher value for sensitivity and specificity. In usual circumstances, when two prediction algorithms display the same threshold value, the threshold value is adjusted so as to compare the accuracies of the two methods. Application of a non-parametric prediction performance measure ROC curve negates this necessity and is a much better technique to measure accuracies between different prediction processes (Powers *et al.*, 2007).

Development of a Web Server

A website was developed for inputting and displaying results using the best performing algorithm. The overall design of the web page was done in HTML and CSS. Back end coding was done with PHP, JAVA, and PERL. A user guide page was created with the instructions for inputting and annotating the results.

RESULTS AND DISCUSSION

Results of Independent Dataset Test

Overall performances of all developed modules were evaluated by independent dataset test. The performance results of all five feature extraction methods are shown in Table 1. While analyzing the test results, we can assume that the algorithm Naïve Bayes has 100% sensitivity with a value of 1 for MCC, for Tripep, and Hybrid 2 method. A classification is said to be ideal with the value of MCC equal to 1 and precision, specificity and sensitivity close

Table 1: Independent dataset results of CnMAPK$_{Pred}$ with five methods

Method	Algorithm	Independent data sets				
		Sn	Sp	Acc	F	MCC
Amino acid	Naive Bayes	80	90	85	94	0.70
	SMO	70	100	85	100	0.73
	KStar	80	90	85	94	0.70
	Simple Logistic	100	90	95	94	0.90
	VFI	100	90	95	94	0.90
Dipep	Naïve Bayes	64	80	90	88	0.82
	SMO	80	70	75	82	0.50
	KStar	90	70	80	82	0.61
	Simple Logistic	82	90	86	94	0.72
	VFI	100	80	90	94	0.82
Tripep	Naïve Bayes	100	100	100	100	1
	SMO	100	90	95	94	0.90
	KStar	100	96	98	97	0.96
	Simple Logistic	100	80	90	88	0.82
	VFI	90	70	80	82	0.61
Hybrid 1	Naïve Bayes	80	73	76.5	84	0.53
	SMO	60	76	68	86	0.36
	KStar	73	89	81	90	0.63
	Simple Logistic	80	90	85	94	0.70
	VFI	100	70	85	82	0.73
Hybrid 2	Naïve Bayes	100	100	100	100	1.00
	SMO	100	90	95	94	0.90
	KStar	100	90	95	94	0.90
	Simple Logistic	80	70	75	82	0.53
	VFI	100	100	100	100	1.00

SMO: Sequential minimal optimization, VFI: Voting feature intervals, MCC: Mathew's Correlation Coefficient

to 100% (Guda *et al.*, 2004). To further calculate the performance of the classifier, receiver operator curve was plotted for each feature of the Naïve Bayes algorithm.

From ROC, it is clear that the Tripep method has an area under the curve of 1 (Figure 1). The above results show that the Naïve Bayes algorithm has a precision of 100% and a feature vector of dimension 8000. Thus, Naïve Bayes algorithm proves that it can predict MAPK kinases the best when compared to other methods. Performance comparison between Naïve Bayes algorithm and the other methods is shown in Figure 2.

Results of Cross-Validation Data Test

The cross-validation results are shown in Table 2. The accuracy of 10-fold cross-validation and LOOCV with independent datasets does not match each other. Cross-validation results with a maximum precision of 99%, an MCC value of 0.9 and F-measure of 99 was obtained with the Hybrid-1 and Hybrid-2 methods. Compared to other features, these features display a large number of features dimension.

Comparison of CnMAPK$_{Pred}$ with Other Machine Learning Methods

A comparison was done between Naïve Bayes algorithm and other classifiers, namely SMO, KStar, Simple Logistic,

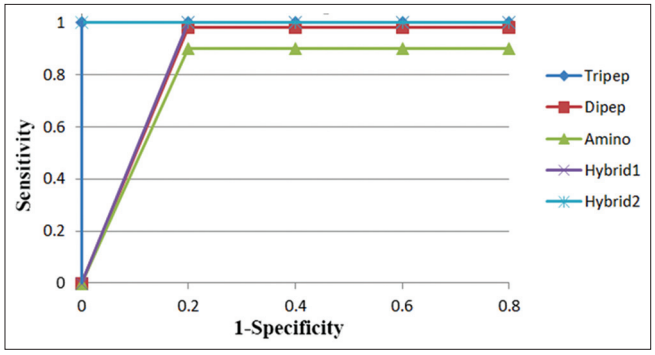

Figure 1: Receiver operating characteristic curve of Naive Bayes Algorithm of different methods for CnMAPK$_{Pred}$

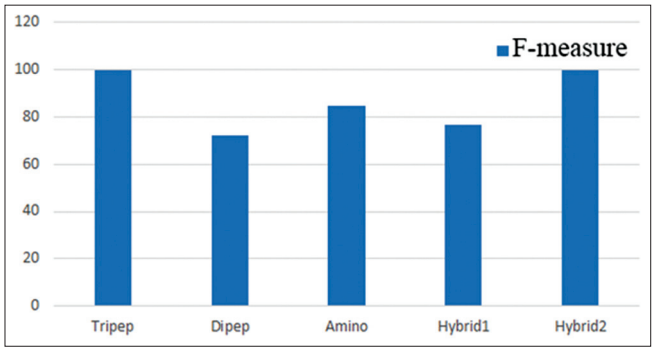

Figure 2: Comparison of accuracy of five methods with Naive Bayes classifier

and VFI. The development of models was carried out using all five feature extraction methods, and the same datasets were used for Naïve Bayes algorithm. Tripeptide method has a lesser feature vector dimension when it is compared to Hybrid-1 and Hybrid-2 methods. After a comparison of methods and algorithms, Naïve Bayes algorithm was observed to obtain an optimum classification for tripeptide method with a vector dimension of 8000. Naïve Bayes has long been rated above other algorithms in many areas especially when there are a relatively lesser number of features to predict (Cao *et al.*, 2003). It is a robust algorithm and displays a much better accuracy when compared to algorithms like the C4.5 (Friedman *et al.*, 1997). The independence assumption between features keeps probability estimate variance at a minimum. Another advantage of this algorithm is the relative comfort with which missing values are handled in the set which is neglecting them assuming that the missing data is random (Hand and Yu, 2001). In case some attribute interdependence does exist, Semi-naïve Bayes algorithms can be employed like the backward sequential elimination and the tree augment Naïve Bayes (TAN) methods (Zheng and Webb, 2005).

Sequence Similarity Search

To find the efficiency of CnMAPK$_{Pred}$, we compared it with results obtained through standalone BLAT. The comparison of results showed that there were not many valid hits, and only a low accuracy of 67.75% could be obtained (Table 3). These comparison results point out that the efficiency of tools used for sequence comparison is relatively low when compared to the use of machine learning algorithms.

Performance on Other Plants

The ability of CnMAPK$_{Pred}$ developed in this study was analyzed using three different plants namely, *A. thaliana*, *Oryza sativa,* and *Theobroma cacao* which have a total of 208 MAPK proteins. If the predictor showed any species-specific feature of protein, the performance of predictor on other plants should give results of diminished accuracy. To justify this, CnMAPK$_{Pred}$ tool was run with Naïve Bayes algorithm for the three diverse plants. The performance of the tool with regard to predictions done on the three plants is shown in Table 4. The prediction results showed an accuracy of 25%, 31%, and 33%, respectively, highlighting that CnMAPK$_{Pred}$ is indeed a species-specific tool. Species-specific tools can have more accuracy in its results when compared to other general tools because they are tailor- made for predictions on the data set of that species. They could also be used for the prediction

Table 2: Cross-validation results of $CnMAPK_{Pred}$ with five methods

Method	Algorithm	10-fold cross-validation					LOOCV				
		Sn	Sp	Acc	F	MCC	Sn	Sp	Acc	F	MCC
Amino acid	Naïve Bayes	39	40	39.5	57	−0.21	62	23	42.5	37	−0.16
	SMO	50	48	49	64	−0.02	50	60	55	75	0.10
	KStar	80	60	70	75	0.41	70	60	65	75	0.30
	Simple logistic	43	50	46.5	66	−0.07	97	80	88.5	88	0.78
	VFI	51	48	49.5	64	−0.05	50	60	55	75	0.10
Dipep	Naïve Bayes	65	86	75.5	92	0.52	78	80	79	88	0.61
	SMO	99	98	98.5	98	0.97	90	98	94	98	0.88
	KStar	80	74	77	85	0.42	86	87	86.5	93	0.73
	Simple logistic	92	68	80	80	0.62	80	70	75	82	0.50
	VFI	80	76	78	86	0.56	70	56	63	71	0.26
Tripep	Naïve Bayes	56	61	58.5	75	0.17	46	63	54.5	77	0.16
	SMO	99	99	99	99	0.98	90	98	94	98	0.88
	KStar	100	96	98	97	0.96	100	90	95	94	0.90
	Simple logistic	97	67	82	75	0.67	67	82	74.5	90	0.50
	VFI	96	92	94	96	0.88	76	83	79.5	90	0.59
Hybrid-1	Naïve Bayes	100	90	95	99	0.90	96	833	89.5	90	0.80
	SMO	60	52	56	68	0.12	47	63	55	77	0.10
	KStar	50	50	50	66	0.00	52	53	52.5	69	0.05
	Simple logistic	96	50	73	66	0.52	90	45	67.5	62	0.39
	VFI	93	80	86.5	88	0.74	80	84	82	91	0.64
Hybrid-2	Naïve Bayes	100	99	99.5	99	0.99	100	90	95	94	0.90
	SMO	68	64	66	78	0.32	60	90	70	94	0.52
	KStar	50	50	50	66	0.0	56	50	53	66	0.06
	Simple logistic	92	60	76	75	0.55	58	70	64	82	0.28
	VFI	100	90	95	94	0.90	90	96	93	97	0.86

SMO: Sequential minimal optimization, VFI: Voting feature intervals, MCC: Mathew's correlation coefficient, LOOCV: Leave-one-out cross-validation

Table 3: Prediction results of $CnMAPK_{Pred}$ domain predictor with similarity search (10-fold validation)

Number of sequences given	Correctly predicted	Accuracy
50	35	70
50	40	80
50	37	74
50	25	50
50	30	60
50	27	54
50	39	78
50	38	76
		67.75

Table 4: Prediction results of $CnMAPK_{Pred}$ with three plants using Naïve Bayes algorithm

Plants	Number of sequences given	Correctly predicted	Accuracy
Arabidopsis	108	28	25
Oryza sativa	38	12	31
Eragrostis tef	62	21	33

of species-specific features like the AtSubP (*Arabidopsis* subcellular localization predictor) tool which predicts the amino acid composition and sorting signals in *A. thaliana* (Kaundal *et al.*, 2010). The only drawback of using a tool that is species specific is the availability of a smaller dataset for training.

Description of Web Server

A web server has been implemented based on the best performing module in this study named "$CnMAPK_{Pred}$", which is available at http://210.212.229.52/Cnmapk. The server runs on SUN server X2200 M2 under the windows environment. This environment allows users to submit their sequences in standard FASTA format or by uploading a file (Figure 3). The server displays a user-friendly result page within a few seconds in a tabular format.

CONCLUSION

A major challenge in proteomics is the identification and annotation of all proteins on a genomic scale. There are a number of tools and programs developed for understanding the functional importance of proteins. However, a drawback of these tools is that they are not specific protein predicting programs. In this study, we implement a new tool called $CnMAPK_{Pred}$ which allows a user to predict MAPK proteins in coconut. This tool delicately shows satisfactory prediction results, and the selected algorithm is suited for this particular protein. The overall performance evaluation indicates that this project will significantly contribute in proteomics related studies for a deep level study of functionally important proteins.

REFERENCES

Bassel GW, Glaab E, Marquez J, Holdsworth MJ, Bacardit J. Functional network construction in *Arabidopsis* using rule-

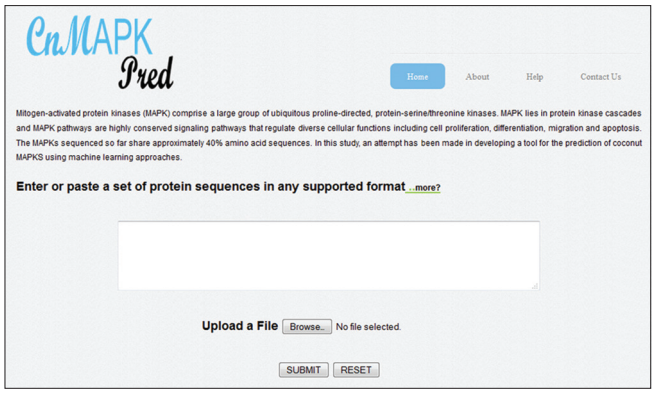

Figure 3: User input screen of $CnMAPK_{Pred}$

based machine learning on large-scale data sets. Plant Cell 2011;23:3101-16.

Cao J, Panetta R, Yue S, Steyaert A, Young-Bellido M, Ahmad S. A naive bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. Bioinformatics 2003;19:234-240.

Cleary JG, Trigg LE. K*: An instance-based learner using an entropic distance measure. Proceedings of the 12th International Conference on Machine Learning 1995;5:108-14.

Cristina MS, Petersen M, Mundy J. Mitogen-activated protein kinase signaling in plants. Ann Rev Plant Biol 2010;61:621-49.

Demiröz G, Güvenir HA. Classification by voting feature intervals. In: Van Someren M, Widmer G, editors. Machine Learning: ECML-97. Vol. 1224 of the Series Lecture Notes in Computer Science. Germany: Springer Berlin Heidelberg; 1997. p85-92.

Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997;29:131-63.

Garcia AV, Al-Yousif M, Hirt H. Role of AGC kinases in plant growth and stress responses. Cell Mol Life Sci 2012;69:3259-67.

Guda C, Fahy E, Subramaniam S. MITOPRED: A genome-scale method for prediction of nucleus-encoded mitochondrial proteins. Bioinformatics 2004;20:1785-94.

Hand DJ, Yu K. Idiot's Bayes—not so stupid after all? Int Statist Rev 2001;69:385-98.

Hanks SK, Hunter T. Protein kinases 6. The eukaryotic protein kinase superfamily: Kinase (catalytic) domain structure and classification. FASEB J 1995;9:576-96.

Hastie T, Tibshirani R. Classification by pairwise coupling. Ann Statist 1998;26:451-71.

John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. In: Philippe B, Steve H, editors. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 1995. p. 338-345.

Kaundal R, Saini R, Zhao PX. Combining machine learning and homology-based approaches to accurately predict subcellular localization in *Arabidopsis*. Plant Physiol 2010;154:36-54.

Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KR. Improvements to platt's SMO algorithm for SVM classifier design. Neural Comp 2001;13:637-49.

Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. Trends in Plant Sci 2014;19:798-808.

Pitzschke A, Schikora A, Hirt H. MAPK cascade signalling networks in plant defence. Curr Opin Plant Biol 2009;12:421-6.

Platt J. Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods: Support Vector Learning. Cambridge, Massachusetts, London and England: The MIT Press. p373.

Powers DM. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol 2011;2:37-63.

Rennie JD, Shih L, Teevan J, Karger DR. Tackling the poor assumptions of Naive Bayes text classifiers. ICML 2003;3:616-23.

Schaeffer HJ, Weber MJ. Mitogen-activated protein kinases: Specific messages from ubiquitous messengers. Mol Cell Biol 1999;19:2435-2444.

Sinha AK, Jaggi M, Raghuram B, Tuteja N. Mitogen-activated protein kinase signaling in plants under abiotic stress. Plant Signal Behav 2011;6:196-203.

Sommer C, Gerlich DW. Machine learning in cell biology–Teaching computers to recognize phenotypes. J Cell Sci 2013;126:5529-39.

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. Nucl Acids Res 1982;10:2997-3011.

Sumner M, Frank E, Hall M. Speeding up logistic model tree induction. In: Alipio MJ, Luis T, Pavel B, Rui C, Joao G, editors. Knowledge Discovery in Databases: PKDD 2005. Berlin-Heidelberg: Springer; 2005. p. 675-83.

Taj G, Agarwal P, Grant M, Kumar A. MAPK machinery in plants: Recognition and response to different stresses through multiple signal transduction pathways. Plant Signal Behav 2010;5:1370-8.

Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S. Machine learning and its applications to biology. PLoS Comp Biol 2007;3:e116.

Wrzaczek M, Hirt H. Plant MAP kinase pathways: How many and what for? Biol Cell 2001;93:81-7.

Zheng F, Webb GI. A comparative study of semi-Naive Bayes methods in classification learning. In: Simeon J, editors. Proceedings of the Fourth Australasian Data Mining Conference, Aus DM05. 2005. p. 141-56.