OPEN ACCESS Freely available online

PLOS ONE

Comparative Analyses between Retained Introns and Constitutively Spliced Introns in *Arabidopsis thaliana* Using Random Forest and Support Vector Machine

Rui Mao[1,2,3], Praveen Kumar Raj Kumar[3], Cheng Guo[3], Yang Zhang[1,2]*, Chun Liang[3,4]*

1 College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi, China, **2** College of Information Engineering, Northwest A&F University, Yangling, Shaanxi, China, **3** Department of Biology, Miami University, Oxford, Ohio, United States of America, **4** Department of Computer Sciences and Software Engineering, Miami University, Oxford, Ohio, United States of America

Abstract

One of the important modes of pre-mRNA post-transcriptional modification is alternative splicing. Alternative splicing allows creation of many distinct mature mRNA transcripts from a single gene by utilizing different splice sites. In plants like *Arabidopsis thaliana*, the most common type of alternative splicing is intron retention. Many studies in the past focus on positional distribution of retained introns (RIs) among different genic regions and their expression regulations, while little systematic classification of RIs from constitutively spliced introns (CSIs) has been conducted using machine learning approaches. We used random forest and support vector machine (SVM) with radial basis kernel function (RBF) to differentiate these two types of introns in *Arabidopsis*. By comparing coordinates of introns of all annotated mRNAs from TAIR10, we obtained our high-quality experimental data. To distinguish RIs from CSIs, We investigated the unique characteristics of RIs in comparison with CSIs and finally extracted 37 quantitative features: local and global nucleotide sequence features of introns, frequent motifs, the signal strength of splice sites, and the similarity between sequences of introns and their flanking regions. We demonstrated that our proposed feature extraction approach was more accurate in effectively classifying RIs from CSIs in comparison with other four approaches. The optimal penalty parameter C and the RBF kernel parameter γ in SVM were set based on particle swarm optimization algorithm (PSOSVM). Our classification performance showed F-Measure of 80.8% (random forest) and 77.4% (PSOSVM). Not only the basic sequence features and positional distribution characteristics of RIs were obtained, but also putative regulatory motifs in intron splicing were predicted based on our feature extraction approach. Clearly, our study will facilitate a better understanding of underlying mechanisms involved in intron retention.

Citation: Mao R, Raj Kumar PK, Guo C, Zhang Y, Liang C (2014) Comparative Analyses between Retained Introns and Constitutively Spliced Introns in *Arabidopsis thaliana* Using Random Forest and Support Vector Machine. PLoS ONE 9(8): e104049. doi:10.1371/journal.pone.0104049

Editor: Yi Xing, University of California, Los Angeles, United States of America

Received February 24, 2014; **Accepted** July 6, 2014; **Published** August 11, 2014

Funding: This research was partially supported by China Scholarship Council (Award #201206305024 to RM), Biology Department and Office for the Advancement of Research and Scholarship (OARS) of Miami University in Ohio, and NIGMS (1R15GM094732-01A1 to CL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: zhangyang@nwsuaf.edu.cn (YZ); liangc@miamioh.edu (CL)

Introduction

As an essential post-transcriptional process, alternative splicing (AS) can increase transcriptome plasticity and protein diversity [1]. There are primarily three types of AS: intron retention, exon skipping, and alternative choices of 5′ and 3′ splice sites (5′ss and 3′ss, respectively) of introns [2]. The frequency and types of AS differ significantly between vertebrates and invertebrates [3]. For example, only ~19% of multi-exon genes are alternatively spliced in fruit fly, while it is ~95% in human [4,5]. In vertebrates and especially mammals, most alternatively spliced genes possess exons that are entirely spliced out or truncated, and intron retention is the least prevalent form of AS [6–8]. In invertebrates and plants, in contrast, more introns have their retention in mature mRNAs [3,7,9,10]. A recent genome-wide study in *Arabidopsis* reports that ~42% of the multi-exon genes undergo AS with ~40% of those genes having retained introns (RIs) but only 3% having spliced exons [11]. Furthermore, it is likely that the number of AS genes identified in plants will keep increasing with the increased number of tissue-specific transcriptome studies. *Syed et al.* [12] reports that the AS events being found have risen from 1.2% to 61% over the past decade in *Arabidopsis*. Accumulating evidence indicates alternative splicing in invertebrates and plants might have different mechanisms in comparison with vertebrates and especially mammals, and the extent and complexity of intron retention in plants still need to be specifically characterized.

Transcript samples with RIs that are examined by RT-PCR are shown to co-purify with polyribosomes, suggesting that these intron retention events are not the result from incomplete splicing but are found in their nuclear exports [13]. Some researches show that specific abiotic stresses can impact on RIs. By analyzing the splicing process of a cold-regulated gene encoding ribokinase (7H8) protein, *Mastrangelo et al.* [14] suggests that 7H8 cold-dependent intron retention is a general trait in cereals. *Palusa et al.* [15] reports that various abiotic stresses affect the splicing pattern of serine/arginine-rich (SR) genes in *Arabidopsis*. On the other hand, there are many studies indicating that intron retention is a major AS phenomenon in plants [13,16,17], most of which

concentrate on the positional distribution of RIs in 3′ UTR, 5′ UTR and CDS regions. However, it still lacks research on characterization, comparison and prediction of two types of introns using large amount of data by machine learning approaches in plants. Therefore, further works are required to deepen our understanding of RIs and unravel the underlying molecular and biological mechanisms.

Machine learning approaches have been widely applied to knowledge extraction from biological experimental data [18]. For classification of various problems in the domain of bioinformatics, prior studies suggest that SVM outperform k-nearest neighbors, neural networks and decision trees [19–21]. In SVM applications, the radial basis kernel function (RBF) that has only one kernel parameter γ is widely adopted [22]. Unlike the linear kernel, it can handle data with nonlinear relations between class labels and features [23]. Only under certain parameters, the sigmoid kernel is valid and demonstrated to behave like RBF [24]. Additionally, the polynomial kernel has more kernel parameters and demands more training time than RBF, and it can easily fall into numerical difficulties with the degree increase [23]. Therefore, RBF is selected and used in our study. In the SVM training procedure with RBF kernel, both γ and the penalty parameter C settings are shown to significantly influence the classification accuracy [25]. Particle swarm optimization (PSO), a meta-heuristic optimization algorithm that simulates the social behavior of bird flocking or fish schooling [26], proves to be an appropriate approach in finding better parameters of SVM [27]. On the other hand, random forest has been reported as another competitive classification algorithm and received increasing interests [28,29]. After surveys of random forest applications in bioinformatics for the recent decade, *Boulesteix et al*. [30] summarizes that random forest offers attractive features such as direct handling of high-dimensional data and advantages in parameters selection. Especially compared with SVM, it is easier for random forest to obtain excellent performance using the default parameterization without tuning parameters in general [31,32]. Recent works show that random forest classifiers obtain better performance comparable to SVM in some bioinformatics applications including classification of cancer microarray data [33], identification of DNA-binding proteins [34], and prediction of miRNA targets [35].

Using random forest and in-house implemented PSOSVM that utilizes PSO to optimize parameters C and γ of SVM, our study was set up to detect systematically the differences between two types of introns, and characterize and categorize them accurately. Our proposed feature extraction approach is novel and hybrid, including three aspects: basic intron sequence features; frequent short linear sequence motifs; and features extracted from splice sites and the flanking sequences of introns. In our study, performances of random forest and PSOSVM to classify RIs and CSIs were analyzed and compared, and the results of classification based on different feature sets suggested that our feature extraction approach had a distinct advantage.

Materials and Methods

Dataset

RIs are defined if the introns are spliced out in at least one isoform (mRNA) but entirely retained in at least one other isoform for the same genes. In addition, for multiple RIs founded in different isoforms of the same genes, if the differences in the 5′ splice sites (or the 3′ splice sites) of these RIs are less than 6 bp, we define these RIs as redundant ones. Hence the longest one is selected among them for downstream data analysis. CSIs are defined as ones that are always spliced out in all isoforms of individual genes.

Based on TAIR10 gene annotation, coordinates of introns in genome sequences were determined using TAIR10_GFF3_genes.gff (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/) by a Perl script. Then using GMAP [36], we extracted RIs and CSIs sequences, splice sites and flanking exons sequences of introns in *Arabidopsis* from the genome sequence files (ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/). R *quantile()* function was employed to generate intron length quantiles for analyzing the intron length distribution in *Arabidopsis*.

Feature extraction approach

Our new hybrid feature extraction approach combines the following three aspects:

(A) Basic features extraction. On one hand, we consider some of the most common global features of nucleotide sequences, such as intron length, nucleotide occurrence probabilities of A, C, G and T in introns, AT content and GC content. On the other hand, we determine local features of segmental nucleotides composition [37], which provide crucial complementary to the global features and are defined as **segmental probabilities of four nucleotides correlation factors** (θ_{AG},θ_{AC},θ_{AT},θ_{GC}, θ_{GT},θ_{CT}), as shown below:

For a L-length nucleotide sequence (S_L):

$$S_L = R_1R_2R_3R_4R_5R_6R_7...R_L \tag{1}$$

$$m = ceiling(L/x) \tag{2}$$

Here x is set to 20 in our work, because the length of the shortest intron sequence is 20 bp in our datasets. m is the smallest integer not less than (L/x).

S_L is divided into m sections as following:

$$\underbrace{R_1R_2...R_{20}}_{1}\underbrace{R_{21}...R_{40}}_{2}..\underbrace{R_{20m-19}...R_L}_{m}$$

Each section includes 20 bp except the last section, which includes $(L-20m+20)$ bp.

$$\left.\begin{aligned}
\theta_{AG} &= \frac{1}{m}\sum_{i=1}^{m}\left\|P_i^A - P_i^G\right\| \\
\theta_{AC} &= \frac{1}{m}\sum_{i=1}^{m}\left\|P_i^A - P_i^C\right\| \\
\theta_{AT} &= \frac{1}{m}\sum_{i=1}^{m}\left\|P_i^A - P_i^T\right\| \\
\theta_{GC} &= \frac{1}{m}\sum_{i=1}^{m}\left\|P_i^G - P_i^C\right\| \\
\theta_{GT} &= \frac{1}{m}\sum_{i=1}^{m}\left\|P_i^G - P_i^T\right\| \\
\theta_{CT} &= \frac{1}{m}\sum_{i=1}^{m}\left\|P_i^C - P_i^T\right\|
\end{aligned}\right\} \tag{3}$$

Here P_i^A,P_i^C,P_i^G,P_i^T denote probabilities of the corresponding 4 bases (A, C, G, T) in the i^{th} section respectively.

(B) Frequent motifs extraction. Because of the differences between RIs and CSIs, some subsequences appear more

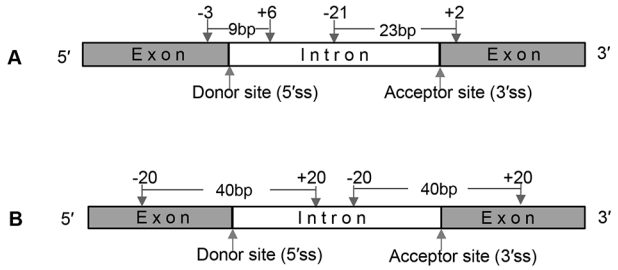

Figure 1. Feature extraction approaches for calculating signal strength of splice sites and similarity of intron and the flanking exons. A. The sequence extraction approach for calculating signal strength of splice sites; B. The sequence extraction approach for calculating increment of diversity (ID).
doi:10.1371/journal.pone.0104049.g001

frequently in RIs than CSIs, or vice versa. In this paper, these motifs need to be more frequent in either RIs or CSIs but not frequently occur in both RIs and CSIs. We searched l-mer subsequences using sliding window with the step size of 1, and extracted all subsequences from 2 to 5-mer because l-mer subsequences occur with low frequencies if l is greater than 5. For example, the mean frequency of 6-mer subsequences is low (2.01E-05). In order to discover frequent motifs from the above-mentioned l-mer subsequences, evaluation indicators are required and defined as following:

$$F_{S_L}(x(l)) = \frac{T_{S_L}(x(l))}{W_{S_L}(l)} \tag{4}$$

$$W_{S_L}(l) = L - l + 1 \tag{5}$$

Here, L refer to the length of S_L (Eq. 1), $x(l)$ designates l-mer subsequence, $T_{S_L}(x(l))$ denotes the occurrence number of $x(l)$ in S_L while $W_{S_L}(l)$ denotes the number of all l-mer subsequences within S_L. So $F_{S_L}(x(l))$ means the frequency of $x(l)$ in S_L, which will be the value of feature vector if $x(l)$ is determined as a frequent motif.

Table 1. The parameter values or ranges of PSOSVM.

Parameter	Value or Range
t(the number of iterations)	10
S(the number of particles)	100
D(dimensions of particle)	2
C_1	1.49618
C_2	1.49618
w	0.7298
C	$(2^{-8}, 2^{10})$
γ	$(2^{-8}, 2^{8})$

The rule-of-thumb settings of C_1, C_2 and w are cited from [74].
doi:10.1371/journal.pone.0104049.t001

Initialize swarm (S) at random within the range of C and γ

Do {
  For each particle (C_i, γ_i) of the S
    Do {
      Evaluation the fitness of (C_i, γ_i)
      If the fitness of (C_i, γ_i) is better than the fitness of p_i^t
        Then Update p_i^t
      End If
    }
  Choose p_i^t with the best fitness among S as p_{gbest}^t
  For each particle (C_i, γ_i) of S
    Do {
      Update the v_i^t and x_i^t of each particle (C_i, γ_i) by Eq. 16
    }
}
Until termination criterion is met

Output p_{gbest}^t

Figure 2. The pseudo-code of PSOSVM. The details of Eq. 16 are illustrated in Materials and Methods.
doi:10.1371/journal.pone.0104049.g002

$$S(x(l)) = \frac{C(x(l))}{n} \tag{6}$$

Dataset ($\{S_{L_i}\}$) include n nucleotide sequences. In $\{S_{L_i}\}$, $C(x(l))$ refers to the number of sequences in which $x(l)$ is discovered. $S(x(l))$ is used to describe the confidence of $x(l)$ in $\{S_{L_i}\}$. In this paper, frequent motifs must have higher value of $S(x(l))$ in either RIs or CSIs.

$$T(x(l)) = \sum_{i=1}^{n} T_{S_{L_i}}(x(l)) \tag{7}$$

$$W(l) = \sum_{i=1}^{n} W_{S_{L_i}}(l) \tag{8}$$

$$F(x(l)) = \frac{T(x(l))}{W(l)} S(x(l)) \tag{9}$$

$T(x(l))$ denotes the occurrence number of $x(l)$ in $\{S_{L_i}\}$, and $W(l)$ denotes the total number of l-mer subsequence included in $\{S_{L_i}\}$. $F(x(l))$ represents the frequency of $x(l)$ in $\{S_{L_i}\}$.

$$\alpha(x(l)) = \frac{F^B(x(l)) - F^A(x(l))}{F^B(x(l)) + F^A(x(l))} \tag{10}$$

In Eq. 10, $F^B(x(l))$ is the frequency of $x(l)$ in dataset of CSIs, and $F^A(x(l))$ is the frequency of $x(l)$ in dataset of RIs. $\alpha(x(l))$ represents the relative difference of $x(l)$ between CSIs and RIs datasets. The positive value of $\alpha(x(l))$ means a higher frequency of $x(l)$ in CSIs than in RIs, the negative value of $\alpha(x(l))$ means the opposite case. So, we need to consider the value of $\alpha(x(l))$ and

$S(x(l))$ as a whole, and select appropriate thresholds of $\alpha(x(l))$ and $S(x(l))$ to decide frequent motifs.

(C) Splice sites and the flanking sequences of introns features extraction. To quantify the signal strength of 5′ and 3′ splice sites, we extracted 9 bases for donor sites (−3∼+6) and 23 bases for acceptor sites (−21∼+2) from introns and their flanking exons (see details in Figure 1A), and then calculated frequencies of nucleotide A, C, G and T, which were selected as the parameters of position weight matrix (PWM) [38]. The PWM is defined as following:

$$P_{ib} = f_{ib}/N \tag{11}$$

$$W_{ib} = \ln(P_{ib}/P_{0b}) \tag{12}$$

Here, P_{ib} is the position probability matrix. N is the total number of sequences in the training sets. b represents any of the four nucleotides: A, C, G, and T. f_{ib} denotes the occurrence number of b in the i^{th} position of the N aligned sequences along the splice sites. P_{0b} is equal to 0.25, and W_{ib} denotes the PWM value of b in the i^{th} position. For a n-length sequence, the PWM scoring function (SF) is defined as:

$$SF = \sum_{i=1}^{n} W_{ib} \tag{13}$$

SF denotes the quantitative value of the signal strength of splice site. The greater value of SF means the more probability of constructive splicing sites [39].

All of the sequences extracted from −20 to +20 bp at donor (acceptor) sites were separated into two datasets from splice sites (see details in Figure 1B): one exon sequences dataset and one intron sequences dataset. Increment of diversity (ID) is used to depict the similarity level of these two datasets [40]. The difference between RIs and their flanking sequences datasets (or CSIs and their flanking sequences datasets) can be quantitatively described by ID.

Let X represents d-dimensional category space X: $\{n_1,n_2,...,n_d\}$, the standard diversity measure for X is defined as:

$$D(X) = D(n_1,n_2,...,n_d) = N\ln N - \sum_{i=1}^{d} n_i \ln n_i \tag{14}$$

Here d represents the total number of trimers, n_i is the absolute frequency of the i^{th} trimer in nucleotide sequence, N is equal to $\sum_{i=1}^{d} n_i$. RIs have the similar trimer usage with the exons, which is different from CSIs where trimer frequencies are obviously different between introns and flanking exon regions [41].

For the two d-dimensional sources X: $\{n_1,n_2,...,n_d\}$ and Y: $\{m_1,m_2,...,m_d\}$, ID depicts the similarity between the X and Y. It is defined as:

$$ID(X,Y) = D(X+Y) - D(X) - D(Y) \tag{15}$$

Here $D(X+Y)$ is the measure of diversity of the mixed source $X+Y$: $\{n_1+m_1,n_2+m_2,...,n_d+m_d\}$.

By the above-mentioned feature extraction approach, the sequence information in our dataset was changed into feature vector using R codes that utilize "seqinr" package (http://cran.r-project.org/web/packages/seqinr/index.html).

Random Forest

Random forest is an ensemble classifier that consists of many independent decision trees [28]. Each tree is created by bootstrap samples of the original training data using a randomly selected subset of features [42]. At each split about 37% of the training data, named as "out of bag" (OOB) samples, is not used to construct but evaluate the performance of each classification tree [33]. The other remainder, named as "in-bag" samples, is used to construct each classification tree. Then individual trees are combined through a voting process to provide an unbiased prediction. Compared with other classification approaches such as decision tree, it possesses internal cross-validation [43] and could be more accurate and tolerant to noises [35]. The random forest algorithm is available in Weka [44].

PSOSVM

SVM classifier, as a typical 2-class classifier, is to calculate an optimal linear separating plane that separates two classes of the dataset [45]. For non-linearly separable cases, samples are mapped into a high-dimensional feature space where a separating hyper plane can be found, and proper kernel function is sought to realize this nonlinear mapping [46].

In our study we used RBF kernel. Considering two samples $x_i,x_j \in R^d (i \neq j)$, the RBF kernel is calculated using $K(x_i,x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$, where d denotes the number of dimensions of input feature vector and γ (>0) represents the width of RBF [47]. In general, the performance of SVM is determined by parameters (C, γ). The grid search algorithm is a traditional method to find the best (C, γ) [48]. However, it is difficult to obtain a satisfactory outcome because of too limited parameter pairs to search from the huge size of possible search space by applying this method. *Lin et al.* [26] introduces PSO for parameter determination and feature selection of SVM, and experimental results demonstrate that the classification accuracy of SVM optimized by PSO performs better than many other parameter optimal approaches [49].

PSO consists of particles in the population that search for the best position by following its best solution [50]. A particle is considered as a point in a D-dimension space, and its status is represented based on its position and velocity. Let $x_i^t = (x_{i1}^t, x_{i2}^t, ..., x_{iD}^t)$ and $v_i^t = (v_{i1}^t, v_{i2}^t, ..., v_{iD}^t)$ represent the D-dimensional position and velocity of particle i at iteration t respectively. Let $p_i^t = (p_{i1}^t, p_{i2}^t, ..., p_{iD}^t)$ represents the best personal

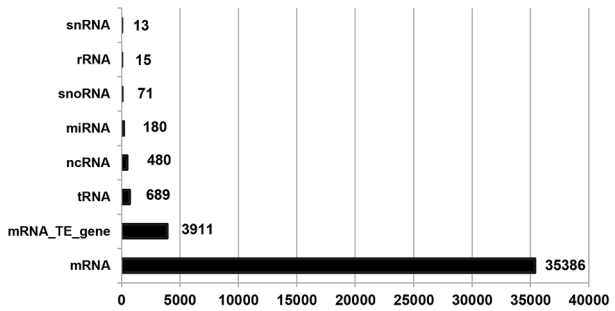

Figure 3. Numbers of various RNA types annotated in TAIR10 gene annotation for *Arabidopsis*. Each horizontal bar (with the number) indicates the number for a given RNA type.
doi:10.1371/journal.pone.0104049.g003

Table 2. Distribution of RIs and CSIs in *Arabidopsis*.

Introns Categories	RIs	CSIs
All RNAs	2,811	113,098
mRNAs	2,762	110,304
ChrC, ChrM	0	42
Chr1, Chr2, Chr3, Chr4, Chr5	2,762	110,262
Redundant Cases	229	0

All RNAs means the 8 types of RNAs described in Figure 3. Redundant cases could only happen in RIs, the detailed description sees Materials and Methods.
doi:10.1371/journal.pone.0104049.t002

solution that particle i has obtained until iteration t, and p^t_{gbest} indicates the best global solution obtained from p^t_i in the population at iteration t. To search for the optimal solution, each particle updates its velocity and position as following:

$$\begin{aligned} v^{t+1}_{id} &= w \times v^t_{id} + C_1 \times rand() \times (p^t_{id} - x^t_{id}) \\ &\quad + C_2 \times rand() \times (p^t_{gbest} - x^t_{id}), \\ x^{t+1}_{id} &= x^t_{id} + v^t_{id} \\ d &= 1,2,...,D \end{aligned} \tag{16}$$

Here C_1 denotes the cognition learning factor, C_2 denotes the social learning factor, *rand*() is positive random number which is uniformly sampled from the interval [0,1].

In this study, parameters of our proposed PSOSVM were set as shown in Table 1, and the pseudo-code of the PSOSVM was illustrated in Figure 2. We implemented PSOSVM algorithm in the eclipse platform integrated with Weka (http://www.cs.waikato.ac.nz/ml/weka/) and LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm). The program of our PSOSVM was written in java.

In order to select optimal parameters C and γ in the population, the fitness as an evaluation indicator in PSOSVM was necessary. Here the fitness of (C_i,γ_i) (Figure 2) was set to be the averaged accuracy of the SVM classifier on the training dataset via 10-fold cross-validation (10FCV) experiment.

Performance assessment

Several assessment measures were used to evaluate the classification performance using random forest and PSOSVM in this study. All of them were deduced from the numbers of true positives (*TP*), false positives (*FP*), true negatives (*TN*) and false negatives (*FN*) [51]:

$$\text{Sensitivity} = TP/(TP+FN) \tag{17}$$

$$\text{Specificity} = TN/(TN+FP) \tag{18}$$

$$\text{Accuracy} = (TP+TN)/(TP+FN+TN+FP) \tag{19}$$

$$\text{F-Measure} = 2 \times TP/(2 \times TP+FP+FN) \tag{20}$$

Here Accuracy (Eq. 19) represents the rate of overall correct classifications. F-Measure (Eq. 20) is often used as a single-value benchmark that characterizes classification performance. A receiver operating characteristics (ROC) curve plots True Positive Rate (i.e., Sensitivity) versus False Positive Rate (i.e., 1-Specificity) [52], providing a valuable tool to summarize the accuracy of predictions. The area under the ROC curve (AUC) is used to quantitatively compare the performances of different predictive models without regarding to class distribution or error costs. So we also evaluated the performance using AUC. Moreover, in our experimental data, we utilized probability estimates instead of −1/+1 class labels [53] for each test instance to generate more accurate ROC curve and AUC for PSOSVM.

Results

Experimental dataset

In TAIR10 gene annotation for *Arabidopsis*, there are 28,775 genes, 3,903 transposable element genes and 924 pseudogenes. All these genes except pseudogenes have been used for further analysis, and they have a total of 40,745 annotated RNAs, which can be categorized into 8 different RNA types (Figure 3). It is clear from Figure 3 that most of the annotated RNAs are mRNAs (86.85%, 35,386 out of 40,745). As shown in Table 2, we found a total of 2,811 RIs and 113,098 CSIs in *Arabidopsis*. Interestingly, no RI was detected in chloroplast (ChrC) and mitochondrion (ChrM) while only 42 CSIs cases were found in these organelle genomes. For the 8 different RNA types, both RIs (98.26%, 2,762 out of 2,811) and CSIs (97.53%, 110,304 out of 113,098) were detected overwhelmingly in mRNAs whereas they (RIs: 1.74%, 49 out of 2,811 and CSIs: 2.47%, 2,794 out of 113,098) were rarely discovered among other 7 RNA types. Therefore, all the RIs (2,762) and CSIs (110,262 = 110,304-42) detected in mRNAs within chromosomes Chr1–Chr5 constituted our data set for downstream analysis.

Different from human genome that has much longer introns (5,500 bp in average) [54], *Arabidopsis* has much shorter introns. As shown in Table 3, the average lengths of RIs and CSIs are 145 bp and 160 bp respectively, and introns length varies greatly within a range from 8 to 10,234 bp. Based on the intron length distribution generated by *quantile()* in terms of the given probabilities (0.02, 0.2, 0.4, 0.6, 0.8, 0.98), 96% RIs and CSIs were found within the range from 44 to 501 bp and from 70 to 631 bp respectively. This suggested that extremely large introns (i.e., RIs: 2,075 bp and CSIs: 10,234 bp, 9,724 bp, 7,384 bp) and extremely small introns (i.e., those less than 20 bp) became outliers, which would cause a negative effect on classification. Consequently, we obtained the high-quality dataset including 2,520 RIs and 110,254 CSIs after removing these outliers (i.e., 13

Table 3. Average size, range and sample qualtiles of RIs and CSIs.

Introns Categories	Average size (bp)	Range [Min,Max] (bp)	Quantile (bp)					
			0.02	0.2	0.4	0.6	0.8	0.98
RIs	145	[10–2,075]	44	81	92	112	182	501
CSIs	160	[8–10,234]	70	83	92	110	195	631

Quantile represents *quantile()* function in R. For given probabilities [0.02, 0.2, 0.4, 0.6, 0.8, 0.98], *quantile()* returns estimates of corresponding distribution quantiles based on sort order.
doi:10.1371/journal.pone.0104049.t003

RIs and 8 CSIs) and 229 redundant RIs (see the definition in Materials and Methods).

Supervised machine learning approaches for the identification of RIs and CSIs require a set of labeled samples [55]. In this study, RIs were regarded as positive samples and CSIs were regarded as negative samples. However, the proportion of positive to negative samples was approximately 1:44, which was unbalanced and the performance of classification tended to be biased towards the negative class. To address this issue, under-sampling proves to be an efficient method for classifying unbalanced dataset [56]. We randomly selected three sets of 2600 CSIs from negative samples, by which we conducted our experiments and obtained similar results. So in this paper, we randomly chose one such set of 2,600 CSIs and integrated with 2,520 RIs as our final experimental dataset.

A new hybrid feature extraction approach for classification between RIs and CSIs

As shown in Table 4, our hybrid feature extraction approach obtained 37 features (combining **A+B+C** features) for each intron in the experimental dataset. **A** denotes basic features, including both global features (e.g., Length, nucleotide occurrence probabilities of A, C, G and T, AT content, GC content) and local features (e.g., $\theta_{AG},\theta_{AC},\theta_{AT},\theta_{GC},\theta_{GT},\theta_{CT}$). **B** denotes frequent motifs features, which are selected from all 2 to 5-mer motifs based on Eq. 4–Eq. 10, and have relatively high values of $|\alpha(x(l))|$ and $S_{True}(x(l))$ or $S_{False}(x(l))$. Among the selected frequent motifs, some of them (i.e., cc, gg, cg, ccg, cga, cgg, ggag, gggt, gaag, ttcg) have negative values of $\alpha(x(l))$ and higher values of $S_{True}(x(l))$. Whereas, others (i.e., ta, at, atgt, taat, tatat, atatt, aaata, ttata, attat) possess positive values of $\alpha(x(l))$ and higher values of $S_{False}(x(l))$. **C** denotes the signal strength features of the splice sites (SFvalue, SFaccvalue) and the similarity level features (IDdonv, IDacceptv) of two datasets, which include sequences from -20 to -1 and from +1 to +20 sites for 5′ and 3′ splice sites (Figure 1B).

Besides our hybrid feature extraction approach, we also built complete features (**52**) and optimized features (**27**) to classify RIs and CSIs (Table 4). All trimer sequences have more obvious differences between RIs and CSIs than dimers, and they also present higher frequencies of occurrence in our datasets than tetramers and pentamers. So we sorted values of $|\alpha(x(l))|$ among all trimers and selected top 15 trimers with higher values of $|\alpha(x(l))|$. By integrating the frequencies of these 15 trimers with our combined **A+B+C** features, the complete features were obtained and defined as the **52** feature set. Moreover, we also employed the PSOSearch method to optimize the complete 52 feature set for getting better classification accuracy with less features. PSOSearch is a feature optimal selection method that implements the PSO algorithm. It is available in Weka 3.7.3. In the optimizing process of PSOSearch, the accuracy of random forest classifier was utilized to compare the classification performance of different feature sets. Finally, the optimized features were obtained and defined as the **27** feature set. The last feature is class label with True representing RIs and False representing CSIs.

Evaluation of our hybrid feature extraction approach in comparison to other four feature sets

In this work, because of the diversity of different features (e.g., intron length, SFvalue and frequencies of frequent motifs), we firstly employed *scale* function to normalize values of individual features. Then, we selected 60% samples from the experimental dataset to verify the proposed feature extraction approach. Finally, the normalized feature vectors were adopted as inputs to classify

Table 4. Feature vectors of experimental dataset.

Feature types	Feature vector
Basic Features **[A]**	Length; AT content; GC content; nucleotide occurrence probabilities of A, C, G and T; θ_{AG},θ_{AC},θ_{AT},θ_{GC},θ_{GT},θ_{CT}
Frequent motifs features **[B]**	cc, gg, cg, ccg, cga, cgg, ggag, gggt, gaag, ttcg; ta, at, atgt, taat, tatat, atatt, aaata, ttata, attat
Splice sites and the flanking sequences features **[C]**	SFvalue, SFaccvalue; IDdonv, IDacceptv
Complete features [52]	Combined features (**A+B+C**) and 15 frequencies of trimmers (agg, ata, atg, cgc, cta, gcg, gga, ggg, gta, taa, tac, tag, tat, tcg, tta)
Optimized features [27]	Length, g, t, AT, θ_{AC},θ_{AT},θ_{GC}, cg, ta, cga, cta, gga, tac, tag, tta, gaag, ttcg, atgt, taat, attat, tatat, aaata, SFvalue, SFaccvalue, IDdonv, IDacceptv
Class label	True (RIs); False (CSIs)

doi:10.1371/journal.pone.0104049.t004

RIs and CSIs by employing random forest and PSOSVM respectively.

By using PSO, the optimal parameters C and γ were selected and applied to test the performance of SVM classifier via 10-fold cross-validation. But for random forest, due to the "out-of-bag" error estimation, it is unnecessary to utilize cross-validation to obtain an unbiased estimate of the test set error [33]. We split 90% of samples for training whereas the remainder is used for testing the performance of random forest classifier. As shown in Table 5, the square root of the whole number of features is set for the parameter *numFeatures*, and the other parameter (*numTrees*) of random forest was set from 30 to 50 with a step size of 2 to find the optimal value using grid search algorithm.

In order to demonstrate the performance of our hybrid feature extraction approach, we employed five different feature sets to classify on our dataset: (1) **A** feature set, (2) **A+C** feature set, (3) our combined **A+B+C** feature set, (4) complete **52** feature set and (5) optimized **27** feature set (see Table 4). For each feature set, random forest and PSOSVM were carried out to do classification. The values of optimal parameters and performances of both two classifiers are shown in Table 5. Clearly, the combined **A+B+C** feature set showed better classification performances than other four feature sets for both random forest (i.e., Accuracy = 0.808, F-Measure = 0.808 and AUC = 0.900) and PSOSVM (Accuracy = 0.774, F-Measure = 0.774 and AUC = 0.844). On the other hand, based on these three assessment measures, the random forest classifier always achieved better classification performance than PSOSVM. The differential performance between these two classifier reached 0.056 obtained by AUC assessment measure using our combined feature set (Figure 4).

In Table 5, the performances of AUC appear to be better than those of Accuracy and F-Measure for all feature sets using random forest and PSOSVM. The performances of Accuracy and F-Measure are equal in all feature sets except **A**, in which the performance of F-Measure increases by 0.001 than that of Accuracy for random forest while the opposite case happens for PSOSVM. Interestingly, these five different feature sets display the same change trend of classification performance in terms of these three assessment measures for both random forest and PSOSVM. We focus on F-Measure (Figure 5) to illustrate this trend.

As shown in Figure 5, compared with our combined feature set, it is impossible to obtain better performance for applying sectional feature sets (e.g., **A** or **A+C**), complete **52** feature set, or optimized **27** feature set. This result suggests, not only for PSOSVM but also for random forest, that our hybrid feature extraction approach selected useful features for better classification between RIs and CSIs.

The influences of short motifs, splice sites and flanking exon sequences in RIs

When we further investigated the influence of the feature sets in classifying RIs and CSIs, we discovered that **C** feature set made

Table 5. Optimal parameters and performances of random forest and PSOSVM using five different feature sets.

Algorithm	Feature set	Parameter (*numFeatures*)	Parameter (*numTrees*)	Accuracy	F-Measure	AUC
Random forest	**A**	4	42	0.771	0.772	0.867
	A+C	4	42	0.785	0.785	0.897
	Combined **A+B+C**	**6**	**42**	**0.808**	**0.808**	**0.900**
	Complete **52**	7	42	0.782	0.782	0.898
	Optimized **27**	5	42	0.788	0.788	0.891
Algorithm	**Feature set**	**Parameter (C)**	**Parameter (γ)**	**Accuracy**	**F-Measure**	**AUC**
PSOSVM	**A**	187.29	15.64	0.742	0.741	0.838
	A+C	1.0	5.5	0.771	0.771	0.842
	Combined **A+B+C**	**59.46**	**0.17**	**0.774**	**0.774**	**0.844**
	Complete **52**	1.40	0.22	0.771	0.771	0.843
	Optimized **27**	1.27	0.52	0.763	0.763	0.840

doi:10.1371/journal.pone.0104049.t005

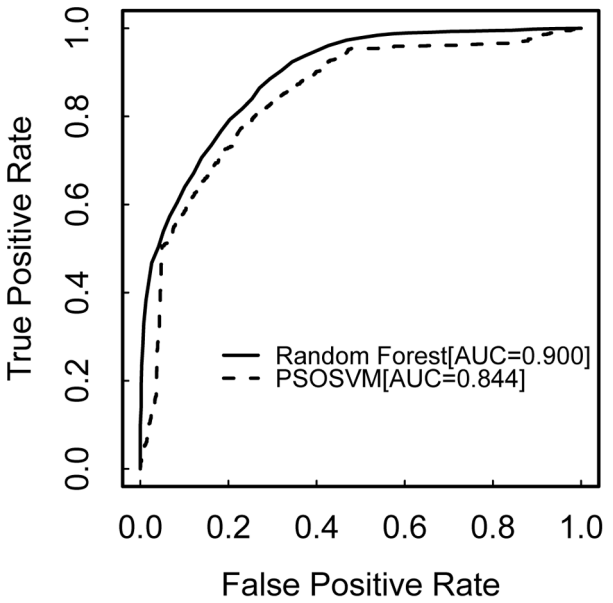

Figure 4. The ROC curves of random forest versus PSOSVM. The ROC curve of random forest is shown by the solid line and PSOSVM by the dashed line. The classification accuracy of these two methods is measured by AUC (the area under the ROC curve). Random forest gains significant advantages compared to PSOSVM (i.e., 0.900 versus 0.844).
doi:10.1371/journal.pone.0104049.g004

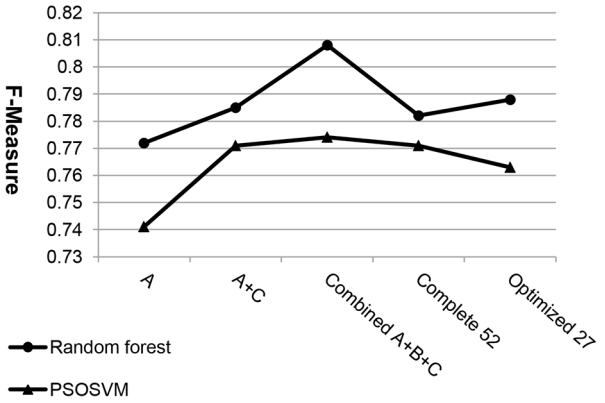

Figure 5. Performance of random forest and PSOSVM (F-Measure) in five different feature sets. Classification accuracy is assessed with F-Measure. Each solid round dot represents the accuracy of random forest and each triangle means the accuracy of PSOSVM for a given feature set. Compared with the other feature sets, our combined **A+B+C** feature set obtains the optimal classification performance by using both classifiers.
doi:10.1371/journal.pone.0104049.g005

the greatest contribution to improve the classification performance: for example, 3% F-Measure increase using PSOSVM, and 3% AUC increase using random Forest for **A+C** feature set in comparison with **A** feature set (Table 5). As shown in Table 6, RIs have lower signal strength of splice sites (SFvalue = 3.930, SFaccvalue = 5.075) than CSIs (SFvalue = 4.806, SFaccvalue = 6.363). In addition, RIs have smaller values of IDdonv (17.934) and IDacceptv (17.891) than CSIs (IDdonv = 18.412, IDacceptv = 18.385), which suggests that intron sequences and flanking exon sequences for both donor sites (5′ splice sites) and acceptor sites (3′ splice sites, see Figure 1B) have higher similarity in RIs than in CSIs. The significant differences among these four features (SFvalue, SFaccvalue, IDdonv and IDacceptv) were detected between RIs and CSIs using one-way ANOVA ($P < 0.0001$, see Table 6). This result demonstrated that these four features were indeed effective in classification between two kinds of introns.

Meanwhile, we also found that some short motifs were relatively frequent but quite different between the RIs and CSIs. So we extracted **B** feature set, and the results showed that they indeed helped us improve the classification performance, especially by using random forest (e.g., 2.3% F-Measure and Accuracy increase for our **A+B+C** feature set in comparison with **A+C** feature set, see Table 5). As showed in Figure 6, some short motifs (e.g., cc, gg, cg, ccg, cga, cgg, ggag, gggt, gaag, ttcg) have higher frequencies in RIs than CSIs whereas others (e.g., ta, at, atgt, taat, tatat, atatt, aaata, ttata, attat) occur higher frequencies in CSIs than RIs.

Discussion

Different from previous bioinformatics analyses of AS in *Arabidopsis* [11,57], we used the most recent and well-annotated gene data from TAIR10 to extract our experimental intron dataset that consists of 2,520 RIs and 110,254 CSIs, and found RIs and CSIs showed distinctive characteristics in their sequences. We not only discovered similar features including shorter intron length, lower AT content and higher GC content in RIs with previous reports [13,58], but also found θ_{GC} (14.3% versus 12.4%) was obviously higher and θ_{CT} was conversely lower (23.0% versus 25.5%) in RIs than in CSIs. This indicates that difference between G and C contents for segmental intron sequences in RIs is greater than that in CSIs, whereas the difference between C and T contents for segmental intron sequences is higher in CSIs than that in RIs. As for the terminal dinucleotide splice signals of introns, there was no surprise that the consensus GT-AG introns (i.e., introns that begin with GT and end with AG) held 99% of CSIs and 96.7% of RIs. The second largest class, GC-AG introns, appeared more frequently in RIs than CSIs (2.61% versus 0.90%). This finding suggests that in *Arabidopsis* the unusual GC-AG introns appear to be more frequent in RIs than CSIs while the canonical GT-AG introns are richer in CSIs than RIs. Another

Table 6. The mean value and P value of SFvalue, SFaccvalue, IDdonv and IDacceptv.

	SFvalue	SFaccvalue	IDdonv	IDacceptv
The mean value in RIs	3.930	5.075	17.934	17.891
The mean value in CSIs	4.806	6.363	18.412	18.385
P value (One–way ANOVA)	2.2e-16	2.2e-16	6.488e-07	3.545e-07

P value was calculated by applying F-test in one-way ANOVA based on experiment dataset included RIs and CSIs. The influences of classification among four features are all significant ($p<0.0001$).
doi:10.1371/journal.pone.0104049.t006

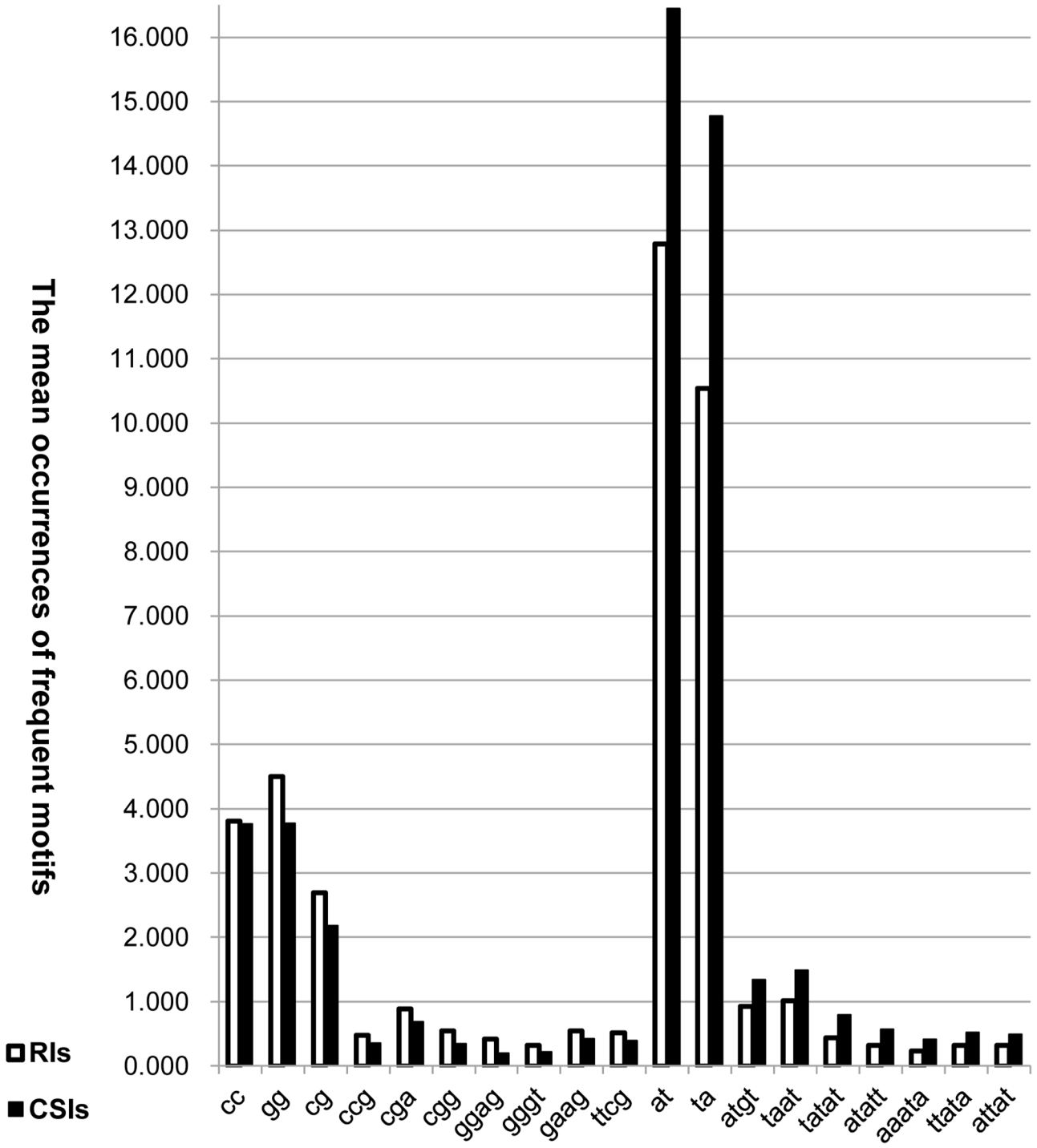

Figure 6. The mean occurrences of B frequent motifs between RIs and CSIs. In the left side of the histogram there are ten frequent motifs that have higher occurrences in RIs than in CSIs. In the right site of the histogram there are nine frequent motifs that have higher occurrences in CSIs than in RIs.
doi:10.1371/journal.pone.0104049.g006

interesting phenomenon in our data analysis is that more than half of RIs (58.4%) occurs in CDS, CDS+3′UTR or CDS+5′UTR regions. Such positional distribution characteristic of RIs indicates the potential that these introns are partly or entirely translated to proteins. Previous studies demonstrate that growing examples of cellular mRNAs with RIs express functional proteins by avoiding degradation through the nonsense-mediated decay (NMD) [59–61]. Our data analysis also provides a support for this trend by a high rate of RIs existing in coding regions.

It is well known that cis-acting sequences or motifs [62], such as enhancers and silencers in exons or introns, play significant roles for the regulation of AS. Plenty of studies indicate that exonic splicing enhancers and silences (ESEs and ESSs), most of which are known to bind SR proteins of the spliceosome, affect intron

excision [63,64]. *Pertea et al.* [65] has identified 84 putative exonic splicing enhancers (hexamers) in *Arabidopsis* by a computational approach. Although intronic splicing enhancers and silences (ISEs and ISSs) are less understood than ESEs and ESSs, a previous study also suggests [66] that these intronic splicing regulatory motifs also commonly impact on AS in mammals. Based on our feature extraction approach of **B** feature set, we analyzed all ggg-containing motifs with length from 3 to 5 bp included g triples (ggg, a well-established mammalian ISEs [67]), and found the mean value of $\alpha(x(\text{ggg-containing motifs}))$ was -0.358, which indicated that ggg-containing motifs occurred more frequently in RIs than in CSIs. So the above result suggests these ggg-containing motifs, such as "gggt", "gggtt" and "tgggt", play a role of ISSs in *Arabidopsis*, instead of the role of ISEs in mammals. Of all ggg-containing motifs, "gggt" proves to contribute in distinguishing RIs from CSIs by our classification methods. In addition, the result of our extraction approach of **B** feature set also discovers that the mean value of $\alpha(x(\text{ggag-containing motifs}))$ was -0.539, which indicated that ggag-containing motifs also have higher occurrences in RIs than in CSIs. In our study, the frequent motifs "ta, at, atgt, taat, tatat, atatt, aaata, ttata, attat" suggest some at/ta-rich motifs (i.e., ones include linear repeat or combination of "at" or "ta" (at least two "at" or "ta")) may be ISEs in *Arabidopsis*. We checked all at/ta-rich motifs with length 4 and 5 bp and obtained the mean value of $\alpha(x(\text{at/ta-rich motifs}))$ was 0.276, which illustrated at/ta-rich motifs had more frequently in CSIs than RIs. Furthermore, as the outstanding representatives of these at/ta-rich motifs, "taat", "tatat", "atatt", "ttata" and "attat" have been proved to help recognizing the CSIs in our data analysis. Overall, ggg-containing and ggag-containing motifs seem to be ISSs because of their obvious abundance in RIs than in CSIs. On the other hand, at/ta-rich motifs appear to be ISEs because of their significant abundance in CSIs than RIs in *Arabidopsis*, which would potentially promote the identification of intronic splicing regulatory elements in plants.

Our results clearly demonstrate that random forest offers more advantageous classification performance than PSOSVM on five different feature sets. Performances of these two kinds of classifier are influenced by their respective parameters. Our experience showed that the parameter optimization was easier to implement for random forest ($numFeatures = \sqrt{\text{numbers of features}}$, where *numTrees* is obtained by grid search), and the optimized parameters were beneficial to obtain stable classifier performance. In contrast, different values of (C, γ) would cause large variation in the classifier performance of SVM [22]. Although we employed PSO to search the optimal parameters and have obtained better classification performance in comparison with the result using traditional grid search method, the classification performance of SVM may be further improved if the parameters could avoid trapping into local optima [68]. Unlike SVM, individual decision trees in random forest automatically utilize informative features more frequently in training process and achieve independent predictions, which were combined to gain accurate prediction of the forest [30,69]. Therefore random forest presents significant superiority in failure tolerances and robustness, which plausibly explain the consistent advantageous performance of random forest classifier for all five feature sets in our study.

In this study, we utilized current TAIR10 mRNA (transcript or isoform) annotation in *Arabidopsis*, which does not provide any quantitative expression information (i.e., highly expressed versus rarely expressed mRNA) for alternate isoforms derived from the same genes. It is likely that highly expressed retained introns have different signal strength than retained introns with low expression levels. Therefore, utilizing RNA-Seq data to extract and incorporate expression information in intron level will definitely facilitate the development of more accurate and robust classifier by machine learning strategies. In fact, a recent RNA-Seq data analysis already shows evidence for novel transcripts and alternative splicing events in *Arabidopsis* that are not annotated in TAIR10 [70]. As more and more RNA-Seq and their meta-data (e.g., including environmental treatments, developmental stages and sampled tissues) are becoming available, more novel isoforms and previously un-annotated RIs will be evident in *Arabidopsis*, which can help us enhance the classification performance by providing more members within the RIs class. Moreover, we can do further classification of RIs that might be related to different environmental and/or developmental cues. Obviously, more RIs with different meta-data can be further analyzed to extract stress-, tissue-, or growth stage-specific features so that we can better understand how RIs are affected by both external and internal conditions in plants. On the other hand, RNA secondary structures have been demonstrated to affect alternative splicing [11,71,72]. Recently, the first *in vivo* genome-wide RNA structure map in *Arabidopsis* [73] highlights the importance of RNA secondary structures in alternative splicing (including intron retention). Therefore, a great challenge is how to accurately and effectively incorporate RNA secondary structures as features to enhance the performance and accuracy of our classifier. Without a doubt, a comprehensive feature extraction including both linear sequence features and RNA secondary structure features will definitely facilitate our understanding of how RIs are regulated in plants.

Conclusions

The primary contribution of this work is our novel hybrid feature extraction approach that reveals overall features of introns, splice sites and flanking exons. These features can be utilized to effectively categorize and differentiate between RIs and CSIs. The experiments on five different feature sets verified that our combined **A+B+C** feature set could obtain the optimal classification performance by applying random forest and PSOSVM classifiers after tuning parameters. Follow-up analysis of these features has revealed interesting information about RIs in comparison with CSIs:

(1) In average RIs have shorter length (145 bp versus 160 bp), higher GC content (35.76% versus 32.43%) and lower AT content (64.24% versus 67.57%) than CSIs.

(2) RIs show different features of segmental nucleotides composition, such as higher θ_{GC} and lower θ_{CT} locally.

(3) RIs possess lower signal strength of 5′ and 3′ splice sites (SFvalue, SFaccvalue), and terminal dinucleotide GC-AG appears a higher frequency in RIs than CSIs.

(4) The RIs show higher similarity with their flanking exons than CSIs.

(5) We here propose ggg-containing and ggag-containing motifs as ISSs as they are enriched in RIs. Accordingly, at/ta-rich motifs seem to be ISEs because of abundant in CSIs.

These features information about RIs can effectively facilitate an understanding of recognition mechanism of RIs in *Arabidopsis*.

Supporting Information

File S1 Detailed introduction for how to extract data (File S2) using our source codes (File S3).
(DOCX)

File S2 All extracted data used in the article.
(ZIP)

File S3 All source codes used in the article.
(ZIP)

References

1. Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, et al. (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol 14: 153–165. doi:10.1038/nrm3525.
2. Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11: 345–355. doi:10.1038/nrg2776.
3. Sammeth M, Foissac S, Guigó R (2008) A General Definition and Nomenclature for Alternative Splicing Events. PLoS Comput Biol 4: e1000147. doi:10.1371/journal.pcbi.1000147.
4. Kim N, Alekseyenko AV, Roy M, Lee C (2007) The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Res 35: D93–D98. doi:10.1093/nar/gkl884.
5. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40: 1413–1415. doi:10.1038/ng.259.
6. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The Transcriptional Landscape of the Mammalian Genome. Science 309: 1559–1563. doi:10.1126/science.1112014.
7. Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O (2005) Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. Gene 364: 53–62. doi:10.1016/j.gene.2005.07.027.
8. Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. Nucleic Acids Res 35: 125–131. doi:10.1093/nar/gkl924.
9. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics 7: 327. doi:10.1186/1471-2164-7-327.
10. Wang B-B, Brendel V (2006) Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci 103: 7175–7180. doi:10.1073/pnas.0602039103.
11. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, et al. (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res 20: 45–58. doi:10.1101/gr.093302.109.
12. Syed NH, Kalyna M, Marquez Y, Barta A, Brown JWS (2012) Alternative splicing in plants–coming of age. Trends Plant Sci 17: 616–623. doi:10.1016/j.tplants.2012.06.001.
13. Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, et al. (2004) Intron retention is a major phenomenon in alternative splicing in Arabidopsis. Plant J Cell Mol Biol 39: 877–885. doi:10.1111/j.1365-313X.2004.02172.x.
14. Mastrangelo AM, Belloni S, Barilli S, Ruperti B, Di Fonzo N, et al. (2005) Low temperature promotes intron retention in two e-cor genes of durum wheat. Planta 221: 705–715. doi:10.1007/s00425-004-1475-3.
15. Palusa SG, Ali GS, Reddy ASN (2007) Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. Plant J Cell Mol Biol 49: 1091–1107. doi:10.1111/j.1365-313X.2006.03020.x.
16. Ner-Gaon H, Fluhr R (2006) Whole-Genome Microarray in Arabidopsis Facilitates Global Analysis of Retained Introns. DNA Res 13: 111–121. doi:10.1093/dnares/dsl003.
17. Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. Genome Res 18: 1381–1392. doi:10.1101/gr.053678.106.
18. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, et al. (2006) Machine learning in bioinformatics. Brief Bioinform 7: 86–112. doi:10.1093/bib/bbk007.
19. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV (2003) Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions. J Chem Inf Comput Sci 43: 2048–2056. doi:10.1021/ci0340916.
20. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21: 631–643. doi:10.1093/bioinformatics/bti033.
21. O'Fallon BD, Wooderchak-Donahue W, Crockett DK (2013) A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. Bioinforma Oxf Engl 29: 1361–1366. doi:10.1093/bioinformatics/btt172.
22. Wang J, Chen Q, Chen Y (2004) RBF Kernel Based Support Vector Machine with Universal Approximation and Its Application. In: Yin F-L, Wang J, Guo C, editors. Advances in Neural Networks – ISNN 2004. Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 512–517. Available: http://link.springer.com/chapter/10.1007/978-3-540-28647-9_85. Accessed 4 November 2013.
23. Hsu C-W, Chang C-C, Lin C-J (2003) A practical guide to support vector classification. Available: https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf. Accessed 27 May 2014.
24. Lin H-T, Lin C-J (2003) A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Submitt Neural Comput: 1–32.
25. Min JH, Lee Y-C (2005) Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Syst Appl 28: 603–614. doi:10.1016/j.eswa.2004.12.008.
26. Lin S-W, Ying K-C, Chen S-C, Lee Z-J (2008) Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Syst Appl 35: 1817–1824. doi:10.1016/j.eswa.2007.08.088.
27. Huang C-L, Dun J-F (2008) A distributed PSO–SVM hybrid system with feature selection and parameter optimization. Appl Soft Comput 8: 1381–1391. doi:10.1016/j.asoc.2007.10.007.
28. Statnikov A, Aliferis CF (2007) Are random forests better than support vector machines for microarray-based cancer classification? AMIA Annu Symp Proc AMIA Symp AMIA Symp: 686–690.
29. Rodriguez-Galiano VF, Chica-Olmo M, Abarca-Hernandez F, Atkinson PM, Jeganathan C (2012) Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sens Environ 121: 93–107. doi:10.1016/j.rse.2011.12.003.
30. Boulesteix A-L, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2: 493–507. doi:10.1002/widm.1072.
31. Masso M, Vaisman II (2010) Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. J Theor Biol 266: 560–568. doi:10.1016/j.jtbi.2010.07.026.
32. Nair V, Dutta M, Manian SS, S RK, Jayaraman VK (2013) Identification of Penicillin-binding proteins employing support vector machines and random forest. Bioinformation 9: 481–484. doi:10.6026/97320630009481.
33. Díaz-Uriarte R, Andrés SA de (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7: 3. doi:10.1186/1471-2105-7-3.
34. Nimrod G, Szilágyi A, Leslie C, Ben-Tal N (2009) Identification of DNA-binding Proteins Using Structural, Electrostatic and Evolutionary Features. J Mol Biol 387: 1040–1053. doi:10.1016/j.jmb.2009.02.023.
35. Mendoza MR, da Fonseca GC, Loss-Morais G, Alves R, Margis R, et al. (2013) RFMirTarget: Predicting Human MicroRNA Target Genes with a Random Forest Classifier. PLoS ONE 8: e70153. doi:10.1371/journal.pone.0070153.
36. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875. doi:10.1093/bioinformatics/bti310.
37. Wu R, Hu Q, Li R, Yue G (2012) A novel composition coding method of DNA sequence and its application. Match-Commun Math Comput Chem 67: 269.
38. Yang W, Li Q (2008) One parameter to describe the mechanism of splice sites competition. Biochem Biophys Res Commun 368: 379–381. doi:10.1016/j.bbrc.2008.01.089.
39. Florea L (2006) Bioinformatics of alternative splicing and its regulation. Brief Bioinform 7: 55–69. doi:10.1093/bib/bbk005.
40. Wang F, Wang Z, Li H, Yang K (2011) Prediction of protein structural classes using the theory of increment of diversity and support vector machine. Wuhan Univ J Nat Sci 16: 260–264. doi:10.1007/s11859-011-0747-6.
41. Marquez Y, Brown JWS, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res 22: 1184–1195. doi:10.1101/gr.134106.111.
42. Statnikov A, Wang L, Aliferis CF (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9: 319. doi:10.1186/1471-2105-9-319.
43. Naidoo L, Cho MA, Mathieu R, Asner G (2012) Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. ISPRS J Photogramm Remote Sens 69: 167–179. doi:10.1016/j.isprsjprs.2012.03.005.
44. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, et al. (2009) The WEKA Data Mining Software: An Update. SIGKDD Explor Newsl 11: 10–18. doi:10.1145/1656274.1656278.
45. Byvatov E, Schneider G (2003) Support vector machine applications in bioinformatics. Appl Bioinformatics 2: 67–77.
46. Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17: 721–728. doi:10.1093/bioinformatics/17.8.721.
47. Scholkopf B, Sung K-K, Burges CJC, Girosi F, Niyogi P, et al. (1997) Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans Signal Process 45: 2758–2765. doi:10.1109/78.650102.
48. Huang C-L, Wang C-J (2006) A GA-based feature selection and parameters optimizationfor support vector machines. Expert Syst Appl 31: 231–240. doi:10.1016/j.eswa.2005.09.024.

Author Contributions

Conceived and designed the experiments: CL RM. Performed the experiments: RM. Analyzed the data: RM. Contributed reagents/materials/analysis tools: RM. Wrote the paper: RM PKRK CG YZ CL.

49. Zhang X, Guo Y (2009) Optimization of SVM Parameters Based on PSO Algorithm. Fifth International Conference on Natural Computation, 2009. ICNC '09. Vol. 1. pp. 536–539. doi:10.1109/ICNC.2009.257.
50. Abdi MJ, Hosseini SM, Rezghi M (2012) A Novel Weighted Support Vector Machine Based on Particle Swarm Optimization for Gene Selection and Tumor Classification. Comput Math Methods Med 2012. Available: http://www.hindawi.com/journals/cmmm/2012/320698/abs/. Accessed 6 November 2013.
51. Liu J, Gough J, Rost B (2006) Distinguishing Protein-Coding from Non-Coding RNAs through Support Vector Machines. PLoS Genet 2: e29. doi:10.1371/journal.pgen.0020029.
52. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30: 1145–1159. doi:10.1016/S0031-3203(96)00142-2.
53. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27: 861–874. doi:10.1016/j.patrec.2005.10.010.
54. Sakharkar MK, Chow VTK, Kangueane P (2004) Distributions of exons and introns in the human genome. In Silico Biol 4: 387–393.
55. Wei L, Yang Y, Nishikawa RM, Jiang Y (2005) A study on several Machine-learning methods for classification of Malignant and benign clustered microcalcifications. IEEE Trans Med Imaging 24: 371–380. doi:10.1109/TMI.2004.842457.
56. Yen S-J, Lee Y-S (2009) Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl 36: 5718–5727. doi:10.1016/j.eswa.2008.06.108.
57. Eichner J, Zeller G, Laubinger S, Rätsch G (2011) Support vector machines-based identification of alternative splicing in Arabidopsis thaliana from whole-genome tiling arrays. BMC Bioinformatics 12: 55. doi:10.1186/1471-2105-12-55.
58. Sakabe NJ, Souza SJ de (2007) Sequence features responsible for intron retention in human. BMC Genomics 8: 59. doi:10.1186/1471-2164-8-59.
59. Torrado M, Iglesias R, Nespereira B, Centeno A, López E, et al. (2009) Intron retention generates ANKRD1 splice variants that are co-regulated with the main transcript in normal and failing myocardium. Gene 440: 28–41. doi:10.1016/j.gene.2009.03.017.
60. Mollet IG, Ben-Dov C, Felicio-Silva D, Grosso AR, Eleuterio P, et al. (2010) Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. Nucleic Acids Res 38: 4740–4754. doi:10.1093/nar/gkq197.
61. Coyle JH, Bor Y-C, Rekosh D, Hammarskjold M-L (2011) The Tpr protein regulates export of mRNAs with retained introns that traffic through the Nxf1 pathway. RNA 17: 1344–1356. doi:10.1261/rna.2616111.
62. Wittkopp PJ, Kalay G (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 13: 59–69. doi:10.1038/nrg3095.
63. Fairbrother WG, Yeh R-F, Sharp PA, Burge CB (2002) Predictive Identification of Exonic Splicing Enhancers in Human Genes. Science 297: 1007–1013. doi:10.1126/science.1073774.
64. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, et al. (2004) Systematic identification and analysis of exonic splicing silencers. Cell 119: 831–845. doi:10.1016/j.cell.2004.11.010.
65. Pertea M, Mount SM, Salzberg SL (2007) A computational survey of candidate exonic splicing enhancer motifs in the model plant Arabidopsis thaliana. BMC Bioinformatics 8: 159. doi:10.1186/1471-2105-8-159.
66. Yeo GW, Nostrand ELV, Liang TY (2007) Discovery and Analysis of Evolutionarily Conserved Intronic Splicing Regulatory Elements. PLoS Genet 3: e85. doi:10.1371/journal.pgen.0030085.
67. Yeo G, Hoon S, Venkatesh B, Burge CB (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc Natl Acad Sci U S A 101: 15700–15705. doi:10.1073/pnas.0404901101.
68. Nakano S, Ishigame A, Yasuda K (2010) Consideration of Particle Swarm Optimization combined with tabu search. Electr Eng Jpn 172: 31–37. doi:10.1002/eej.20966.
69. Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: A survey and results of new tests. Pattern Recognit 44: 330–349. doi:10.1016/j.patcog.2010.08.011.
70. Loraine AE, McCormick S, Estrada A, Patel K, Qin P (2013) RNA-Seq of Arabidopsis Pollen Uncovers Novel Transcription and Alternative Splicing. Plant Physiol 162: 1092–1109. doi:10.1104/pp.112.211441.
71. Solnick D (1985) Alternative splicing caused by RNA secondary structure. Cell 43: 667–676.
72. Jin Y, Yang Y, Zhang P (2011) New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. RNA Biol 8: 450–457.
73. Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, et al. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505: 696–700. doi:10.1038/nature12756.
74. Shi Y, Eberhart R (1998) A modified particle swarm optimizer. The 1998 IEEE International Conference on Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence. pp. 69–73. doi:10.1109/ICEC.1998.699146.