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. 3, Special Issue 7, October 2015**

# Computational Prediction of Magnesium Transporter Proteins in Rice (*Oryzasativa*)

Hemalatha N, [1]Praveen, Rajesh M.K[2], Eliza D.M[1]

Dept of Bioinformatics, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India[1]

Central Plantation Crops Research Institute, Kasargod, India [2]

**ABSTRACT***:* Magnesium (Mg) is essential in production of chlorophyll and photosynthesis in plants. Also, it plays an important role as an ion for signaling, both activating and mediating numerous bio-chemical reactions. In this study, we have developed algorithms for computational prediction of Mg transporter proteins in rice. The classifiers were developed using machine learning algorithms available in WEKA using diverse protein sequence features. Positive and negative dataset for training the classifiers included magnesium transporter protein sequences of *Oryza sativa* (rice) and non-magnesium transporter protein sequences respectively, which were obtained from NCBI protein database. To evaluate the prediction performance of the algorithms, leave one-out cross-validation and independent dataset validation were carried out. Using amino acid composition feature with J48 classifier, we obtained best measures of performance. The above best performing module from present investigation was implemented for computational prediction of Magnesium transporter proteins.

**KEYWORDS:** Machine Learning; Magnesium Transporter; Rice; WEKA;

## I. INTRODUCTION

Rice, *Oryzasativa L*., has a great impact on food security and human nutrition and is a source of staple food. It is a plant species which belongs to the grass family and has 12 chromosomes with a genome size of 430 Mb. Potential yields in rice production are hardly accomplished as a result of the effects of abiotic and biotic stresses. Abiotic stresses impair the growth and productivity of rice-based farming systems. Current day problems like infertile soils, poor water availability, global climate change, urbanization, etc. have intensified the problems. With rapid rise in human population and declining natural resource base, rice production continuously faces the challenge of keeping pace as two of the critical resources being land and water [1]. These situations have generated an escalated amount of attention amidst scientists towards research studies related to tolerance to abiotic stress. In order to cope with these extreme situations, developing varieties by integrating stress tolerance properties into high yielding rice varieties is found to be a highly successful approach. Modern day improvement in molecular and genomic technologies hasled to better understanding of the underlying genetic mechanisms that control the abiotic stress response in plants. Much progress has been achieved in the direction of finding potential stress related genes that are useful in improving the plants tolerance to abiotic stress.

Magnesium exists essentially as $Mg^{2+}$ and is a vital constituent in living systems. It occurs in different cell types of every organism and is the fourth abundant metal ion in cells (per moles). It is intrinsically and deeply involved into cellular metabolism and function because it is such a common free divalent cation[2]. It has many biological functions one of them being is its important part in making all the cell's polyphosphate compounds stable, even those related with the production of RNA and DNA. Presence of $Mg^{2+}$ is required by over 300 enzymes to perform the catalytic action, comprising even those enzymes using or producing ATP. $Mg^{2+}$ is essential in the production of chlorophyll and photosynthesis in plants. Equilibrium of Mg is critical for good health of every organism. Mg also plays an important role as an ion for signalling, both activating and mediating numerous bio-chemical reactions. Regulation of carbon fixation in chloroplasts in calvin cycle is a best example for this. Magnesium is comparatively an element that is present in plenty in earth's mantle, crust while being bioavailable also in the earth's water bodies in a high amount. This vast abundance, along with its beneficial characteristics and extraordinary chemistry, has led to its use as an ion for enzyme activation, catalysis and signalling during the course of evolution of biological systems [3].

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. 3, Special Issue 7, October 2015**

The strange property of ionic magnesium does cause a huge difficulty in its transport inside biological systems such as membranes.Hence, transport proteins are needed to enable the movement of magnesium, both inside and outside of cells and intracellular compartments since cell membranes are impermeable to Mg. Mg transporters are proteins that transport magnesium across the cell membrane.The primary source of energy for the motion of ions in cells is concentration gradient and electric potential ($\Delta\Psi$) across the membrane. A huge number of Mg fluxes are present across the membranes and it is transported across membranes in spite of this mechanistic difficulty. Plants that suffer deficiency of Mg show responses of stress. Initial signs can be observed that indicate both Mg starvation and overexposure in plants is from reduction in photosynthesis rate. This is because of the crucial placement of magnesium ion in the molecule of chlorophyll [4]. Significant reduction in growth and reproductive viability are the later effects of Mg deficiency on plants. Mg deficiency is a harmful plant disorder that appears commonly in sandy, light and strongly acidic soils, where magnesium can beeasily drained away [5]. Losses of RNA transcription, loss of protein and carbohydrate immobility are reported as secondary effects in magnesium deficiency. Lack of Mg ion in the media (soil), can lead to deficiency but more commonly occurs due to inhibition of its uptake. Therefore maintaining a homeostatic condition by Mg Transporter Proteins is very critical for the well being of the organism and thus an effort is made to understand these proteins.

There are several methods for predicting the function of a given protein sequence. The similarity search-based tools have been utilised for functional annotation of proteins where a sequence is searched against an experimentally annotated database and a function is specified to a protein [6]. However, this methodology is unsuccessful when an unknown query protein does not have considerable sequence similarity to proteins existing in the database and sequence similarity does not always lead to functional similarity. Another way to predict the function of proteins is to identify sequence motifs such as signal peptide or nuclear localization signal. Many machine learning technique based methods such as artificial neural networks and support vector machines (SVM) have been developed to predict the function of proteins [7]. Recent advances in the prediction of function from protein sequences, have stressed the need for organism specific prediction tools [8]. In comparison to the general protein prediction methods, organism specific prediction methods are more accurate. To our best knowledge, there is no method currently available for predicting Magnesium Transporter proteins in rice, which are responsible for the transport of magnesium ions.

Machine learning is a way of automatically using training data to create or alter a model which can be later applied to make predictions for new unknown data. Classification is a task of recognizing a set of protein categories on the basis of a training set of data comprising observations (or instances) whose category membership is known based on positive and negative datasets. Machine learning classifiers are capable to distinguish among proteins belonging to different functional classes.

WEKA stands for 'Waikato Environment for Knowledge Analysis'. It is a freely available popular suite of machine learning software developed at the University of Waikato, New Zealand and written in Java (http://www.cs.waikato.ac.nz/ml/weka/). The WEKA workbench includes a compilation of visualization tools and algorithms for data analysis and predictive modeling, along with graphical user interfaces for easy access to these functionalities [9]. WEKA supports several standard data mining tasks, more specifically, clustering, data pre-processing, regression, classification, feature selection and visualization. We have used WEKA machine learning workbench to implement Sequential minimal optimization (SMO), Random forest (RF) and J48 Decision Treealgorithms.

Computational prediction methods are automatic, more accurate and quick particularly for high-throughput studies of large-scale genome sequences. Therefore, a fully automatic identification setup for Mg transporter proteins in rice is a systematic approach in this direction. The machine learning classifiers available in WEKA are developed using diverse features of protein sequences and the performance of these models are evaluated using different statistical validation techniques. Based on the best model obtained "MgTransPred", a computational tool for the prediction of Mg transporter proteins in rice has been developed in the present study, and integrated into a web-based application where users can query their protein sequence for the prediction of magnesium transport function.

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. 3, Special Issue 7, October 2015**

## II. EXPERIMENTAL PROCEDURES

### A. *Datasets*

For training the model, both positive and negative data sets are required, which were retrieved from NCBI. The positive dataset used in the present study, consisted of 90 Mg transporter protein sequences from *Oryza sativa indica*and*Oryza sativa japonica* taken from NCBI after redundancy elimination. Similarly, negative dataset was created by using 60 Non-Mg transporter protein sequences. The sequences were retrieved from Protein Database of NCBI in FASTA format accessible at http://www.ncbi.nlm.nih.gov/protein. Protein sequences were run through NCBI's conserved domain database to confirm their functional class. The positive and negative sequences were divided for training and testing purpose. A total of 25 proteins were randomly selected from the dataset to create a test dataset for independent dataset testing. For training and testing, independent datasets were used which means training dataset and test dataset were completely different.

### B. *Feature Extraction*

Feature extraction is a form of pre-processing in which the original variables are transformed into new inputs for machine learning classification. The different composition techniques used are listed as follows.

i. Amino-acid composition: It is the proportion of individual amino acid occurring in a protein sequence and this feature is of dimension 20 since there are generally 20 types of amino acids occurring in a protein. This feature misses the order of amino acids completely. To calculate the ratio of all 20 individual amino acids, the following equation was used:

Amino acid composition (n) = (Total no. of amino acidsin the protein sequence) /(Total no. of amino acids)

where, n can be any amino acid.

ii. Dipeptide composition: It gives information about a protein sequence encompassing measure of amino-acid composition coupled with local arrangement of two amino acids adjacent to each other giving fixed pattern length of 400 (20x20). Total number of amino acids is 20, so the theoretical number of all possible dipeptides is 400. A matrix of these 400 dipeptides was constructed for each protein and was then given as input to WEKA. The fraction of individual dipeptide was found according to the equation:

Dipeptide composition (n+1) = Total no of all possibleDipeptides / Total no dipeptides

where, dipeptide (n+1) is one among the 400 dipeptides.

iii. Molecular Weight: This feature gives information about the molecular weight of a protein and is of dimension 1.

Molecular Weight = Total sum of molecular weights of individual amino acids.

iv. Group Composition: In this method, we adopted a 8-letter group based on amino acid properties {E1, E2, E3, E4, E5, E6, E7, E8} to represent a protein sequence where E1$\epsilon${G, A, L, V, I} are aliphatic, E2$\epsilon${F,Y,W} are aromatic, E3$\epsilon${S,T} have OH group, E4$\epsilon${D,E} are acidic, E5 $\epsilon${N,Q} are acid amide, E6$\epsilon${R,K,H} are basic, E7$\epsilon${C,M} are sulphuric and E8$\epsilon${P} is cyclic. These exchange groups are effectively classes of amino acids that have similar chemical properties. In this method, count of amino acids based on a given group is considered. This feature has a dimension of 8.

Hybrid methods: Different hybrid approaches were developed by the combination of the above various features of protein sequence.

v. Hybrid 1: In this method, a hybrid module was developed by the combination of amino acid composition and molecular weight features of protein sequence and the fractions are were computed by using features (i) and (iii) respectively. The input vector pattern of this particular module is of 21 dimensions (20 for amino acid and 1 for molecular weight).

vi. Hybrid 2: In this method, a hybrid module was developed by the combination of amino acid composition, molecular weight and group composition features of protein sequence and the ratios were computed by using features (i), (iii) and (iv) respectively. The input vector pattern of this particular module is was of 29 dimensions (20 for amino acid, 1 for molecular weight and 8 for group composition).

vii. Hybrid 3: In this method, a hybrid module was developed by the combination of amino acid composition, molecular weight and dipeptide composition features of a protein sequence and the proportions were computed by using features (i), (ii) and (iii) respectively. The input vector pattern of this module is of 421 dimensions (20 for amino acid, 1 for molecular weight and 400 for dipeptide composition).

viii. Hybrid 4: In this method, a hybrid module was developed by the combination of amino acid composition, molecular weight, group composition and dipeptide composition features of protein sequence and the fractions are

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. 3, Special Issue 7, October 2015

computed by using features (i), (ii), (iii) and (iv) respectively. The input vector pattern of this module is of 429 dimensions (20 for amino acid, 1 for molecular weight, 8 for group composition and 400 for dipeptide composition).

We have written and developed PERL Scripts/programs to extract all these features from the protein sequences and convert them into specific file formats as required for WEKA as inputs.

C. *Performance Evaluation and Parameters*

The methods often used for examining the effectiveness of a predictor, in statistical prediction are single independent dataset test and cross-validation test [10]. In "leave-one-out" cross-validation (LOOCV), every sample in the dataset is separated out in turn as an independent test sample, and rest of the samples are used as training dataset. This procedure is continued until every sample is used as a test sample one time without any repetition. In independent dataset testing, no data from the test appears in the training dataset that is used to train the classifier and the selection of data for testing dataset is done at random. In our present work, we have adopted independent dataset validation and LOOCV methods for evaluating the performance.

We have used WEKA software Version 3.6 for classification, where various features of protein sequence were analyzed to group protein sequences into one of the predefined classes. Both training and test dataset were used to get the classification of the unknown dataset by using various algorithms [11]. The performance of WEKA was optimized in order to analyse the accuracy of classifiers. Here, we have used different approaches, based on composition of the protein sequence, to train the different classifiers [12]. Out of the 76 classifications and regression algorithms available in WEKA software, three methods (viz., SMO, J48 and Random Forest) were used for constructing classifiers.

D. *Evaluation Parameters*

We selected six measurements which are considered frequently for evaluation, viz., sensitivity (Sn), specificity (Sp), accuracy (Ac), precision (Pr), F-Measure and Mathew's Correlation Coefficient (MCC). Sensitivity (Sn) and specificity (Sp) defines correct prediction ratios of positive (+) and negative (-) datasets of Mg transporter proteins respectively. Correct ratio between both positive (+) and negative (-) datasets is defined by Accuracy (Ac). Precision is proportion of predicted positive cases that were correct. In statistical analysis of binary classification, F-measure score (also F1-score or F-score) is an estimate of a test's accuracy. It takes into account both recall (R)and precision (Pr)of the test to calculate F-measure. Ris number of true outcomes divided by number of outcomes that should have been returned and is also called as sensitivity (Sn) and precision is the number of true outcomes divided by total number of returned outcomes. F-measure is considered as a weighted average of recall and precision; It attains its best value at 1 and lowest score of 0.F-measure is harmonic mean of recall and precision [13]. However, MCC is also considered to estimate the performance of the developed prediction tool when there is a lot of difference between the number of negative data and positive datasets. MCC is regarded as the most robust parameter for evaluating methods that perform classification tasks. The range of MCC lies from -1 to 1, and a positive value of MCC indicates better performance of prediction. The real positives are termed as true positives (TP), whereas the rest are termed as false positives (FP), among the data with positive hits by the classifier models.The measurements are expressed in terms of true negative (TN), true positive (TP), false negative (FN) and false positive (FP):

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

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

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

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

$$\text{F-Measure} = 2 * \frac{Rec*Prec}{Rec+Prec} \quad (5)$$

$$\text{MCC} = \frac{(TPXTN)-(FPXFN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN}} \quad (6)$$

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. 3, Special Issue 7, October 2015**

where TP and TN are correctly or truly predicted positive magnesium transporter protein and negative non-magnesium transporter protein, respectively. FP and FN are falsely or wrongly predicted proteins.

## III. RESULTS

Prediction accuracy was assessed by two different validation techniques namely LOOCV and independent dataset validation. For achieving maximum accuracy, eight different feature- based techniques, including four composition-based and four hybrid-based methods, were used to develop models and then evaluated.

A. *Composition and hybrid based approaches in WEKA*

Models in WEKA were created for three classifiers based on features extracted using four different amino acid composition based techniques and four hybrid techniques. The performance of all the classifiers with eight methods was then statistically evaluated. The detailed performance of independent dataset test validation results of WEKA is listed in Table 1. Among these three classifiers, J48 classifier gave best results during testing. The detailed performance of LOOCV results of WEKA is listed in Table 2. The J48 model derived using amino-acid composition, hybrid 1 and hybrid 2 approaches achieved overall maximum F-Measure of 1.00 and MCC of 1.00 in independent dataset test validation (Table 1). In the LOOCV, Random Forest classifier with Hybrid3 composition achieved an accuracy of 99%. Hybrid3 composition had a very high dimension of 421 compared to amino acid composition which had only 20 dimensions. Therefore, we selected the amino acid composition based WEKA's J48 classifier as the best model compared to Random Forest.

B. *Web-based Application*

Based on our above work, we integrated the best performing module i.e amino acid composition- based WEKA's J48 classifier from the present investigation and implemented it by developing a dynamic web-based application named 'MgTransPred'. All CGI scripting of 'MgTransPred' was done using PERL programming language version 5.03 and web interface was designed using HTML to accept end user's protein sequence as a query. The overall development work plan of 'MgTransPred' web-based tool is given in Figure 1. Users can submit a protein sequence using file uploading or cut and paste options for prediction on the submission form page of the web-based application. It enables end users to enter one protein sequence at a time in the standard format such as FASTA as shown in Figure 2. After the analysis is donecomplete, prediction results are displayed in a format which is very user friendly on the output screen in a matter of few seconds. The output page presents summarized results of the prediction analysis. An illustration of prediction output is as given in Figure 3.

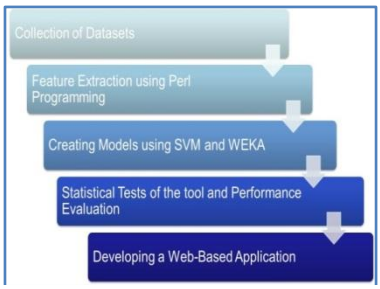


Figure 1:Overall Development Work Plan ofMgTransPred Web-based Tool

Table 1: Comparison of prediction performance of classifiers of WEKA with different composition techniques using independent dataset validation

| Method | Algorithm | Sensitivity | Specificity | Precision | Accuracy | F-Measure | MCC |
|---|---|---|---|---|---|---|---|
| Amino acid | Random Forest | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |
| | J48 | 1 | 1 | 1.00 | 1 | 1.00 | 1.00 |
| | SMO | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |
| Mol. Weight | Random Forest | 0.87 | 0.7 | 0.74 | 0.79 | 0.80 | 0.58 |
| | J48 | 0.87 | 0.8 | 0.81 | 0.84 | 0.84 | 0.67 |
| | SMO | 1 | 0 | 0.50 | 0.5 | 0.67 | - |
| Group | Random Forest | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |

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. 3, Special Issue 7, October 2015

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | J48 | 0.74 | 0.7 | 0.71 | 0.72 | 0.73 | 0.44 |
| | SMO | 1 | 0.8 | 0.83 | 0.9 | 0.91 | 0.82 |
| Dipeptide | Random Forest | 0.93 | 0.9 | 0.90 | 0.92 | 0.92 | 0.83 |
| | J48 | 0.93 | 0.7 | 0.76 | 0.82 | 0.84 | 0.65 |
| | SMO | 1.00 | 0.8 | 0.83 | 0.90 | 0.91 | 0.82 |
| Hybrid 1 | Random Forest | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |
| | J48 | 1 | 1 | 1.00 | 1 | 1.00 | 1.00 |
| | SMO | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |
| Hybrid 2 | Random Forest | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |
| | J48 | 1 | 1 | 1.00 | 1.00 | 1.00 | 1.00 |
| | SMO | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |
| Hybrid 3 | Random Forest | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |
| | J48 | 0.93 | 0.7 | 0.76 | 0.82 | 0.83 | 0.65 |
| | SMO | 1 | 0.8 | 0.83 | 0.90 | 0.91 | 0.82 |
| Hybrid 4 | Random Forest | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |
| | J48 | 0.93 | 0.7 | 0.76 | 0.82 | 0.83 | 0.65 |
| | SMO | 1 | 0.9 | 0.91 | 0.95 | 0.95 | 0.90 |

Table 2: Comparison of prediction performance of classifiers of WEKA with different composition techniques using Leave-one-out cross-validation

| Method | Algorithm | Sensitivity | Specificity | Precision | Accuracy | F-Measure | MCC |
|---|---|---|---|---|---|---|---|
| Amino acid | Random Forest | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.94 |
| | J48 | 0.95 | 0.93 | 0.93 | 0.94 | 0.94 | 0.87 |
| | SMO | 0.87 | 0.83 | 0.83 | 0.85 | 0.85 | 0.70 |
| Mol. weight | Random Forest | 0.89 | 0.86 | 0.86 | 0.87 | 0.87 | 0.75 |
| | J48 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.56 |
| | SMO | 0.60 | 0.40 | 0.50 | 0.50 | 0.55 | 0.00 |
| Group | Random Forest | 0.96 | 0.95 | 0.95 | 0.95 | 0.95 | 0.91 |
| | J48 | 0.92 | 0.90 | 0.90 | 0.91 | 0.91 | 0.82 |
| | SMO | 0.79 | 0.72 | 0.74 | 0.76 | 0.76 | 0.51 |
| Dipeptide | Random Forest | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.94 |
| | J48 | 0.87 | 0.83 | 0.83 | 0.85 | 0.85 | 0.70 |
| | SMO | 0.91 | 0.87 | 0.87 | 0.89 | 0.89 | 0.77 |
| Hybrid 1 | Random Forest | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.94 |
| | J48 | 0.93 | 0.92 | 0.92 | 0.93 | 0.93 | 0.85 |
| | SMO | 0.89 | 0.84 | 0.85 | 0.86 | 0.87 | 0.73 |
| Hybrid 2 | Random Forest | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.94 |
| | J48 | 0.97 | 0.96 | 0.96 | 0.96 | 0.96 | 0.92 |
| | SMO | 0.89 | 0.85 | 0.86 | 0.87 | 0.87 | 0.74 |
| Hybrid 3 | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 |
| | J48 | 0.89 | 0.86 | 0.87 | 0.88 | 0.88 | 0.76 |
| | SMO | 0.9 | 0.86 | 0.87 | 0.88 | 0.88 | 0.76 |
| Hybrid 4 | Random Forest | 0.98 | 0.97 | 0.97 | 0.98 | 0.98 | 0.95 |
| | J48 | 0.89 | 0.862 | 0.87 | 0.88 | 0.88 | 0.76 |
| | SMO | 0.91 | 0.881 | 0.88 | 0.90 | 0.90 | 0.79 |

## IV. DISCUSSION

With advances in genome sequencing technologies and rapid availability of whole genome sequences, tools and resources need to be developed to deduce the information contained in these genome sequences. A big challenge in rice annotation is the unavailability of efficient gene and protein prediction softwares and tools. Moreover rice, which is a reference organism, among the plant species, plays as the biological role for many cereal proteins. Therefore, availability of tools and computer- based systems that can detect and identify biological function from the sequence is remarkable and is necessary for the complete characterization and understanding of a protein that gets expressed. Computer- based programs give accurate and quicker access to various prediction based findings for studying various species including plants.

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. 3, Special Issue 7, October 2015**



Figure 2:An overview of the protein sequence submission form Figure 3:An overview of the output page displaying the predicted results

## V. CONCLUSION AND FUTURE SCOPE

We have used eight different features derived from protein sequences for the identification of biological function of Mg transport in a given protein specific to rice in our present study, but as a scope for future improvement, the potential of prediction module may be refined and expanded by implementing other features of sequences and machine learning algorithms. Another limitation of this method is that it just predicts the single function of a protein, whereas in realistic situation it is observed that a protein may have multiple functions. These are the areas where there is a possibility to work upon in this direction for further development. Here, we have presented a technique for prediction of function implemented in WEKA, whose performance is observed to be extremely satisfactory. Very high prediction accuracies show that our model is potentially helpful for the detection of magnesium transporter proteins in rice.

## REFERENCES

[1] Zeigler RS, 'Rice Research for Poverty Alleviation and Environmental Sustainability in Asia', IRRI (S3),pp. 6-11,2006.
[2] Kehres D and Maguire M, 'Structure, properties and regulation of magnesium transport proteins',BioMetals, Vol.15, Issue 3,pp. 261–270,2002
[3] Gardner R, 'Genes for magnesium transport', Current Opinion in Plant Biology, Vol.6, Issue 3, pp. 263–267, 2003.
[4] Shaul O, 'Magnesium transport and function in plants: the tip of the iceberg', BioMetals, Vol 15, Issue 3, pp. 309–323,2002.
[5] Norman PA and William Hopkins, 'Introduction to Plant Physiology', John Wiley & Sons, New York, 2010.
[6] Altschul, Stephen F., et al.,'Basic local alignment search tool.' Journal of molecular biology, Vol. 215, Issue.3, pp.403-410.1990.
[7] Predrag R et al., 'A large-scale evaluation of computational protein function prediction', Nature Methods, Vol.10, Issue 3, pp. 221-227, 2013.
[8] Schneider and Fechner, 'Advances in the prediction of protein targeting signals', Proteomics, Vol. 4, Issue 6, pp. 1571-1580,2004.
[9] Witten IH and Frank E, 'Data Mining: Practical Machine Learning Tools and Techniques', Morgan Kaufmann, San Francisco,2005.
[10] Baldi P, Brnak S, Chauvin Y, Andersen CAF and Nielsen H, 'Assessing the accuracy of prediction algorithms for classification: an overview',Bioinformatics , Vol.16, pp. 412-424, 2000.
[11] Mark H, Eibe F, Geoffrey H, Bernhard P, Peter R and Ian H, 'The WEKA Data Mining Software', SIGKDD Explorations,Vol. 11, Issue 1, pp. 431-436,2009.
[12] Hemalatha N, Rajesh MK and Narayanan NK., 'An integrative system for Prediction of NAC proteins in rice using different feature extraction methods', International Journal on Soft Computing, Vol. 4, Issue 1, pp. 9-21, 2013.
[13] Predrag R et al., 'A large-scale evaluation of computational protein function prediction', Nature Methods, Vol. 10, Issue 3, pp. 221-227,2013.