Indian Journal of Science and Technology, *Vol 9(12), DOI: 10.17485/ijst/2016/v9i12/70538, March 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Machine Learning Approach for Prediction of Domains of DELLA Proteins, a Key Component of Gibberellic Acid Signaling in Plants

V. Akhil[1], V. Amal[1], N. Hemalatha[2]* and M. K. Rajesh

[1]ICAR-Central Plantation Crops Research Institute, Kasaragod - 671124, Kerala, India
[2]AIMIT, St. Aloysius College, Mangalore - 575002, Karnataka, India; hemasreen71aimit@gmail.com

Abstract

Background/Objective: For the annotation of large scale proteins, generally computational methods or tools are used. One of the drawbacks of these annotation tools is that they are not specific protein prediction programs. **Methods/Analysis:** In this study, we implement a machine-learning algorithm for fast and accurate prediction of DELLA proteins. **Findings:** We developed various modules by using conserved protein domains in DELLA proteins. To evaluate the modules, classifiers like sequential minimum optimization, J48 decision tree, AD tree and logistic algorithms were used. By analyzing the results obtained from independent data set and cross-validation tests, maximum accuracy was achieved by logistic algorithm. The developed tool was tested with various inputs and it showed that the algorithm developed in the study would be helpful in predicting plant DELLA domains. **Applications:** This tool will significantly contribute to functional genome annotation and development of predictors.

Keywords: Algorithms, Coconut, DELLA, Domains, Machine Learning, Prediction

1. Introduction

Gibberellins or Gibberellic Acid (GAs) comprise a group of plant hormones that perform plant growth regulator functions and influence a repertoire of developmental processes in higher plants ranging from dormancy, induction of enzymes, leaf and fruit senescence, stem elongation, flowering, sex expression to germination[1]. DELLA domain-containing proteins have been implicated as negative regulators of GA. The DELLA proteins are localized in the nucleus and restrain plant growth by binding to and inactivating transcription related proteins in the absence of GA[1,2]. The repressor activity of these proteins is relieved by their GA-dependent degradation[3,4]. DELLA repressors mediate degradation of DELLA proteins chiefly through post translational modifications mechanisms[5–8].

The DELLA proteins are classified as members of the GRAS family. However, in addition to consensus motifs found in members of GRAS family proteins, the GA-signal-related DELLA proteins have also been reported to contain additional novel motifs in their amino-terminal region- these are known as 'DELLA domains' which are not found in other members of GRAS family proteins[9]. Two domains of the DELLA repressors, which have been observed to be highly conserved in members like SLENDER RICE1 (SLR1) from rice and GA-INSENSITIVE (GAI) and REPRESSOR-of-*ga1-3* (RGA) from *Arabidopsis*, are the DELLA and VHYNP domains[10].

DELLA proteins show significant protein sequence conservation within species and also between species because of which they tend to have highly similar, if not identical function. The mutations in the DELLA domain are associated with the dwarfing alleles which are present in DELLA repressors of many plants including wheat, maize, barley etc[11]. In *Arabidopsis*, deletion of 17 amino

*Author for correspondence

acids in the conserved DELLA domain of the DELLA repressors results in GA-insensitive, late-flowering and dark-green dwarf plants[12].

Mutations in the DELLA domain have been reported in DELLA repressor gain-of-function variants[13]. Moreover, the deficiency of this domain impairs the GA-dependent interaction of DELLA repressors with GA receptors and this subsequently prevents their proteolysis mediated by the ubiquitin–proteasome pathway[14,15]. It has also been reported that the DELLA domain act as receiver domains for GA receptors which are activated[12].

Experimental and computational methods are two methods for analysis of large scale sequence data. Because of the exponential growth in sequence data, computational methods are much efficient for fast and accurate prediction of large sequences. A computational approach to predict a particular protein from sequence data would typically consist of identifying domains and sequence motifs within the particular protein and creating sequence specificity-based algorithms to predict the protein. We have developed a new prediction method for the prediction of DELLA domains, which is based on WEKA (Waikato Environment for Knowledge Analysis) machine learning techniques in this study. Development of WEKA modules relied on the following protein features: composition of amino acids, dipeptide method, tripeptide method, N and C-terminal methods and hybrid-based methods. We have also validated the performance of the model by using various validation techniques and a web server has also been created on the best model to predict DELLA domains and can assist in the automated genome annotations.

2. Methods

2.1 Dataset

In this study, the data set consisting of 60 DELLA proteins from diverse plant species viz., *Phaseolus vulgaris, Helianthus annuus, Triticum aestivum, Vitis vinifera, Zea mays, Arabidopsis thaliana, Eragrostis tef, Olea europaea, Sorghum bicolor, Ricinus communis, Oryza sativa* and *Ocimum basilicum*, were obtained from NCBI (http://www.ncbi.nlm.nih.gov/). These proteins were confirmed to be DELLA family through PROSITE database (http://prosite.expasy.org/) and PFAM database (http://pfam.xfam.org/). These conformed protein dataset were selected for the creation of the training data set. Non-DELLA proteins constituted the negative data set. The test dataset contained both DELLA and non-DELLA proteins. The DELLA proteins in the training and test datasets were totally different.

2.2 Protein Feature Extraction

2.2.1 Residue Method

Amino acid composition was based on the frequencies of occurrence of various amino acids present in a protein. A vector of dimension 20 was utilized to symbolize the 20 natural amino acids. To summarize the global information present in a protein sequence, computation of dipeptide method was carried out, which comprises of a fixed pattern length of 400 (20*20) dimension. Tripeptide method provides an 8000 (20*400) dimensional feature vector for the protein sequences. The following equations were utilized to compute the fraction of each amino acid, dipeptide and tripeptide methods respectively:

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

$$P(aiaj) = \frac{Na_i a_j}{\sum_{i=1}^{20}\sum_{j=1}^{20} Na_i a_j} \quad (2)$$

$$P(aiajak) = \frac{Na_i a_j a_k}{\sum_{i=1}^{20}\sum_{j=1}^{20}\sum_{k=1}^{20} Na_i a_j a_k} \quad (3)$$

where P (Ai) stands for the fraction of each $(Ai)^{th}$ amino acid, N (Ai) symbolizes the total number of $(Ai)^{th}$ amino acids and denominator denotes the total number of amino acids presents in the specified protein sequences.

2.2.2 Method based on Composition of Terminal Residue

Proteins generally comprise of terminal residue compositions of two types: The N-terminal and C-terminal, which are located on surface of protein. In this method, protein sequences were separated into a number of overlapping fragments and calculation of amino acid composition of each fragment was carried out separately using Equation (1). The N-terminal and C-terminal sequences are important in plant cells as they possess signals responsible for targeting of proteins to diverse sub-cellular localizations. For both the N and C terminals in this study, a residue length of 30 was considered.

2.2.3 Hybrid-based Methods

Combination of diverse methods was carried out to obtain novel ones in hybrid-based approaches. Development of Hybrid-1 method was carried out by a combination of Equation (1) and Equation (2) which represent amino acids and dipeptide features of a protein sequence. Hybrid-2 method was developed by merging Equation (1) and Equation (3), i.e., amino acid and tripeptide features of a protein sequence. The input feature vector pattern possessed a dimension of 420 (20 + 400) and 8020 (20 + 800) respectively.

2.3 The Machine Learning Algorithms

2.3.1 Sequential Minimum Optimization (SMO)

SMO constitutes a theoretically simple, effortlessly executable and rapid machine learning algorithm which possesses enhanced scalable properties for complex problems in comparison to standard and simple training algorithms, the latter requiring additional time while resolving vast quadratic programming optimization problems. When the same quadratic programming is given to SMO, the SMO breaks down the complex problem into a sequence of smaller problems which can be deciphered faster analytically and the use of taxing numerical QP optimization is avoided. A huge amount of training set can easily be handled by SMO as the quantum of memory requirement for SMO is linear to the size of the training dataset.

2.3.2 J48 Decision Tree

J48 constitutes a predictive machine-learning model wherein the values of dependent or the target variable are predicted based on the value of the independent variable. The algorithm uses the characteristic value in the provided data to align values to dependent variable. Various attributes are represented by the internal nodes of a decision tree, whereas the branches in between the nodes indicate the probable values for these attributes. The final value or quantity of the dependent variable is shown in the terminal node.

2.3.3 AD Tree

AD Tree, which stands for alternating decision tree, is a classification technique used in machine learning and has association with boosting. Boosting, a machine learning algorithm, lessens biased-ness in supervised learning. An AD tree comprises of two nodes: (i) A decision node (which details a predicate condition) and, (ii) A prediction node (which includes a single number). An AD tree always possesses prediction node at both roots and leaves.

2.3.4 Logistic Algorithm

Logistic Algorithm is a popular method to model a binary data. This algorithm predicts the maximum probability of occurrence by correlating data to logistic function. Logistic Algorithm is commonly used in several predictor variables. This algorithm provides a multi-dimensional feature space and will give high prediction accuracy. This algorithm focuses to build linear regression models.

2.4 Evaluation Procedures

Three evaluation procedures, viz., ten-fold cross validation, independent dataset test and leave-one-out cross validation were used to gauge the performance of the modules built.

We have utilized use several parameters to evaluate the performance of modules, which are described below:

1. Sensitivity or percentage of coverage is the percentage of DELLA protein accurately predicted as DELLA proteins.

$$Sensitivity = \frac{TP}{TP+FN} x100$$

2. Specificity or percentage of coverage of non-DELLA protein is the percentage of non-DELLA proteins accurately predicted as non-DELLA proteins.

$$Specificity = \frac{TN}{FP+TN} X100$$

3. Accuracy is the percentage of accurately predicted DELLA proteins.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} X100$$

4. F-measure or F-score is considered based on precision and recall, were precision is described as the portion of elements accurately classified as positive of all the elements classified as positive by the algorithm.

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

Recall denotes the portion of elements accurately classified as positive out of all positive elements.

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

The F measure is calculated as follows:

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

5. Mathew Correlation Coefficient (MCC) is judged to be the most vigorous parameter of a class prediction technique. A MCC which equals to one is deemed as a prefect prediction, whereas a MCC which equals to zero is deemed a completely random prediction. A-1 MCC indicates perfect negative correlation.

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

Where TP (True Positive) and TN (True Negative) are accurately predicted DELLA protein (positive) and non-DELLA protein (negative), respectively and FP (False Positive) and FN (False Negative) are incorrectly predicted DELLA and non-DELLA proteins, respectively.

The measures illustrated here have a weakness that they provide performances at a specified threshold. Receiver Operating Characteristic (ROC) is a recognized threshold independent parameter plotted involving True Positive Propagation (TP/TP+FP) and False Positive Propagation (FP+FP+TN). Area Under Curve (AUC) provides a single measure to assess the performance of method. An AUC of 1 indicates that the prediction is perfect.

2.5 Similarity Search Methods

Position-Specific Iterative Basic Local Alignment Tool (PSI-BLAST) is protein sequence profile similarity search engine. This similarity search method builds profiles with the help of BLASTP programme and detects sequence similarity above the given threshold. In this study, a database of 80 DELLA proteins was created to utilize it for standalone sequence similarity searches. This method is capable of finding remote homologies sequence.

2.6 Development of a Web Server

A user friendly web server was developed for inputting and displaying results. Client side coding was developed in HTML and CSS. The server side coding was done with PHP, JAVA and PERL. A user interface with two input fields was developed enabling the input of both amino acid and nucleotide sequence.

3. Results and Discussion

3.1 Evaluation of the Modules

3.1.1 Results of Independent Data Set Test

The performance of all seven feature extraction methods used in prediction of DELLA proteins using independent dataset test is given in Table 1.

Table 1. Independent dataset test results with seven feature extraction methods

Method	Algorithm	Independent dataset test				
		Sn	Sp	Acc	F	MCC
Amino acid	ADtree	100	90	95	92	0.95
	SMO	100	98	99	86	0.90
	Logistic	100	100	100	100	1.00
	J48	100	90	95	100	0.90
Dipep	ADtree	100	80	90	100	0.85
	SMO	100	90	95	86	0.92
	Logistic	100	100	100	100	1.00
	J48	100	90	95	95	0.90
Tripep	ADtree	80	90	85	95	0.90
	SMO	70	100	85	95	0.90
	Logistic	100	80	100	91	0.90
	J48	100	90	95	100	1
N-Terminal	ADtree	100	100	100	83	1
	SMO	80	100	90	91	0.82
	Logistic	90	100	95	95	0.90
	J48	70	100	85	86	0.74
C-Terminal	ADtree	100	80	90	95	0.90
	SMO	100	90	95	80	0.90
	Logistic	100	100	100	100	1.00
	J48	100	97	99	0.9	0.90
Hybrid 1	ADtree	70	90	80	95	0.86
	SMO	80	80	80	95	0.92
	Logistic	100	90	95	90	1.00
	J48	70	100	85	86	0.76
Hybrid 2	ADtree	90	100	95	87	0.82
	SMO	100	80	90	95	1.00
	Logistic	100	100	100	95	1.00
	J48	100	70	85	86	0.73

Using logistic algorithm, a sensitivity of 100%, with a high confidence of MCC 1 and F-measure of 100, could be achieved for amino acid, Hybrid-1 and Hybrid-2 methods. In statistical method, sensitivity and specificity are used to measure the performance of classifiers. Matthew's correlation coefficient provides quality of classifications. If the value for MCC is 1 and sensitivity, specificity and precision are close to 100%, the classification is said to be ideal. The F1-measure computes the test's accuracy. Both precision and recall are used to calculate the F1-measure and a value of 1 is the best F1- score. Receiver Operator Curve (ROC) is another method to calculate the performance of a classifier. Here in DELLA domain predictor, ROC was constructed for all methods of Logistic Algorithm (Figure 1).

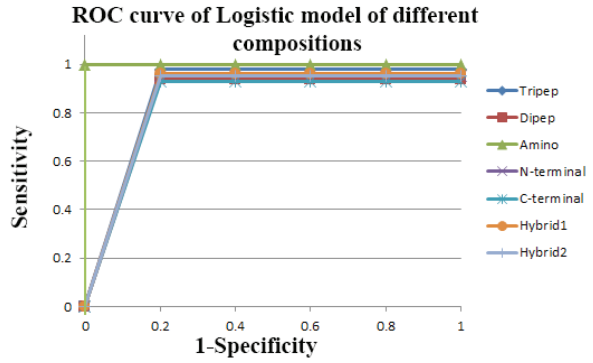

Figure 1. ROC curve of Logistic Algorithm of different methods for DELLA domain predictor.

From the plot, it could be visualized that the amino acid technique, used for prediction of DELLA domains, possessed an AUC of 1. In addition, this technique had 100% precision rate. Thus, it could be concluded that the amino acid method, which has 100% precision rate with fewer features, is the best technique under Logistic Algorithm for prediction of DELLA proteins. A comparison of the performance of seven methods with Logistic Algorithm is provided in Figure 2.

3.1.2 Results of Cross Validation Data Test

Table 2 shows the cross-validation result of DELLA domain predictor with the seven feature extraction methods. Even after both the leave one out and 10-fold cross validations were perfomed, the overall accuracy could not equal the results of independent dataset test. In the case of cross-validation results, a maximum precision rate of 99 %, with MCC of 0.9 and F-measure of 99 for Hybrid-1 and Hybrid-2 method, was obtained with Logistic Algorithm. Sensitivity and specificity also showed 99%.

3.2 Comparison of DELLA Domain Predictor with Other Machine Learning Algorithms

Comparison of the performance of DELLA domain predictor developed using Logistic Algorithm with other classifiers namely AD tree, J48 and SMO, was also carried out in this study. Prediction of the modules was carried out using all the seven feature extraction methods and the same datasets as used for Logistic Algorithm. Amino

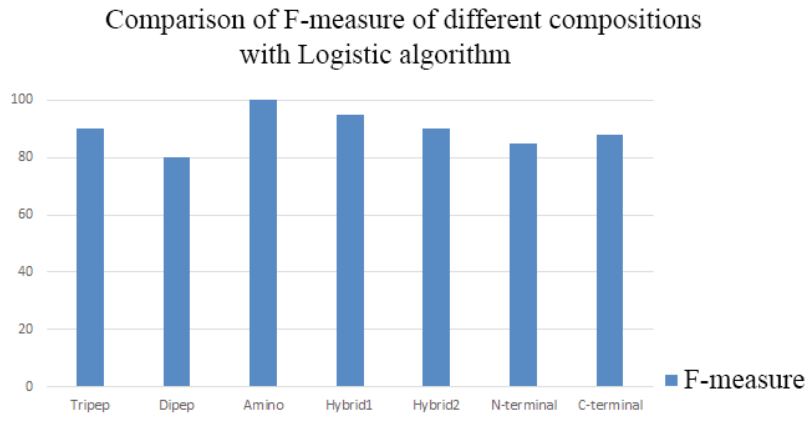

Figure 2. Comparison of accuracy of seven methods with Logistic Classifier.

Table 2. Cross validation results with seven feature extraction methods

Method	Algorithm	10-fold cross validation					LOO Cross validation				
		Sn	Sp	Acc	F	MCC	Sn	Sp	Acc	F	MCC
Amino acid	ADtree	50	50	50	78	0.55	98	97	97.5	99	0.96
	SMO	64	45	55	66	-0.43	56	28	42	65	-0.18
	Logistic	99	100	99.5	99	0.99	100	99	99.5	99	0.97
	J48	99	99	99	100	0.99	100	99	99.5	99	0.97
Dipep	ADtree	80	87	83.5	68	0.55	86	84	85	87	0.76
	SMO	98	64	81	80	0.60	82	75	78.5	80	0.68
	Logistic	100	90	95	98	0.98	89	56	72.5	87	0.78
	J48	52	0	26	67	-0.13	51	0	25.5	68	-0.13
Tripep	ADtree	64	80	72	81	0.70	87	82	84.5	65	0.64
	SMO	80	76	78	79	0.60	64	54	59	80	0.78
	Logistic	100	58	79	68	0.60	84	64	74	89	0.60
	J48	100	97	98.5	99	0.97	100	97	98.5	99	0.97
N-Terminal	ADtree	80	75	77.5	81	0.80	98	98	98	99	0.96
	SMO	46	53	49.5	62	0.50	74	65	69.5	66	-0.12
	Logistic	62	74	68	65	0.74	87	65	76	89	0.84
	J48	80	74	77	78	0.55	82	74	78	79	0.56
C-Terminal	ADtree	65	8	36.5	60	-0.28	87	84	85.5	86	0.78
	SMO	93	84	88.5	86	0.78	93	84	88.5	87	0.76
	Logistic	100	99	99.5	99	0.80	93	87	90	87	0.75
	J48	99	100	99.5	98	0.96	100	96	98	98	0.96
Hybrid 1	ADtree	63	11	37	80	0.06	85	83	84	86	0.70
	SMO	92	84	89	74	0.56	89	45	67	58	0.66
	Logistic	99	99	99	90	0.99	97	100	98.5	98	0.99
	J48	64	9	36.5	66	-0.26	96	98	97	98	0.98
Hybrid 2	ADtree	56	68	62	84	-0.15	100	97	98.5	98	0.97
	SMO	96	84	90	78	0.79	100	97	98.5	98	0.97
	Logistic	100	99	99.5	99	0.99	99	99	99	99	0.99
	J48	100	99	99.5	99	0.99	100	99	99.5	99	0.97

acids methods have less feature vector of lesser dimension when compared to other methods. After the comparison of methods and algorithms, Logistic algorithm, we obtained the most favourable classification for amino acid method, which possessed a lower vector of dimension 20.

3.3 Sequence Similarity Search

PSI-BLAST is a sequence similarity search tool, which uses query as protein sequence to search against non-redundant database of all protein sequences. This similarity tool is run based on profiles which are created by combining all protein sequences based on significant features. To find the efficiently of DELLA domain predictor, we conducted similarity search and the results demonstrate that there were lesser hits which were significant and an accuracy of only 51.5% could be obtained (Table 3). This result shows the lower reliability of sequence similarity tools on comparison to modules based on machine learning algorithms.

3.4 Description of Web Server

A dynamic web server 'DELLA domain predictor' was executed based on all modules in this study. The server

runs on SUN server X2200 M2 under windows environment. The user friendly environment allows user to submit their sequences either in the basic FASTA format or uploading of sequences as a file (Figure 3).

The server provides to display a user-friendly result page within few seconds in a tabular format. The overall architecture is shown in Figure 4.

The algorithm of DELLA domain predictor has been shown below.

Table 3. Prediction results of DELLA domain predictor with similarity search (10 fold cross validation)

Test	No. of sequences given	Correctly predicted	Accuracy
1	25	11	44
2	25	13	52
3	25	14	56
4	25	13	52
5	25	13	52
6	25	13	52
7	25	13	52
8	25	13	52
			51.5

Step 1: Start.
Step 2: Input training datasets.
Step 3: Feature extraction from the training sets using different methods.
Step 4: Create model.
Step 5: Validate test data set using above model.
Step 6: Calculate statistical results and evaluate the machine learning method.
Step 7: Repeat 4 to 6 using other machine learning algorithms.
Step 8: Find the machine learning method with best result and save model.
Step 9: Predict the user input query with best model.
Step10: Stop.

4. Conclusion

A repertoire of tools and resources are being developed for understanding functional characteristics and importance of proteins. A major drawback with these annotations is lack of accurate protein prediction programs. We present, through this study, a new technique for prediction of DELLA domains from plants, which was implemented in WEKA environment. Comparisons of various machine learning approaches were also undertaken. The enhanced accuracy for validation tests reveal that DELLA domain

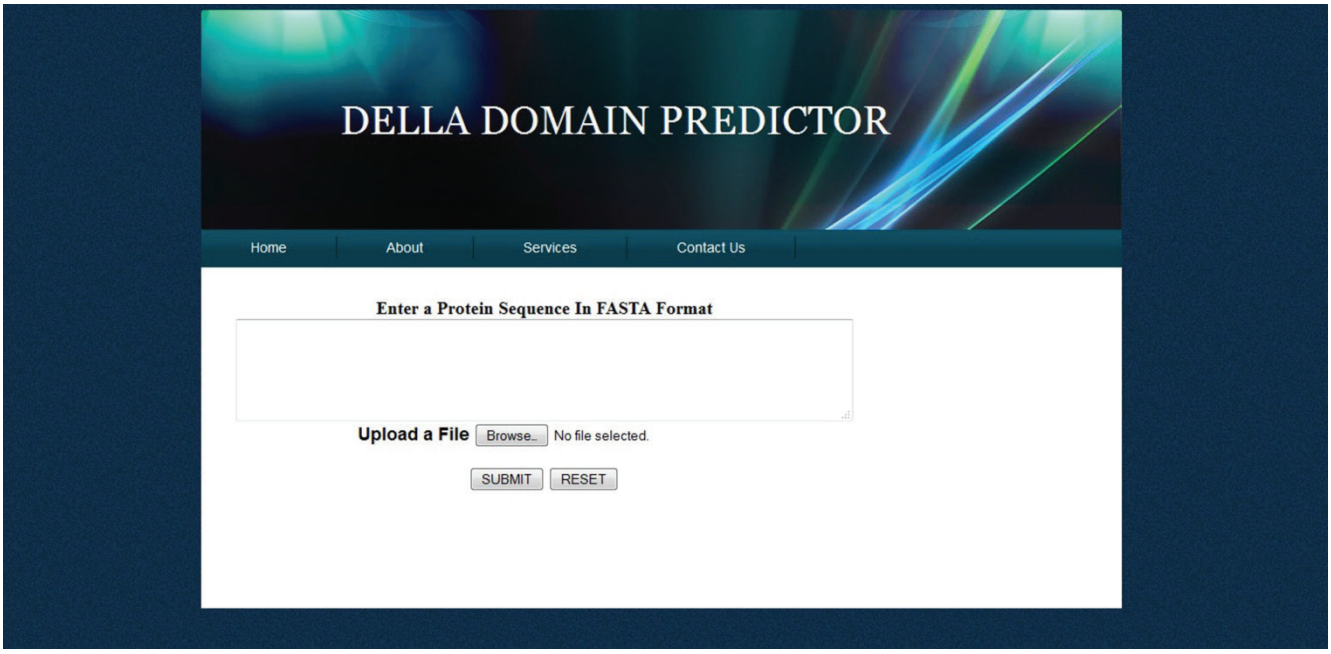

Figure 3. User input screen of DELLA domain predictor.

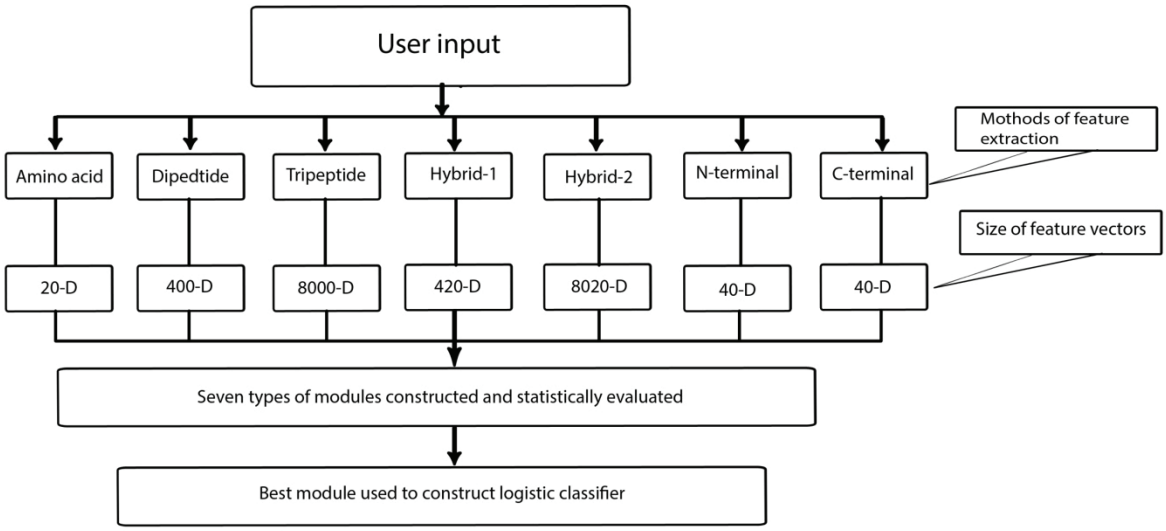

Figure 4. Over all architecture of DELLA domain predictor.

predictor will significantly contribute to genome annotation projects and development of domain prediction tools

5. Acknowledgements

The authors acknowledge the Department of Biotechnology (Sub-Distributed Information Centre), Government of India, for funding this research.

6. References

1. Daviere JM, Achard P. Gibberellin signaling in plants. Development. 2013 Mar; 140(6):1147–51.
2. Hauvermale AL, Ariizumi T, Steber CM. Gibberellin signaling: A theme and variations on DELLA repression. Plant Physiol. 2012 Sep; 160(1):83–92.
3. Silverstone AL, Jung HS, Dill A, Kawaide H, Kamiya Y, Sun TP. Repressing a repressor: Gibberellin-induced rapid reduction of the RGA protein in Arabidopsis. Plant Cell. 2001 Jul; 13(7):1555–65.
4. Dill A, Jung HS, Sun TP. The DELLA motif is essential for gibberellin-induced degradation of RGA. Proc Natl Acad Sci USA. 2001 Nov; 98:14162–7.
5. Sasaki A, Itoh H, Gomi K, Ueguchi-Tanaka M, Ishiyama K, Kobayashi M, Jeong DH, An G, Kitano H, Ashikari M, Matsuoka M. Accumulation of phosphorylated repressor for gibberellin signaling in an F-box mutant. Science. 2003 Mar; 299(5614):1896–8.
6. Gomi K, Sasaki A, Itoh H, Ueguchi-Tanaka M, Ashikari M, Kitano H, Matsuoka M. GID2, an F-box subunit of the SCF E3 complex, specifically interacts with phosphorylated SLR1 protein and regulates the gibberellin-dependent degradation of SLR1 in rice. Plant J. 2004 Feb; 37(4):626–34.
7. Itoh H, Sasaki A, Ueguchi-Tanaka M, Ishiyama K, Kobayashi M, Hasegawa Y, Minami E, Ashikari M, Matsuoka M. Dissection of the phosphorylation of rice DELLA protein, SLENDER RICE1. Plant Cell Physiol. 2005 Aug; 46(8):1392–9.
8. Shimada A, Ueguchi-Tanaka M, Sakamoto T, Fujioka S, Takatsuto S, Yoshida S, Sazuka T, Ashikari M, Matsuoka M. The rice SPINDLY gene functions as a negative regulator of gibberellin signaling by controlling the suppressive function of the DELLA protein, SLR1, and modulating brassinosteroid synthesis. Plant J. 2006 Nov; 48(3):390–402.
9. Gomi K, Matsuoka M. Gibberellin signalling pathway. Current Opinion in Plant Biology. 2003 Oct; 6(5):489–93.
10. Tanaka MU, Hirano K, Hasegawa Y, Kitano H, Matsuoka M. Release of the repressive activity of rice DELLA protein SLR1 by gibberellin does not require SLR1 degradation in the gid2 Mutant. Plant Cell. 2008 Sep; 20(9):2437–46.
11. Sun T. Gibberellin-GID1-DELLA: A pivotal regulatory module for plant growth and development. Plant Physiol. 2010 Oct; 154(2):567–70.
12. Willige BC, Ghosh S, Nill C, Zourelidou M, Dohmann EMN, Maier A, Schwechheimer C. The DELLA domain of GA INSENSITIVE mediates the interaction with the GA INSENSITIVE DWARF1A gibberellin receptor of Arabidopsis. Plant Cell. 2007 Apr; 19(4):1209–20.
13. Chandler PM, Marion-Poll A, Ellis M, Gubler F. Mutants at the Slender1 locus of barley cv. Himalaya: Molecular and

physiological characterization. Plant Physiol. 2002 May; 129(1):181–90.

14. Griffiths J, Murase K, Rieu I, Zentella R, Zhang ZL, Powers SJ, Gong F, Phillips AL, Hedden P, Sun TP, Thomas SG. Genetic characterization and functional analysis of the GID1 gibberellin receptors in Arabidopsis. Plant Cell. 2006 Dec; 18(12):3399–414.
15. Iuchi S, Suzuki H, Kim YC, Iuchi A, Kuromori T, Ueguchi-Tanaka M, Asami T, Yamaguchi I, Matsuoka M, Kobayashi M, Nakajima M. Multiple loss-of-function of Arabidopsis gibberellin receptor AtGID1 completely shuts down a gibberellin signal. Plant J. 2007 Jun; 50(6):958–66.