# Building a Model for Predicting Target Gene Expression in Rice T-DNA Insertional Mutants by Machine Learning Approaches

**Chi-Chou Liao**
National Chung Hsing University College of Life Sciences
**Liang-Jwu Chen**
National Chung Hsing University
**Shuen-Fang Lo**
Academy of Sciences
**Chi-Wei Chen**
National Chung Hsing University
**Jia-Jyun Chen**
National Chung Hsing University
**Yen-Wei Chu** ( ywchu@nchu.edu.tw )
National Chung Hsing University https://orcid.org/0000-0002-5525-4011

**Research article**

**Keywords:** Rice, CaMV 35S enhancer, T-DNA activation tagging, Gene expression, Machine learning

**Posted Date:** April 22nd, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-20492/v1

# Building a Model for Predicting Target Gene Expression in Rice T-DNA Insertional Mutants by Machine Learning Approaches

**Chi-Chou Liao[1†], Liang-Jwu Chen[1,2†], Shuen-Fang Lo[2,3], Chi-Wei Chen[4,5], Jia-Jyun Chen[5] and Yen-Wei Chu[1,2,5,6,7,8*]**

1 Institute of Molecular Biology, National Chung Hsing University, Taichung, Taiwan
2 Agricultural Biotechnology Center, National Chung Hsing University, Taichung, Taiwan
3 Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan
4 Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan
5 Institute of Genomics and Bioinformatics National Chung Hsing University, Taichung, Taiwan
6 Biotechnology Center, National Chung Hsing University, Taichung, Taiwan
7 Ph.D. Program in Translational Medicine, National Chung Hsing University, Taichung, Taiwan
8 Rong Hsing Research Center for Translational Medicine, National Chung Hsing University, Taichung 402, Taiwan

† These authors contributed equally to this work.
* Correspondence: ywchu@nchu.edu.tw; Tel.: +886-4-2284-0338#7041

**Abstract**

**Background:** T-DNA activation-tagging technology is widely used to enhance flanking gene expression near the site of insertion for functional genomics research in rice. However, whether the expression of a gene of interest is enhanced must be validated experimentally.

**Results:** In this study, we built a model to predict gene expression in T-DNA mutants by machine learning approaches, thereby improving the efficiency of screening for activated genes. We gathered experimental consisting of gene expression data in T-DNA mutants and captured the PROMOTER and MIDDLE sequences for encoding. In first-layer models, SVM models were constructed with nine features consisting of information about biological function and local and global sequences. Feature-encoding based on the PROMOTER sequence was weighted by logistic regression. The second-layer models integrated 16 first-layer models with feature selection and the algorithm, which were selected from nine feature selection methods and 65 classified methods, respectively. The accuracy of the final two-layer machine learning model, referred to as was 99.3% based on five-fold cross-validation, and 85.6% based on independent-testing.

**Conclusion:** We discovered that the information within the local sequence had a greater contribution than the global sequence with respect to classification had a good predictive ability for target genes within 20 from the 35S enhancer. Based on the analysis of significant sequences, the G-box regulatory sequence may also play an important role in the mechanism of activation of the 35S enhancer.

**Keywords:** Rice, CaMV 35S enhancer, T-DNA activation tagging, Gene expression, Machine learning

## Background

Rice is one of the most important models of monocotyledon plants for the analysis of plant gene function. Rice is one of three major food crops throughout the world, and it is the staple food of more than half of the world's population. In the past 30 years, rice production has doubled, although the supply of rice is expected to gradually become insufficient with the rapid increase in the world

population, climate change and a shortage of water [1]. It will not be easy to increase food production to the necessary levels. In 2004, the International Rice Genome Sequencing Project (IRGSP) completed the sequencing of the rice genome [2]. The ultimate goal of genome analysis is to realize the structure and function of each gene within an organism. To further confirm the function of and metabolic pathways related to each gene in rice, scientists have focused their efforts on analyzing the rice genome and are committed to promoting rice genome annotation to move rice research into the post-genome era.

T-DNA insertion activation-tagging technology is widely used in the analysis of the function of rice genes [3, 4]. This method results in the construction of four tandem cauliflower mosaic virus (CaMV) 35S enhancers on a T-DNA plasmid; when this T-DNA is inserted into the rice genome, it activates genes that flank the T-DNA insertion site [5]. The CaMV 35S enhancer can activate gene expression in dicots and monocots and is widely used in T-DNA transformation. Gene expression gradually increases with the number of 35S enhancers on T-DNA, which led to the incorporation of four tandem repeat CaMV 35S enhancers for enhanced gene expression with this approach [6-11]. *Agrobacterium*-mediated T-DNA transformation tends to insert one copy of T-DNA, an average of 1.4 loci of T-DNA inserts in transgenic plants [12], reducing the complexity of rice gene research. T-DNA inserted into the rice genome with 35S enhancer resulted in two states. 1) Gene knockdown: when T-DNA is inserted into the coding region or promoter of a gene, it is likely to destroy the structure of the gene, resulting in reduced function or loss of function of the gene. 2) Activation-tagging: T-DNA might enhance the activity of genes that flank the T-DNA insertion site through the effect of the 35S enhancers. Thus, we can make use of T-DNA insertion activation-tagging to study the association between genetic function and morphological traits [5]. However, there has been no basis for determining whether a target gene is activated by the enhancer prior to experimental analyses. There has even been a study indicating that the enhancer can activate genes that are millions of base pairs away from the enhancer [13]. Not all of the genes that flank the T-DNA insertion site are expected to be activated by the 35S enhancer in our experiments. In some T-DNA mutants, the 35S enhancer does not activate the closer gene but instead activates a gene that is farther away from the 35S enhancer [14]. Researchers thus cannot rely on the distance between the enhancer and a particular gene to judge whether that gene would be activated. They must instead determine the activated genes experimentally to explore the related genetic function and morphological traits. Therefore, it is a time-consuming and laborious process to check for the expression of a target gene.

In this study, we used machine learning approach to predict target gene expression in rice T-DNA insertion mutants and improved the efficiency of finding activated target genes. Our team had developed a website platform, EAT-Rice [15], for predicting the expression status of rice genes that flank the T-DNA insertion site in activating mutants. The system of EAT-Rice applied the distance factor from T-DNA insertion site to gene loci to weight feature encoding, and used two kinds of algorithms to build two-layer model of machine learning. In this study, we based on EAT-Rice with a modified sequence capturing method, system architecture and other additional features to build a more comprehensive system for target gene expression prediction in T-DNA insertion mutants.

The datasets used in this study were experimentally validated. We first characterized genes based on their activation by the 35S enhancer, these genes were divided into activated genes and non-activated genes. The system we built refer to the EAT-Rice. We captured the DNA sequence of the promoter and the central region of each activated gene from the start codon of the target gene to the 35S enhancer and used nine features for encoding. We then used LibSVM (Library Support Vector Machine) [16] and LADTree [17] algorithms to build two-layer models of machine learning. In the first-layer model, we carried out a logistic regression to weight the features that depending on the probability of gene activation and the distance from the enhancer to the gene start codon. Moreover, we used the mRMR (Minimum Redundancy Maximum Relevance) [18] method and incremental

feature selection to determine the most relevant features in the second layer. This system is referred to as TIMgo.

The TIMgo performance was 99.3% based on five-fold cross-validation and 85.6% based on independent-testing. TIMgo had >80% accuracy for target genes within 20 kb from the 35S enhancer, but genes that were >20 kb away were still predicted with >60% accuracy. We also discovered that the value of the *k* parameter for Kmer, RevKmer and PseKNC encoding within the PROMOTER sequences was higher than that of MIDDLE sequences. This suggested that for the analysis of longer sequences a greater number of features was needed to improve the prediction performance. Finally, the G-box *cis*-element has an important function in gene activation by the 35S enhancer based on the motif analysis, and among the G-box–associated binding proteins most are bZip (basic region/leucine zipper) transcription factors.

## Methods

### *Sources for T-DNA Mutant Data and Datasets*

The experimental data were collected from 11 rice T-DNA mutants from Liang-Jwu Chen's laboratory at NCHU and 316 mutants from Su-May Yu's research team at the Academia Sinica. These data consisted of the T-DNA insertion point and expression status of flanking genes (as detected by RT-PCR [28]). The expression status of each gene was characterized based on the four following categories: activated gene (Ac), no significant effect (NE), non-detectable (ND) and knockout (Ko). The data distribution for the expression status of these genes is shown in Table 1.

To maintain dataset quality and consistency, we removed the 30 ND genes from the dataset. The collected data included two Ko genes, in which the T-DNA insertion point was located inside the gene, thus disrupting the gene structure and most likely leading to a loss of function. Because Ko genes were not a focus of this study, we removed them from the dataset. We defined NE genes as non-activated (NAc) genes to differentiate them from the Ac genes. Ultimately, data for 453 genes were collected in this study.

A training set was used to determine the performance of the subsequent system. As the ratio of positive data (Ac genes) to negative data (NAc genes) affects the performance of the machine learning [29], this study used EAT-Rice with a 1:1 ratio to carry out the selection of the training dataset. We used data from 300 genes in the training dataset, which was referred to as D300. Data from the remaining 153 genes were used for independent-testing to evaluate system accuracy; this dataset was referred to as D153 (Table 2).

### *Target Gene Sequence Retrieval*

The analyzed genes provided from Liang-Jwu Chen's laboratory and Su-May Yu's team were annotated according to the Rice Genome Automated Annotation System (RiceGAAS)[30] and the MSU Rice Genome Annotation Project (TIGR)[31, 32] rice gene annotation database. We hypothesized that we could predict the expression status of a target gene by analyzing the sequence of Ac and NAc genes. Thus, with reference to the EAT-Rice construction process and the enhancer-related hypothesis mechanisms [33, 34], we extracted nucleotide sequences for each gene from two regions: 1) a 1500-bp region upstream relative to the translation start site (TLS), referred to as the PROMOTER region, and 2) a central region of 300 bp centered between the TLS of the target gene and the 35S enhancer, referred to as the MIDDLE region (Supplementary Figure S1).

### *Feature Encoding*

In this study, we encoded information about nine features of the sequences: five sequence information codes and four biological functional codes. The sequence codes consisted of two local sequence codes, two global sequence codes and a code to reflect both the local and global sequence

information simultaneously. The local sequence characteristics consisted of Kmer and Reverse complementary kmer (RevKmer) values, which were coded by the DNA composition; such characteristics have been successfully applied toward human gene regulatory sequence prediction [35, 36] and enhancer identification [37], among others. The two global sequence codes, dinucleotide-based auto-cross covariance (DACC) and trinucleotide-based auto-cross covariance (TACC), were coded by calculating the sequence autocorrelation as global sequence characteristics; this type of feature has been used to predict sequence-based protein-protein interactions [38]. Another coding method, PseKNC, has been used to identify promoters in prokaryotes [39] and incorporates the contiguous local sequence-order information and the global sequence-order information into the feature vector. The biological characteristics included the presence of CpG-islands (CGIs), regulatory *cis*-elements (Motif) and conformational and physicochemical properties of dinucleotide and trinucleotide sequences (DNP and TNP, respectively). Each of these features is described in more detail below.

CpG-Islands (CGIs)

DNA methylation on CpG-islands reduces or silences gene expression based on enhancer-promoter interactions [40, 41]. For this analysis, we used the EMBOSS Newcpgreport tools of EMBL-EBI to predict CpG-islands and encoded their corresponding number, length, distance from the TLS, CpG ratio and OE (observed/expected) value, resulting in the feature CGIs were as follows:

$$\text{CGI\_Number} = \begin{cases} j,\ j \in \text{N} \\ 0, \text{otherwise} \end{cases} \quad (1)$$

$$\text{CGI\_LengthRatio} = \frac{\text{length of CGI}}{\text{length of sequence}} \quad (2)$$

$$\text{CGI\_Dis} = |\text{TLS} - \text{CGI location site}| \quad (3)$$

$$\text{CGI\_CGRatio} = \frac{\text{CpG percent in CGI}}{\text{CGI\_Number}} \quad (4)$$

$$\text{CGI\_OE} = \frac{\text{number of CpGs in CGI}}{(\text{number of Cs in CGI}) \times (\text{number of Gs in CGI})} \quad (5)$$

Number coding was represented by the number of CGIs (j) predicted in the sequence (Equation 1). Length coding consisted of the ratio of the CGIs length dived by the PROMOTER or MIDDLE sequence length (Equation 2). Distance was encoded as the distance from the CpG-island to the TLS (Equation 3). The CG ratio was calculated as the ratio of CpG fragments in the CpG-island by dividing total number of CGIs (Equation 4). The OE value indicates the ratio of the number of CpGs present in the CpG-island relative to the expected value of CpG fragments and was calculated by dividing the number of CpG fragments in the sequence by the product of the number of Cs and the number of Gs in the CpG-island (Equation 5).

Regulatory *cis*-Elements (Motif)

To consider that the transcription factor binding sites (TFBSs) of rice have been confirmed may not be comprehensive enough yet, we therefore incorporated proven TFBSs from other plants. Data for 2087 motifs were collected from PLACE [42] and the RegSite database (http://linux1.softberry.com/berry.phtml?topic=regsitelist). The tool Find Individual Motif Occurrences (FIMO)[43] in the MEME suite was used to scan for regulatory sequences in the PROMOTER region, and the scanning results were encoded by FIMO [44, 45]. This features encoding are listed as follows:

$$\text{Motif\_Number}_{(i)} = \begin{cases} j,\ j \in \text{N} \\ 0,\ \text{otherwise} \end{cases},\ i \in \{1,2,\cdots,2087\} \quad (6)$$

$$\text{Motif\_Conserve}_{(i)} = \frac{M_i \text{ alignment score in promoter}}{\text{Motif\_Number}(i)} \quad (7)$$

$$\text{Motif\_Orientation}_{(i)} = \frac{pos \text{ in Motif\_Number}_{(i)}}{\text{Motif\_Number}_{(i)}} \quad (8)$$

$$\text{Motif\_Dis}_{(i)} = \frac{\left|\text{geneTLS - Motif location site}\right|}{\text{Motif\_Number}_{(i)}} \quad (9)$$

The number of regulatory elements was coded by the number (*j)* of motifs found in the PROMOTER (Equation 6). The conservation score was calculated by FIMO; we summed up the conserved scores of specific motifs and divided by the number of motifs in the sequence (Equation 7). As motifs can be located on both the DNA coding strand (codons) and the template strand (anti-codons), the orientation characteristic was calculated to determine the proportion of motifs on the coding strand. We thus used the number of motifs on the coding strand (i.e., positive motifs, pos) as the numerator, and the denominator is the number of all motifs (Equation 8). The distance characteristic was determined based on the distance (in base pairs) from each motif to the TLS, which was summed for all motif sites within a given sequence, divided by the number of motifs (Equation 9). In these equations, *i* indicates the kinds of motifs, Mi indicates a specific motif and the geneTLS refers to the translation start site of a target gene.

Kmer and Reverse Complementary Kmer (Kmer and RevKmer)

Kmer refers to the local sequence information and indicates a subsequence containing *k* neighboring nucleic acids in a DNA sequence. Using a coding strand as the template, the Kmer feature will scan for the number of occurrences of the nucleic acid subsequence in the template. For example, when *k* is 2, the subsequence composition of a Kmer will be called a 2-mer, which contains 16 subsequences (based on the four nucleotides, G, A, T and C). In the case of the AA dinucleotide, if this subsequence appeared twice in the DNA template, it would be encoded as 2; if it was not present in the template, it would be encoded as 0. In eukaryotes, the average length of TFBSs is 10 bps [46], which suggests that the number of *k* neighboring nucleic acids in this study could be increased. We encoded the sequence with 3- to 6-mer, 3- to 7-mer, 3- to 8-mer and 3- to 9-mer, which produced 5440, 21824, 87360 and 349504 different nucleotide compositions, respectively. The Kmer encoding was carried out based on the number of occurrences in the template sequence (Equation 10).

Reverse complementary kmer (RevKmer) is a variant of kmer- in which the kmers are not expected to be strand specific, so reverse complements are collapsed into a single value. In this study, the RevKmer feature was encoded in the same manner as Kmer and produced 2760, 10952, 43848 and 174920 nucleotide compositions for the 3- to 6-mer, 3- to 7-mer, 3- to 8-mer and 3- to 9-mer, respectively. RevKmer encoding was carried out according to the number of occurrences in the template sequence (Equation 11).

$$\text{Kmer\_Number}_{(i)} = \begin{cases} j, & j \in \text{N} \\ 0, & \text{otherwise} \end{cases}, \quad i \in \{1, 2, \cdots, 349504\} \quad (10)$$

$$\text{RevKmer\_Number}_{(i)} = \begin{cases} j, & j \in \text{N} \\ 0, & \text{otherwise} \end{cases}, \quad i \in \{1, 2, \cdots, 174920\} \quad (11)$$

where *i* indicates kinds of nucleotide combinations and *j* indicates the number of matches with pattern (*i*) in a specific sequence. If there are no matches for a pattern in a specific sequence, this is coded as zero.

Nucleotide Conformational and Physicochemical Properties (DNP and TNP)

The nucleotide conformation and physicochemical properties of dinucleotide**s** and trinucleotides were also encoded. DiProDB provides information about 125 properties of dinucleotides, and these 125 properties were integrated into 15 characteristics through a statistical Principal components analysis (PCA) method [47]. The value of each property is based on the dinucleotide as a unit, and each property has 16 values corresponding to all possible dinucleotide combinations. We used the property of the dinucleotide to produce a training model with 240 dimensions; this feature is referred to as the DNP (dinucleotide conformation and physicochemical properties) (Equation 12). PseKNC-General (the general form of pseudo k-tuple nucleotide composition) is a tool that provides the conformation and physicochemical properties of oligonucleotides [48]. In this study, 12 trinucleotide properties were used for coding. There were 64 combinations of trinucleotides, which generated a training model with 768 dimensions based on the 12 trinucleotide properties; this feature is referred to as the TNP (trinucleotide conformation and physicochemical properties) (Equation 13).

$$\mathrm{DNP\_Value}_{(i,j)} = \frac{S(d_i)\times \mathrm{F}_j(d_i)}{\text{sequence legnth - 1}},\ i \in \{1,2,\cdots,16\},\ j \in \{1,2,\cdots,15\},\ d_i \in \mathrm{D},\ \mathrm{F}_j \in F \quad (12)$$

$$\mathrm{TNP\_Value}_{(i,j)} = \frac{S(d_i)\times \mathrm{F}_j(d_i)}{\text{sequence legnth - 1}},\ i \in \{1,2,\cdots,64\},\ j \in \{1,2,\cdots,12\},\ d_i \in \mathrm{D},\ \mathrm{F}_j \in F \quad (13)$$

For the encoding of DNP, *i* indicates the kind of dinucleotide (two-nucleotide combination), and *j* indicates the kind of physicochemical structure of the dinucleotide. D is the collection of 16 dinucleotides (*i*), F is the 15 integrated dinucleotide properties (*j*), S(D) is the number of times that the 16 dinucleotides appear in the target sequence, F(D) is the value of the 15 properties corresponding to 16 dinucleotides (Equation 12). For the encoding of TNP, D is the collection of 64 trinucleotides (*i*), F is the 12 trinucleotide properties (*j*), S(D) is the number of times that the 64 dinucleotides appear in the target sequence, F(D) is the value of the 12 properties corresponding to the 64 dinucleotides (Equation 13). The length of target sequence minus one indicates the maximum number of dinucleotides could match on sequence.

Autocorrelation (DACC and TACC)

Pse-in-One provides a pseudo-components mode reflecting the correlation between two dinucleotides or trinucleotides within a DNA sequence via their physicochemical properties [49]. In this study, we used dinucleotide-based auto-cross covariance (DACC) and trinucleotide-based auto-cross covariance (TACC) as provided by Pse-in-One for encoding (Equations 14-16).

In this study, DACC was based on the 15 properties from DiProDB and the *lag* value was 4, generating a training model with 900 dimensions. TACC used the 12 Pse-in-One built-in properties and the *lag* value was 4; it generated a training model with 576 dimensions. There formulas were show as follows:

$$D = R_1R_2R_3R_4R_5R_6\cdots R_L \quad (14)$$

$$\mathrm{DACC}_{(u_1,\,u_2,\,lag)} = \sum_{i=1}^{L-lag-1} \frac{\left(P_{u_1}(R_iR_{i+1}) - \overline{P_{u_1}}\right)\left(P_{u_2}\left(R_{i+lag}R_{i+lag+1}\right) - \overline{P_{u_2}}\right)}{L-lag-1} \quad (15)$$

$$\mathrm{TACC}_{(u_1,\,u_2,\,lag)} = \sum_{i=1}^{L-lag-2} \frac{\left(P_{u_1}(R_iR_{i+1}R_{i+2}) - \overline{P_{u_1}}\right)\left(P_{u_2}\left(R_{i+lag}R_{i+lag+1}R_{i+lag+2}\right) - \overline{P_{u_2}}\right)}{L-lag-2} \quad (16)$$

for a DNA sequence D with *L* nucleic acid residues, where $R_1$ represents the nucleic acid residue at the sequence position 1, $R_2$ the nucleic acid residue at position 2 and so on (Equation 14). DACC and TACC measure the correlation of the same physicochemical index between two dinucleotides or

trinucleotides separated by a distance of *lag* along the sequence (Equations 15, 16). In these equations, $u_1$ and $u_2$ are two different physicochemical indices, $P_u(R_iR_{i+1})$ and $P_u(R_iR_{i+1}R_{i+2})$ are the numerical value of the physicochemical index $u$ for the dinucleotide $R_iR_{i+1}$ or trinucleotide $R_iR_{i+1}R_{i+2}$ at position $i$ and $\overline{P_u}$ is the average value for the physicochemical index value $u$ along the whole sequence. The number of features represented by DACC and TACC, will be the lag value multiplied by the square of the properties number.

Pseudo K-Tuple Nucleotide Composition (PseKNC)

Pseudo k-tuple nucleotide composition (PseKNC) is one of the encoding modes supplied by Pse-in-One. It incorporates both the contiguous local sequence order information (like Kmer and RevKmer) and the global sequence order information (like DACC and TACC) into the feature vector of the DNA sequence.

$$D = R_1R_2R_3R_4R_5R_6 \cdots R_L \tag{17}$$

$$\text{PseKNC}_{(u)} = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j}, u \in \{1,2,\cdots 4^k\} \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j}, u \in \{4^{k+1},(4^{k+1}+1),\cdots,(4^{k+1}+\lambda)\} \end{cases} \tag{18}$$

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\{\frac{1}{\mu}\sum_{v=1}^{u}[P_v(R_iR_{i+1}) - P_v(R_{i+j}R_{i+j+1})]\}^2, j \in \{1,2,\cdots,\lambda\}, \lambda < L \tag{19}$$

For a DNA sequence D with $L$ nucleic acid residues, where $R_1$ represents the nucleic acid residue at the sequence position 1, $R_2$ the nucleic acid residue at position 2 and so on (Equation 17). PseKNC will calculate the occurrence frequency ($f$) of dinucleotides in the DNA sequence and the correlation between two oligonucleotides that are 1 to $\lambda$ nucleotides apart from each other. In equation 18, $f_u$ is the occurrence frequency of dinucleotides in the DNA sequence, which is normalized to $\sum_{i=1}^{4^k} f_i = 1$; $w$ is the weight factor; $\theta_j$ represents the correlation factor that reflects the sequence-order correlation between all two dinucleotides that are j nucleotides away from each other along a DNA sequence; $\mu$ is the number of physicochemical indices; $P_v(R_iR_{i+1})$ represents the numerical value of the dinucleotide located at $i$th position ($R_iR_{i+1}$) of the $v$th ($v$ = 1, 2, …, $\mu$) physicochemical property (Equation 19). The feature number of PseKNC will be $\lambda$ multiplied by 4 to the power $k$. In this study, the PseKNC feature was determined with a $\lambda$ value of 4, $w$ is 0.2, and $k$ is from 2 to 6.

***Significant Sequence Fragments Analysis***

Because there are numerous features in this first-layer model, the complexity of the model is relatively high. To reduce the interference of excessive noise, we used independent two-sample t-tests (implemented in R) to select features from the high-dimension models. We used the occurrence of specific oligonucleotides in the Ac and NAc groups to generate the t-test (Supplementary Figure S2) and retained the oligonucleotides with $P < 0.05$ to encode these significant fragments.

***Model Evaluation and Cross-Validation***

We used a five-fold cross-validation method and independent-testing data to evaluate the predictive performance of the model. Our evaluation methods included Accuracy (Acc), Sensitivity (Sn), Specificity (Sp) and Matthews Correlation Coefficient (MCC). Acc is used to estimate the prediction accuracy of the global prediction capability, with values closer to 100% indicating the better overall predictive performance of a model (Equation 20). Sn and Sp evaluate the accuracy of

the prediction of positive and negative data, respectively (Equations 21, 22). When the number of positive and negative data differs, Acc is not a good evaluation indicator. MCC is, however, suitable for assessing a dataset in which there is an imbalance between positive data and negative data (Equation 23). When the MCC score is closer to 1, the prediction capability is better; a score closer to −1 indicates a worse prediction capability.

$$\mathrm{Acc} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (20)$$

$$\mathrm{Sn} = \frac{TP}{TP + FN} \quad (21)$$

$$\mathrm{Sp} = \frac{TN}{TN + FP} \quad (22)$$

$$\mathrm{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \quad (23)$$

***Framework of TIMgo***

TIMgo is a two-layer machine learning model constructed for predicting the effect of 35S enhancer on the expression of target gene (Figure 1). The D453 were dived into training dataset (D300) and independent-testing data (D153). The DNA sequence of Promoter and Middle were retrieved for analyzing between NAc and Ac gene. In the first-layer module, the SVM models were constructed within 9 feature encoding methods. And the significant sequences were analyzed by student's t-test and a model of logistic regression was used to assist training, which based on the relationship between distance from the 35S enhancer to the target gene and states of gene expression. The features encoded from Promoter region were weighted by logical regression model for probability of gene activation. Then, we adopted feature selection by LIBSVM built-in tool in the partial SVM models. The prediction results of first layer module were integrated in the second layer model, and mRMR [50] were used to feature selection and built the LAD tree model. At the last, we evaluated the prediction efficacy of TIMgo with D153 independent-testing dataset.

**Results**

***Correlation between Gene Activation and Distance from the 35S Enhancer to the TLS***

The distance between the enhancer and a target gene cannot be directly used to determine whether the target gene will be activated, although it does have some relevance for determining gene activation [19, 20]. A target gene is more likely to be activated if it is closer to the enhancer [21]. We characterized each of the 453 genes in the entire dataset (D453) based on the distance from the CaMV 35S enhancer on the inserted T-DNA to TLS and calculated the ratio of Ac genes and NAc genes. We found a negative correlation between this distance and gene activation. Genes closer to the 35S enhancer had a greater probability of activation ($P < 0.001$) (Supplementary Figure S3). The results are the same as those indicated in a previous study [15](Figure 2, Supplementary Table S1).

Among the D453 dataset, there were 94 sets of duplicated data which consist of multiple genes, and the PROMOTER sequences corresponding to these genes were identical. Each of the experimental data in this study represented the effect of a single insertion event on its target gene. In the experimental data collected in this study, when the same gene was detected for multiple T-DNA insertion events, the PROMOTER sequences from those genes were identical. For different T-DNA insert events, the 35S enhancer may result in different states of expression for the same target gene, which will lead to contradictory results while building the machine learning model. To distinguish between these PROMOTER sequences, we used logistic regression to build a regression model of the

distance coefficient and the target gene activation probability (Supplementary Equation S1). In this study, the values calculated by logistic regression were used to weight the promoter sequence feature, so that the same sequence could be distinguished when quantified based on numerical values.

### *Comparison of Kmer and RevKmer Combined with Motif*

In the Kmer and RevKmer feature models, a t-test was used to calculate the number of occurrences of specific sequence fragments in Ac and NAc genes, respectively, from sequence lengths ($k$) of 3–9 nucleotides. The specific sequence fragments with $P < 0.05$ were then used for encoding. These fragments were combined as 3–6, 3–7, 3–8 and 3–9 combinations for Kmer and RevKmer. The Motif feature was used to carry out a similar analysis. The Kmer and RevKmer features associated with the PROMOTER region were combined with the Motif feature (Supplementary Table S2). The features from Kmer, RevKmer, Kmer + Motif and RevKmer + Motif were used to build SVM models, and the best model was selected for the second-layer model integration (Supplementary Table S3).

Before combining the Motif with Kmer or RevKmer, the Acc scores of the SVM models of Kmer and RevKmer were 55–85%, whereas the Acc scores of the Motif model**s** were 52–75%. After combining the Motif with Kmer or RevKmer, the Acc scores were 78–86%, and the Acc consistently increased with the $k$ value for Kmer and RevKmer (Table 3).

### *First-Layer Models Evaluation*

In the first-layer models, nine feature coding methods and two types of sequences were used to construct into 16 feature models (Supplementary Table S4). The prediction ability of each feature model was evaluated with five-fold cross-validation and independent-testing with the D153 data (Table 4). For the Pre-in-One feature encoding, one gene sequence from the training dataset (D300) did not conform to the encoding requirements. Therefore, in the DACC, TACC and PseKNC model, this information was removed from the training data, and the training dataset consisting of the remaining 299 genes was referred to as D299. The PseKNC models used $k$ values of 2–7, and eight models each were established for the PROMOTER and MIDDLE sequences. A PseKNC model with $k = 6$ that was selected among the PROMOTER models had an Acc of 75.3% with five-fold cross-validation. The PseKNC model with $k = 2$ that was selected among the MIDDLE models had an Acc of 59.5% (Supplementary Table S5).

In the evaluation results of the first-layer feature models (Table 4), the Kmer, RevKmer, Kmer + Motif and RevKmer + Motif had the best predictive performance based on the Kmer feature provided. Their Acc values were 79–88.3% with five-fold cross-validation. With independent-testing, their Acc values were 80.4–84.3%, with the exception for RevKmer, which was 67.3%. The PseKNC model built using the PROMOTER sequence was slightly inferior to the model built using Kmer-related features. The Acc and MCC values for PseKNC were 75.3% and 0.53 with cross-validation, respectively, and 56.2% for Acc and 0.165 for MCC with independent-testing. The DACC, TACC, DNP, CGIs and TNP constructed by the PROMOTER sequence and the PseKNC constructed **by** the MIDDLE sequence had lower predictive performance, with Acc values of 58.2–69.9% and MCC values of 0.164–0.398. Among these 16 models, CGIs and TNP constructed using the MIDDLE sequence were the least accurate in cross-validation, with an Acc of ~47%. Their Acc values for independent-testing were 11.8% and 62.1%, respectively. In terms of overall predictive performance, the PROMOTER sequence is thus more important than the MIDDLE sequence, and Kmer, RevKmer, Kmer + Motif and RevKmer + Motif features have the highest correlation with the activation of genes.

### *Comprehensive Feature Selection in Second-Layer Model*

The second-layer model integrated the prediction results from the 16 feature models in the first layer as features to build model by machine learning and obtained the ultimate prediction result. The features used in the second-layer model of this study included predictive results and positive and

negative predictive confidence scores, generating 48 features. We used incremental feature selection and an SVM model with cross-validation to carry out comprehensive feature selection among these 48 features to pick out the best feature combinations with nine feature selection methods. The top 33 features of the mRMR [18] were selected as the best feature combination with the highest Acc and the fewest features (Figure 3, Supplementary Table S6). Among the 33 selected features, we knew that the encoding contributed to classification is DACC and Kmer related principally, the PseKNC and TACC are secondary, and the few are CGIs, TNP and DNP.

### *Second-Layer Model Evaluation*

We assessed the best-suited machine learning algorithm for the second-layer model through the WEKA [22] analysis platform. In this study, we use the 65 algorithms provided by WEKA to establish the model separately and evaluated the effectiveness of these models with cross-validation (Supplementary Table S7). In this experiment, the LADTree algorithm was used to construct the second-layer integration model according to the above conditions. The Acc was 99.3%, MCC was 98.7% and Sn and Sp were 0.993. In independent-testing, the model Acc reached 85.6%, MCC was 35.3%, Sn was 0.891 and Sp was 0.533. Among the testing data, there were only 15 negative data, such that each predictive result with these data would lead to a substantial impact on the overall predictive effectiveness assessment. Among these models built with multiple algorithms, Sp values ranged from 0.467 to 0.733, which corresponded to a difference of only six correctly predicted negative data.

### *Correlation between Predictive Accuracy and Distance from 35S Enhancer to TLS*

To analyze the relationship between distance and TIMgo prediction accuracy, the training dataset and independent-testing dataset were grouped according to the distance between the TLS and 35S enhancer (Figure 4). In cross-validation, Acc was 99.3%, and predictions for only two genes were incorrect (Table 5); these two genes were 10–15 kb away from the 35S enhancer. In independent-testing, the prediction accuracy for genes within 20 kb from the 35S enhancer was >84%. For genes located >20 kb from the 35S enhancer, the prediction accuracy decreased with increasing distance, but still was >60% (Table 6).

## Discussion

### *Differences of the Framework between TIMgo and EAT-Rice*

In a previous study, the PROMOTER region for most genes was defined as the upstream region from the transcription start site (TSS)[23]. For the EAT-Rice analysis, we collected gene data that only include the information of TLS, however, the promoter region usually includes the upstream sequence of the TSS, which was based on a 1000-bp region upstream of the TLS. Upstream sequence of the TSS contain the 5' untranslated region of the mRNA and sequences downstream of the TSS may also be involved with transcription factor regulation of gene expression [24]. Therefore, we considered an average length of 500 bp for 5' untranslated regions in rice and the 1000-bp upstream of the TSS as the condition. We used the 1500-bp sequence upstream of the TLS as the PROMOTER in this study.

For our prediction models, we retained the EAT-Rice CGIs and DNP (dinucleotide conformation and physicochemical properties encoding) and increased the TNP coding with the DNP coding concept. We also used the Pse-in-One tool to generate code for DACC, TACC and PseKNC. Given the strand specificity of Kmer, we added RevKmer coding, and the Motif coding of the PROMOTER region was combined with Kmer and with RevKmer. The ranges of overall predictive accuracy for Kmer + Motif and RevKmer + Motif models were small, which indicated that Motif was complementary with Kmer and RevKmer, the combination of these two features could improve the classification ability. Predictive accuracy increased with the length of $k$ for both Kmer and RevKmer, because that Motif feature consisted of experimentally validated regulatory sequences, but the

number of proven regulatory sequences in plants is limited, whereas Kmer and RevKmer considered all the sequence combinations that provided higher data integrity than Motif, so using longer Kmer and RevKmer should lead to better prediction performance. Although Kmer and RevKmer had higher data integrity than Motif, the complexity of the Kmer and RevKmer data increased exponentially with the increase in sequence length, resulting in processing time that was too lengthy. Therefore, we used Kmer (and RevKmer) with limited k length and retained Motif with longer sequences, to preserve important regulatory sequence data and reduce the computational complexity significantly.

***Specific Regulatory Sequences within Genes Activated by the 35S Enhancer***

To find out whether a specific regulatory sequence was related to gene activation in the T-DNA insertion mutants, we analyzed the 2087 motifs with a t-test. We found that there were 181 regulatory sequences that had significant difference in their occurrence frequency between Ac and NAc genes. Among these 181 regulatory sequences, 20 were G-box and G-box related sequences. The G-box contains a core region, CACGTG, and flanking sequences that are composed of other nucleotides. The G-box binding protein has different binding preferences and affinities according to the different flanking sequences in the G-box. bZip (basic region/leucine zipper) transcription factors account for the majority of G-box binding proteins. Transcription regulation in plants is often affected by G-box sequences, such as stress hormones (e.g., abscisic acid), seed germination, protein storage and the light response [25-27]. Thus, the G-box may have important biological significance in the regulation of gene expression by the 35S enhancer and may affect whether the 35S enhancer will activate a target gene in rice.

***Correlation between Length of Sequence and Nucleotide Length Parameter***

In the feature coding of TIMgo, the coding of Kmer, RevKmer and PseKNC can be adjusted based on the nucleotide length parameter (*k*). We needed to find a suitable nucleotide length parameter for encoding. For these three kinds of coding, the *k* value selected for the PROMOTER region was greater than that for the MIDDLE region. A higher value for *k* results in a higher number of features being generated, which requires that more features need to be improved to increase the predictive accuracy of the PROMOTER region, relative to the MIDDLE region. Thus, an excessive number of features would reduce the predictive performance of the model. From the optimal *k* value for the MIDDLE sequence, we could see that a higher number of features did not necessarily make the classification better. By comparing the optimal *k* value selected for the PROMOTER and MIDDLE region, we note that a longer sequence does seem to require more features to make the classification better. Moreover, among the local, global and local + global sequence characteristics used to build the TIMgo, the local sequences had a greater contribution with respect to identifying activation of the target genes (Table 4).

***Performance Comparison of TIMgo and EAT-Rice***

To confirm that the model constructed by the framework of TIMgo is superior to that of EAT-Rice, the training dataset and testing dataset used to develop EAT-Rice were used to build models in the TIMgo framework and to evaluate TIMgo for comparing their predictive performance. The training dataset used with EAT-Rice had data for 280 validated genes, and these 280 data points were separated into two subsets (subset1, subset2) with 180 validated genes [15]. The independent-testing dataset used with EAT-Rice had 48 validated genes. Two training datasets (subset1 and subset2) were used to build training models within the framework of TIMgo, and the predictive efficacy of EAT-Rice and TIMgo was evaluated with an independent-testing dataset consisting of an additional 48 validated genes (Table 7). Using subset1 as the training dataset and using the EAT-Rice system to establish the model, the Acc in the independent-testing was 72.9%; the Acc for TIMgo was 79.2%, and the Sp value of TIMgo was 12.8% higher than that of EAT-Rice. Using subset2 as the training dataset, the Acc with independent-testing was 77.1% for EAT-Rice and 77.6% for TIMgo. In the case of using

the same training dataset and testing dataset, the accuracy of the TIMgo framework is better than EAT-Rice.

## Conclusions

In this study, we analyzed the DNA sequence and constructed a two-layer model system using the machine learning method to predict whether the 35S enhancer would affect the expression of a target gene in T-DNA insertion mutants. The first layer of the system was built with the PROMOTER and MIDDLE sequences and was encoded using nine features. We analyzed significant sequence fragments in Motif, Kmer and RevKmer and weighted the PROMOTER based on a logistic regression analysis of the distance between the 35S enhancer and the TLS of each gene. Some of the first-layer SVM models were built with LIBSVM feature selection. The second-layer model used the mRMR feature selection tool to select the predicted values from the 16 models in the first layer, and these were integrated with the LADTree algorithm as the second-layer model. The predictive performance of TIMgo had Acc of 99.3% with cross-validation and of 85.6% with independent-testing, and the predictive efficiency of TIMgo was better than that of EAT-Rice. TIMgo can more accurately predict the activation of genes located within 20 kb of the 35S enhancer. We analyzed the 2087 motifs and found that there was a significant difference in the frequency of G-box sequences between Ac and NAc genes, suggesting that the G-box may play an important role in the mechanism of 35S enhancer–activation of genes. Our model has improved the predictive ability of determining target gene activation based on the CaMV 35S enhancer in rice T-DNA insertion mutants. This system could help researchers to improve their efficiency when screening rice genes.

## List of abbreviations

cauliflower mosaic virus: CaMV; transcription start site: TSS; translation start site: TLS; Di- or Trinucleotide conformation and physicochemical properties: DNP/TNP; non-activated: NAc; activated: Ac; Rice Genome Automated Annotation System: RiceGAAS; MSU Rice Genome Annotation Project: TIGR; Di- or trinucleotide-based auto-cross covariance: DACC/TACC; CpG-islands: CGIs; Regulatory *cis*-elements: Motif; Transcription factor binding sites: TFBSs; Reverse Complementary Kmer: RevKmer; Pseudo K-Tuple Nucleotide Composition: PseKNC; Accuracy: Acc; Matthews Correlation Coefficient: MCC; Sensitivity: Sn; Specificity: Sp; Support Vector Machine: SVM; Minimum redundancy maximum relevance feature selection: mRMR

## Ethics approval and consent to participate

Not applicable

## Consent for publication

Not applicable

## Availability of data and materials

Data cannot be shared publicly because of this data contain each gene expressed information are from the C4 Rice Project – is founded by Bill & Melinda Gates Foundation, which is not published yet, so we only provide the gene sequence for interest researcher to confirm. It can be obtained in supplementary material (Please see the additional file 2 of "Additional Material").

## Competing interests

The authors declare no conflict of interest.

## Funding

This research was funded by the following: (a) Ministry of Science and Technology, Taiwan, R.O.C. under grant numbers 106-2221-E-005-077-MY2, 107-2634-F-005-002 and 107-2321-B-005-013. This funding supported the salaries of research assistants and graduate students. (b) National Chung Hsing University and Chung-Shan Medical University under grant number NCHU-CSMU-10705. This funding was used to purchase research

equipment and office supplies for the printing fee. (c) Advanced Plant Biotechnology Center from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. This funding was used for English editing and publication fees. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Author Contributions

C.-C.L. and J.-J.C. contributed to data collection, design of experimental processes and system architecture, and C.-W.C. provided expertise on machine learning. C.-C.L. drafted the manuscript. L.-J.C. and S.-F.L. supported the experimental data and data interpretation. L.-J.C. and Y.-W.C. conceived of the study goal, supervised the study and provided advice with respect to the study direction. All authors read and approved the manuscript.

## Acknowledgements

Not applicable

## Additional Material

**Additional file 1: Supplementary Material, including additional tables and figures**
**Additional file 2: Sequences Datasets for Training and Testing.** Gene sequences were stored in a FASTA format, that include 2 types of sequences, PROMOTER and MIDDLE, in train dataset and testing dataset.

## References

1. Ray DK, Mueller ND, West PC, Foley JA: **Yield Trends Are Insufficient to Double Global Crop Production by 2050**. *PLoS One* 2013, **8**(6):e66428.
2. IRGSP: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.
3. Jeong DH, An S, Kang HG, Moon S, Han JJ, Park S, Lee HS, An K, An G: **T-DNA insertional mutagenesis for activation tagging in rice**. *Plant Physiol* 2002, **130**(4):1636-1644.
4. Yang Y, Li Y, Wu C: **Genomic resources for functional analyses of the rice genome**. *Curr Opin Plant Biol* 2013, **16**(2):157-163.
5. Hsing YI, Chern CG, Fan MJ, Lu PC, Chen KT, Lo SF, Sun PK, Ho SL, Lee KW, Wang YC *et al*: **A rice gene activation/knockout mutant resource for high throughput functional genomics**. *Plant Mol Biol* 2007, **63**(3):351-364.
6. Odell JT, Nagy F, Chua NH: **Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter**. *Nature* 1985, **313**(6005):810-812.
7. Fang RX, Nagy F, Sivasubramaniam S, Chua NH: **Multiple cis regulatory elements for maximal expression of the cauliflower mosaic virus 35S promoter in transgenic plants**. *Plant Cell* 1989, **1**(1):141-150.
8. Kardailsky I, Shukla VK, Ahn JH, Dagenais N, Christensen SK, Nguyen JT, Chory J, Harrison MJ, Weigel D: **Activation tagging of the floral inducer FT**. *Science* 1999, **286**(5446):1962-1965.
9. Weigel D, Ahn JH, Blazquez MA, Borevitz JO, Christensen SK, Fankhauser C, Ferrandiz C, Kardailsky I, Malancharuvil EJ, Neff MM *et al*: **Activation tagging in Arabidopsis**. *Plant Physiol* 2000, **122**(4):1003-1013.
10. Huang S, Cerny RE, Bhat DS, Brown SM: **Cloning of an Arabidopsis patatin-like gene, STURDY, by activation T-DNA tagging**. *Plant Physiol* 2001, **125**(2):573-584.
11. Ichikawa T, Nakazawa M, Kawashima M, Muto S, Gohda K, Suzuki K, Ishikawa A, Kobayashi H, Yoshizumi T, Tsumoto Y *et al*: **Sequence database of 1172 T-DNA insertion sites in Arabidopsis activation-tagging lines that showed phenotypes in T1 generation**. *Plant J* 2003, **36**(3):421-429.

12. Jeon JS, Lee S, Jung KH, Jun SH, Jeong DH, Lee J, Kim C, Jang S, Yang K, Nam J *et al*: **T-DNA insertional mutagenesis for functional genomics in rice**. *Plant J* 2000, **22**(6):561-570.
13. Li G, Ruan X, Auerbach Raymond K, Sandhu Kuljeet S, Zheng M, Wang P, Poh Huay M, Goh Y, Lim J, Zhang J *et al*: **Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation**. *Cell* 2012, **148**(1–2):84-98.
14. Ren S, Johnston JS, Shippen DE, McKnight TD: **TELOMERASE ACTIVATOR1 induces telomerase activity and potentiates responses to auxin in Arabidopsis**. *Plant Cell* 2004, **16**(11):2910-2922.
15. Liao CC, Chen LJ, Lo SF, Chen CW, Chu YW: **EAT-Rice: A predictive model for flanking gene expression of T-DNA insertion activation-tagged rice mutants by machine learning approaches**. *Plos Comput Biol* 2019, **15**(5):e1006942.
16. Chang CC, Lin CJ: **LIBSVM: A Library for Support Vector Machines**. *Acm T Intel Syst Tec* 2011, **2**(3).
17. Boros E, Crama Y, Hammer PL, Ibaraki T, Kogan A, Makino K: **Logical analysis of data: classification with justification**. *Ann Oper Res* 2011, **188**(1):33-61.
18. Peng H, Long F, Ding C: **Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy**. *IEEE Trans Pattern Anal Mach Intell* 2005, **27**(8):1226-1238.
19. vanderGeest AHM, Hall TC: **The beta-phaseolin 5' matrix attachment region acts as an enhancer facilitator**. *Plant Mol Biol* 1997, **33**(3):553-557.
20. Jagannath A, Bandyopadhyay P, Arumugam N, Gupta V, Burma PK, Pental D: **The use of a Spacer DNA fragment insulates the tissue-specific expression of a cytotoxic gene (barnase) and allows high-frequency generation of transgenic male sterile lines in Brassica juncea L.** *Mol Breeding* 2001, **8**(1):11-23.
21. Marenduzzo D, Faro-Trindade I, Cook PR: **What are the molecular ties that maintain genomic loops?** *Trends in Genetics* 2007, **23**(3):126-133.
22. Holmes G, Donkin A, Witten IH: **Weka: A machine learning workbench**. In: *Intelligent Information Systems, 1994 Proceedings of the 1994 Second Australian and New Zealand Conference on: 1994*. IEEE: 357-361.
23. Chang WC, Lee TY, Huang HD, Huang HY, Pan RL: **PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups**. *BMC Genomics* 2008, **9**:561.
24. Heyndrickx KS, Van de Velde J, Wang C, Weigel D, Vandepoele K: **A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana**. *Plant Cell* 2014, **26**(10):3894-3910.
25. Marcotte WR, Jr., Russell SH, Quatrano RS: **Abscisic acid-responsive sequences from the em gene of wheat**. *Plant Cell* 1989, **1**(10):969-976.
26. Mason HS, DeWald DB, Mullet JE: **Identification of a methyl jasmonate-responsive domain in the soybean vspB promoter**. *Plant Cell* 1993, **5**(3):241-251.
27. Donald RG, Cashmore AR: **Mutation of either G box or I box sequences profoundly affects expression from the Arabidopsis rbcS-1A promoter**. *EMBO J* 1990, **9**(6):1717-1726.
28. Ohan NW, Heikkila JJ: **Reverse transcription-polymerase chain reaction: an overview of the technique and its applications**. *Biotechnol Adv* 1993, **11**(1):13-29.
29. Akbani R, Kwek S, Japkowicz N: **Applying support vector machines to imbalanced datasets**. *Lect Notes Comput Sc* 2004, **3201**:39-50.
30. Sakata K, Nagamura Y, Numa H, Antonio BA, Nagasaki H, Idonuma A, Watanabe W, Shimizu Y, Horiuchi I, Matsumoto T *et al*: **RiceGAAS: an automated annotation system and database for rice genome sequence**. *Nucleic acids research* 2002, **30**(1):98-102.

31. Yuan Q, Ouyang S, Liu J, Suh B, Cheung F, Sultana R, Lee D, Quackenbush J, Buell CR: **The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists**. *Nucleic acids research* 2003, **31**(1):229-233.
32. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L *et al*: **The TIGR Rice Genome Annotation Resource: improvements and new features**. *Nucleic acids research* 2007, **35**(Database issue):D883-887.
33. Singer SD, Cox KD, Liu ZR: **Both the constitutive Cauliflower Mosaic Virus 35S and tissue-specific AGAMOUS enhancers activate transcription autonomously in Arabidopsis thaliana**. *Plant Mol Biol* 2010, **74**(3):293-305.
34. Singer SD, Cox KD, Liu Z: **Enhancer-promoter interference and its prevention in transgenic plants**. *Plant cell reports* 2011, **30**(5):723-731.
35. Noble WS, Kuehn S, Thurman R, Yu M, Stamatoyannopoulos J: **Predicting the in vivo signature of human gene regulatory sequences**. *Bioinformatics* 2005, **21 Suppl 1**:i338-343.
36. Gupta S, Dennis J, Thurman RE, Kingston R, Stamatoyannopoulos JA, Noble WS: **Predicting human nucleosome occupancy from primary sequence**. *Plos Comput Biol* 2008, **4**(8):e1000134.
37. Lee D, Karchin R, Beer MA: **Discriminative prediction of mammalian enhancers from DNA sequence**. *Genome Res* 2011, **21**(12):2167-2180.
38. Guo Y, Yu L, Wen Z, Li M: **Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences**. *Nucleic acids research* 2008, **36**(9):3025-3030.
39. Lin H, Deng EZ, Ding H, Chen W, Chou KC: **iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition**. *Nucleic acids research* 2014, **42**(21):12961-12972.
40. Volpe TA, Kidner C, Hall IM, Teng G, Grewal SIS, Martienssen RA: **Regulation of Heterochromatic Silencing and Histone H3 Lysine-9 Methylation by RNAi**. *Science* 2002, **297**(5588):1833-1837.
41. Antequera F, Boyes J, Bird A: **High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines**. *Cell* 1990, **62**(3):503-514.
42. Higo K, Ugawa Y, Iwamoto M, Korenaga T: **Plant cis-acting regulatory DNA elements (PLACE) database: 1999**. *Nucleic acids research* 1999, **27**(1):297-300.
43. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif**. *Bioinformatics* 2011, **27**(7):1017-1018.
44. Beer MA, Tavazoie S: **Predicting gene expression from sequence**. *Cell* 2004, **117**(2):185-198.
45. Yuan Y, Guo L, Shen L, Liu JS: **Predicting gene expression from sequence: a reexamination**. *Plos Comput Biol* 2007, **3**(11):e243.
46. Stewart AJ, Hannenhalli S, Plotkin JB: **Why transcription factor binding sites are ten nucleotides long**. *Genetics* 2012, **192**(3):973-985.
47. Friedel M, Nikolajewa S, Suhnel J, Wilhelm T: **DiProDB: a database for dinucleotide properties**. *Nucleic acids research* 2009, **37**(Database issue):D37-40.
48. Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou K-C: **PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions**. *Bioinformatics* 2015, **31**(1):119-120.
49. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC: **Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences**. *Nucleic acids research* 2015, **43**(W1):W65-71.

50. Hanchuan P, Fuhui L, Ding C: **Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2005, **27**(8):1226-1238.

## Figures

**Figure 1. Flowchart of the TIMgo predictive system.** TIMgo was built in two-layer model, primary module included nine feature-encoding and secondary modules was integrated nine results of primary modules. The red dash line indicates the system core architecture.

**Figure 2. Correlation between distance and gene activation.** The data were sorted by the distance between the 35S enhancer and the TLS, and the ratio of Ac to NAc genes in each group was calculated. The *x* axis is the distance from the 35S enhancer to the TLS of a target gene; the *y* axis is the proportion of Ac and NAc genes in each group.

**Figure 3. Accuracy trend in the second-layer feature selection.** The x-axis represents how many features the models used, and y-axis represents the accuracy of model had been built with some of features. In this study, nine feature selection methods were used.

**Figure 4. Accuracy trend of TIMgo for cross-validation and independent-testing of data within different distances.** Train represents the Acc from 5-fold cross validation with D299; Test represents the Acc from independent testing with D153. The *x* axis indicates each distance interval, and the *y* axis indicates the predictive accuracy.

## Tables

**Table 1. Data distribution of flanking analyzed genes in rice T-DNA mutants**

| Data source | Number of mutant lines | Gene expression status | | | | Validated genes[a] |
|---|---|---|---|---|---|---|
| | | Ac | NE | ND | Ko | |
| **NCHU**[b] | 11 | 26 | 22 | 17 | 0 | 65 |
| **Academia-Sinica**[c] | 316 | 262 | 143 | 13 | 2 | 420 |
| **Total** | 327 | 288 | 165 | 30 | 2 | 485 |

[a] Validated genes indicate the target genes that were detected by RT-PCR.
[b] NCHU, experimental data were collected from Liang-Jwu Chen's laboratory.
[c] Academia-Sinica, experimental data were collected by Su-May Yu's research team.
Ac: activated gene; NE: non-activated gene; ND: non-detectable gene; Ko: knockout gene

**Table 2. Data distribution of the training dataset and independent-testing datase**

| Data sources | Training dataset (D300) | | Testing dataset (D153) | |
|---|---|---|---|---|
| | Ac | NAc | Ac | NAc |
| **NCHU** | 20 | 20 | 6 | 2 |
| **Academia-Sinica** | 130 | 130 | 132 | 13 |
| **Total** | 150 | 150 | 138 | 15 |

**Table 3. SVM model evaluation of Kmer, RevKmer and Motif features**

| Feature | *k*[a] | Without Motif | | | | | With Motif | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sp | Sn | Acc | MCC | AUC | Sp | Sn | Acc | MCC | AUC |
| **Kmer** | | | | | | | | | | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6 | 0.727 | 0.660 | 0.693 | 0.388 | 0.790 | 0.793 | 0.773 | 0.783 | 0.567 | 0.881 |
| | 7 | 0.867 | 0.733 | 0.800 | 0.605 | 0.891 | 0.833 | 0.787 | 0.810 | 0.621 | 0.897 |
| | 8 | 0.753 | 0.353 | 0.553 | 0.116 | 0.653 | 0.833 | 0.847 | 0.840 | 0.680 | 0.936 |
| | 9 | 0.847 | 0.853 | 0.850 | 0.700 | 0.932 | 0.867 | 0.853 | 0.860 | 0.720 | 0.937 |
| | 6 | 0.713 | 0.607 | 0.66 | 0.322 | 0.727 | 0.780 | 0.773 | 0.777 | 0.553 | 0.857 |
| **RevKmer** | 7 | 0.847 | 0.760 | 0.803 | 0.609 | 0.879 | 0.793 | 0.773 | 0.783 | 0.567 | 0.881 |
| | 8 | 0.773 | 0.327 | 0.550 | 0.112 | 0.649 | 0.840 | 0.800 | 0.820 | 0.641 | 0.915 |
| | 9 | 0.747 | 0.880 | 0.813 | 0.632 | 0.906 | 0.840 | 0.847 | 0.843 | 0.687 | 0.929 |

[a] *k* refers to the maximum *k* value used in Kmer and RevKmer, with a range of 3-*k* nucleotides in length for each analysis.

**Table 4. Performance of the first-layer features with the SVM models**

| Feature encoding | Sequence | Cross-validation | | | | | Independent-testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sp | Sn | Acc | MCC | AUC | Sp | Sn | Acc | MCC | AUC |
| **CGIs** | PROMOTER | 0.713 | 0.487 | 0.600 | 0.205 | 0.585 | 0.533 | 0.406 | 0.418 | -0.037 | 0.482 |
| | MIDDLE | 0.773 | 0.180 | 0.477 | -0.058 | 0.472 | 1.000 | 0.022 | 0.118 | 0.047 | 0.650 |
| **DNP** | PROMOTER | 0.560 | 0.647 | 0.603 | 0.207 | 0.643 | 0.267 | 0.717 | 0.673 | -0.011 | 0.451 |
| | MIDDLE | 0.593 | 0.620 | 0.607 | 0.213 | 0.600 | 0.600 | 0.536 | 0.543 | 0.081 | 0.487 |
| **TNP** | PROMOTER | 0.560 | 0.613 | 0.587 | 0.174 | 0.622 | 0.533 | 0.681 | 0.667 | 0.135 | 0.574 |
| | MIDDLE | 0.647 | 0.300 | 0.473 | -0.057 | 0.474 | 0.267 | 0.659 | 0.621 | -0.047 | 0.450 |
| **Kmer + Motif** | PROMOTER | 0.867 | 0.853 | 0.860 | 0.720 | 0.937 | 0.733 | 0.855 | 0.843 | 0.435 | 0.791 |
| **RevKmer + Motif** | PROMOTER | 0.840 | 0.847 | 0.843 | 0.687 | 0.929 | 0.733 | 0.812 | 0.804 | 0.378 | 0.836 |
| **Kmer** | MIDDLE | 0.920 | 0.847 | 0.883 | 0.769 | 0.942 | 0.667 | 0.862 | 0.843 | 0.401 | 0.864 |
| **RevKmer** | MIDDLE | 0.853 | 0.727 | 0.790 | 0.585 | 0.882 | 0.533 | 0.688 | 0.673 | 0.140 | 0.665 |
| **DACC** | PROMOTER | 0.671 | 0.727 | 0.699 | 0.398 | 0.786 | 0.467 | 0.594 | 0.582 | 0.037 | 0.546 |
| | MIDDLE | 0.765 | 0.58 | 0.672 | 0.351 | 0.741 | 0.533 | 0.493 | 0.497 | 0.016 | 0.475 |
| **TACC** | PROMOTER | 0.604 | 0.580 | 0.592 | 0.184 | 0.603 | 0.133 | 0.630 | 0.582 | -0.148 | 0.416 |
| | MIDDLE | 0.597 | 0.567 | 0.582 | 0.164 | 0.578 | 0.467 | 0.457 | 0.458 | -0.046 | 0.451 |
| **PseKNC** | PROMOTER | 0.899 | 0.607 | 0.753 | 0.529 | 0.845 | 0.733 | 0.543 | 0.562 | 0.165 | 0.591 |
| | MIDDLE | 0.564 | 0.527 | 0.595 | 0.191 | 0.617 | 0.667 | 0.580 | 0.588 | 0.147 | 0.545 |

**Table 5. Performance of the LADTree model in the second-layer**

| | TP | FP | TN | FN | Sn | Sp | Acc | MCC |
|---|---|---|---|---|---|---|---|---|
| **Cross-validation** | 149 | 1 | 148 | 1 | 0.993 | 0.993 | 0.993 | 0.987 |
| **Independent-testing** | 123 | 7 | 8 | 15 | 0.891 | 0.533 | 0.856 | 0.353 |

**Table 6. Predictive accuracy of TIMgo for different distance groups**

| | Distance from the 35S enhancer (kb) | | | | | | |
|---|---|---|---|---|---|---|---|
| **Dataset** | 0-2 | 2-5 | 5-10 | 10-15 | 15-20 | 20-25 | >25 |
| **Training set** | 1.00 | 1.00 | 1.00 | 0.97 | 1.00 | 1.00 | 1.00 |
| **Testing set** | 0.89 | 0.91 | 0.84 | 0.86 | 0.93 | 0.71 | 0.60 |

**Table 7. Comparison of TIMgo and EAT-Rice with independent-testing evaluation**

| System | Subset1 | | | | Subset2 | | | |
|---|---|---|---|---|---|---|---|---|
| | Sp | Sn | Acc | AUC | Sp | Sn | Acc | AUC |
| **EAT-Rice** | 0.591 | 0.846 | 0.729 | 0.794 | 0.591 | 0.923 | 0.771 | 0.832 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **TIMgo** | 0.727 | 0.846 | 0.792 | 0.874 | 0.783 | 0.767 | 0.776 | 0.844 |

# Figures

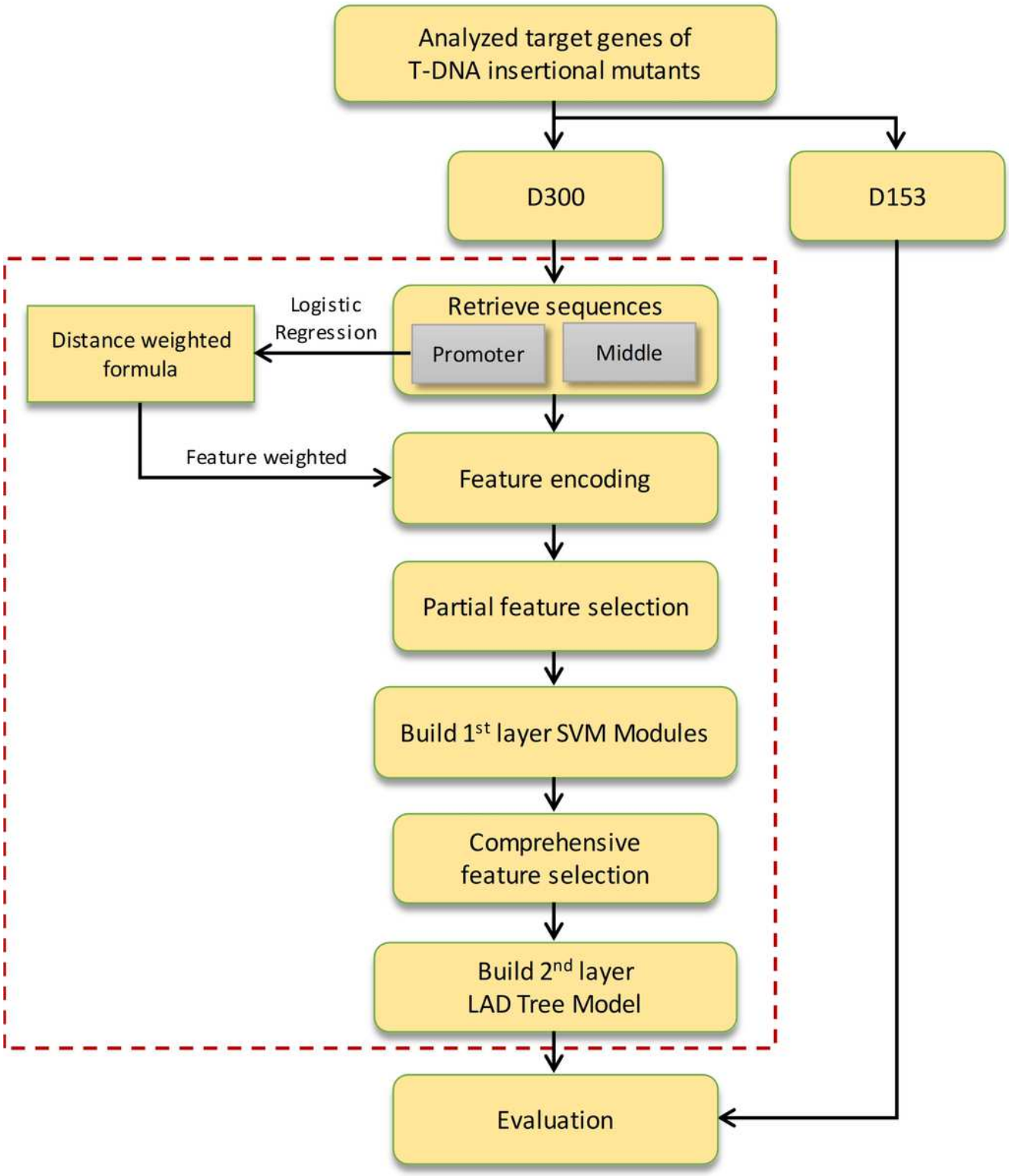


**Figure 1**

Flowchart of the TIMgo predictive system. TIMgo was built in two-layer model, primary module included nine feature-encoding and secondary modules was integrated nine results of primary modules. The red dash line indicates the system core architecture.

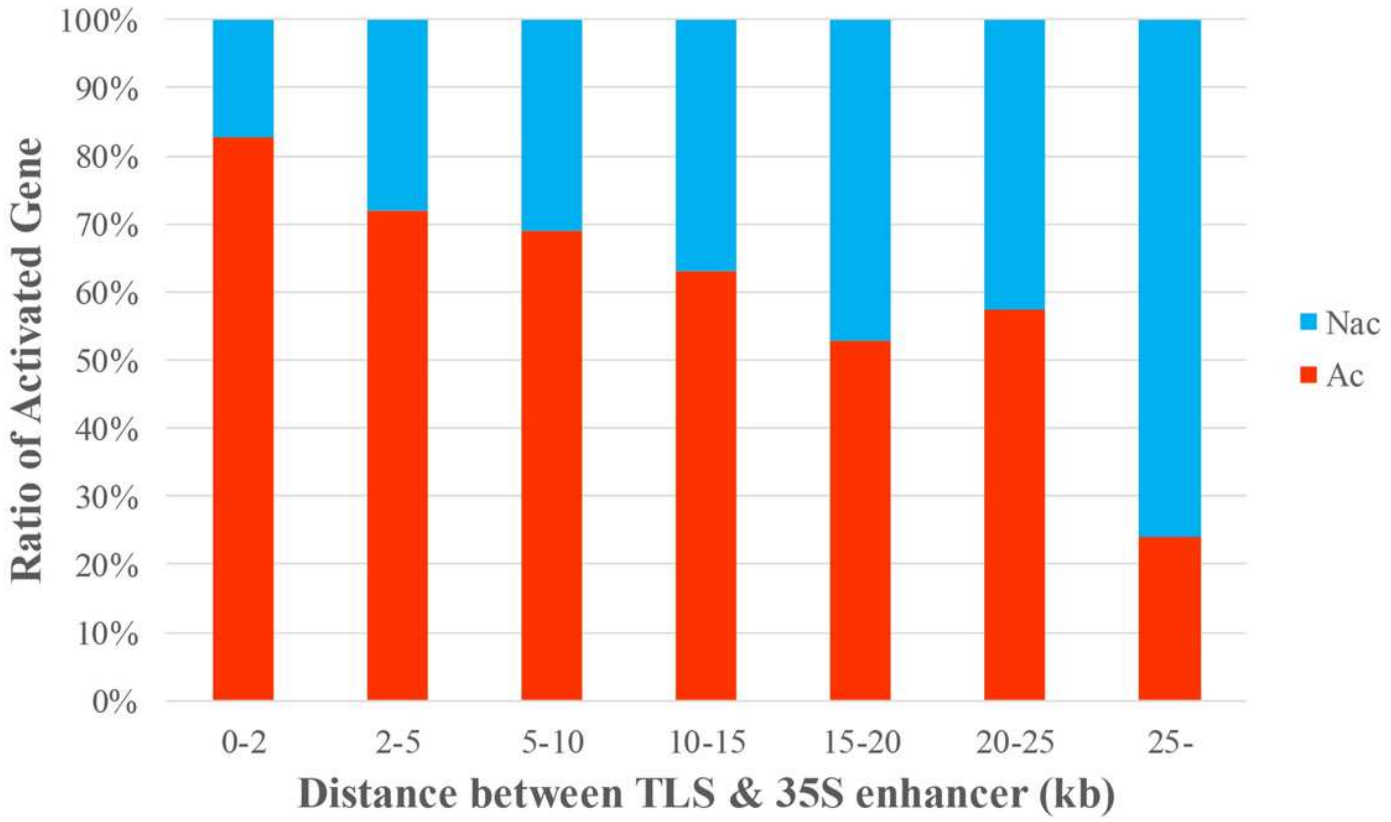


**Figure 2**

Correlation between distance and gene activation. The data were sorted by the distance between the 35S enhancer and the TLS, and the ratio of Ac to NAc genes in each group was calculated. The x axis is the distance from the 35S enhancer to the TLS of a target gene; the y axis is the proportion of Ac and NAc genes in each group.

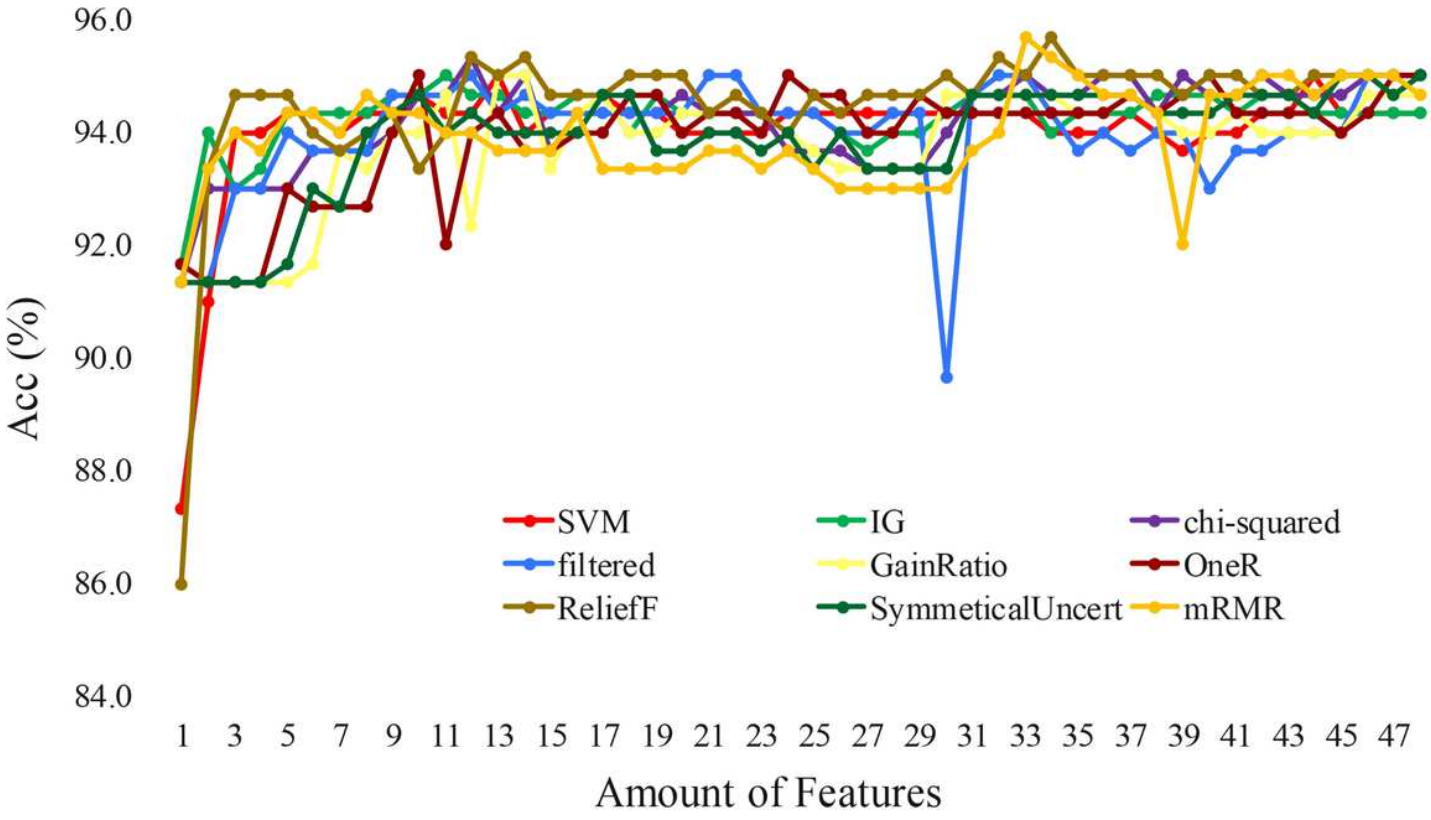


**Figure 3**

Accuracy trend in the second-layer feature selection. The x-axis represents how many features the models used, and y-axis represents the accuracy of model had been built with some of features. In this study, nine feature selection methods were used.

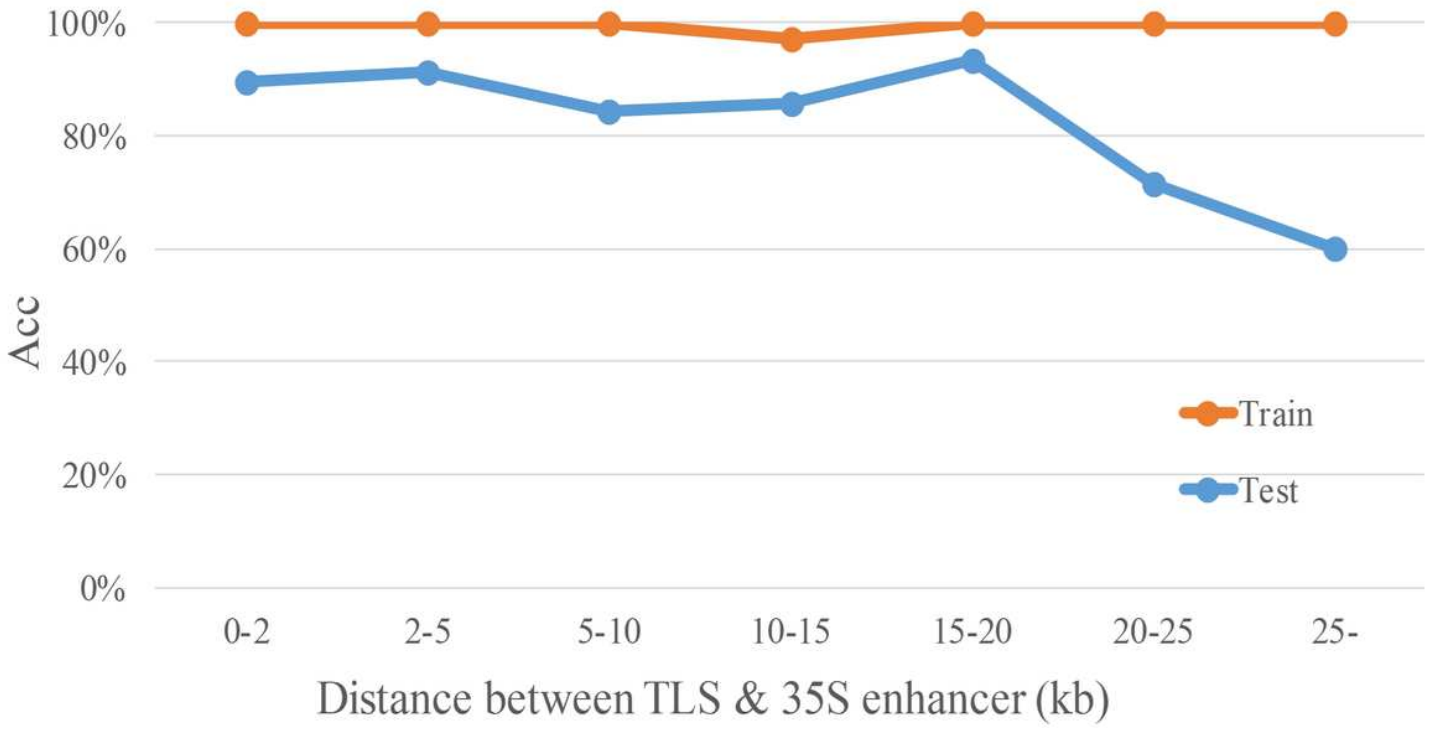


**Figure 4**

Accuracy trend of TIMgo for cross-validation and independent-testing of data within different distances. Train represents the Acc from 5-fold cross validation with D299; Test represents the Acc from independent testing with D153. The x axis indicates each distance interval, and the y axis indicates the predictive accuracy.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TIMgoSupplementary.docx
- SupplementaryFiles.rar