ERISdb: A Database of Plant Splice Sites and Splicing Signals

Michał Wojciech Szczęśniak1,*, Michał Kabza1, Rafał Pokrzywa2, Adam Gudys2 and Izabela Makalowska1,*

1Laboratory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University, 61-614 Poznan, Poland
2Institute of Informatics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland
*Corresponding authors: Michał Wojciech Szczęśniak, E-mail, miszcz@amu.edu.pl; Izabela Makalowska: E-mail, izabel@amu.edu.pl.

(Received November 7, 2012; Accepted January 2, 2013)

Splicing is one of the major contributors to observed spatio-temporal diversification of transcripts and proteins in metazoans. There are numerous factors that affect the process, but splice sites themselves along with the adjacent splicing signals are critical here. Unfortunately, there is still little known about splicing in plants and, consequently, further research in some fields of plant molecular biology will encounter difficulties. Keeping this in mind, we performed a large-scale analysis of splice sites in eight plant species, using novel algorithms and tools developed by us. The analyses included identification of orthologous splice sites, polypyrimidine tracts and branch sites. Additionally we identified putative intronic and exonic cis-regulatory motifs, U12 introns as well as splice sites in 45 microRNA genes in five plant species. We also provide experimental evidence for plant splice sites in the form of expressed sequence tag and RNA-Seq data. All the data are stored in a novel database called ERISdb and are freely available at http://lemur.amu.edu.pl/share/ERISdb/.

Keywords: MicroRNA • Splice sites • Splicing signals • U12 introns.

Introduction

One of the most prominent features of eukaryotic genes is that they possess quite a complex structure, which is attributed to the presence of spliceosomal introns. In photosynthetic eukaryotes, the vast majority of protein-coding genes (up to 90%) contain introns (Barbazuk and McGinnis 2008, Labadorf et al. 2010). The presence of introns has notable functional consequences for the cell. For instance, the primary transcripts are longer, up to hundreds of thousands of bases, than the actual coding sequence, which requires a lot more substrates and energy in the process of transcription. It therefore does not seem surprising that introns play fundamental roles, being key elements in the process of alternative splicing, which is a critical contributor to the transcriptome and proteome complexity in most eukaryotes. In the process of alternative splicing, primary transcripts from intron-containing genes are spliced by differential selection of splice sites in a spatio-temporal manner, leading to production of multiple mature mRNAs from a single gene (Pan et al. 2008, Kalsotra and Cooper 2011). Protein isoforms produced in this way may possess altered functions (Stamm et al. 2005). Additionally, alternative splicing plays a key role in gene regulation through regulated production of splice variants with a premature termination codon that are degraded in nonsense-mediated decay (Palusa and Reddy 2010). It also may lead to production of alternative splice forms that contain or lack microRNA (miRNA) target sequences (Tan et al. 2007). As a result, post-transcriptional regulation through alternative splicing constitutes an elaborate mechanism to fine-tune gene expression and provide proteome diversity.

The (alternative) splicing is performed by a large ribonucleoprotein complex called the spliceosome. In many eukaryotes, including most plant and animal species, there are two types of spliceosomes, the major and minor ones. The major spliceosome is responsible for splicing of the vast majority of introns in both plants and animals, and interacts with so-called U2 introns. The latter participates in splicing of U12 introns and spills out ~0.3% of introns in human (Lavine and Durbin 2001) and ~0.15% in Arabidopsis thaliana (Zhu and Brendel 2003). U2 introns usually possess GT and AG terminal dinucleotides at their 5’ and 3’ termini, respectively. In U12 introns, the terminal dinucleotides are more divergent, with GT–AC and AT–AC being the most prevalent dinucleotides. A distinctive feature of U12 introns is that they possess a well-conserved donor site with a consensus sequence RTATCCCTT as well as a distinct branch site with a consensus sequence TTCCTT RAY (Dietrich et al. 1997). Relatively high evolutionary conservation of U12 introns suggests that they might play important roles in the cell, and indeed they have been implicated in several
molecular phenomena (Patel et al. 2002, Hirose et al. 2004, Hastings et al. 2005). So far, a few studies have been performed to search for U12 introns in eukaryotes, but there is still little known about them in plants (Alioto 2007).

Correct recognition of splice sites by the spliceosome is critical for proper excision of introns from a primary transcript. There are three canonical splicing signals that guide a spliceosome to splice sites. The first one constitutes splice sites themselves, one at the 5’ end of an intron (donor site) and the second at the 3’ end (acceptor site), with the most important roles being played by highly conserved intronic terminal dinucleotides. The two latter elements are the polypyrimidine tract (PPT), rich in C and U and usually 15–20 bases long, and the branch site containing a so-called branch point nucleotide, required to produce a lariat intermediate, a key step in the splicing process. In plants, there is another element, the UA-rich tract, which is required for effective splicing of U2 introns (Goodall and Filipowicz 1989) and improves the splicing of U12 introns (Lewandowska et al. 2004). Although these core sequence features are quite conserved across species, they alone are not sufficient to define exons and introns and recruit the splicing machinery. In fact, some plant introns lack PPTs or UA tracts. Additional intronic and exonic sequences, usually referred to as splicing regulatory elements (SREs) or cis-acting elements, are important for both constitutive and alternative splicing. The SREs function as either splicing enhancers or suppressors and affect splice site choice by interacting with proteins that are collectively called trans-acting factors. Depending on the location of SREs and their effect on splicing, they are grouped into four classes: intronic splicing enhancers (ISEs), intronic splicing silencers (ISSs), exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs) (Wang and Burge 2008, Wang et al. 2009).

In recent years, much attention was paid to a class of non-coding RNAs called miRNAs, which resulted in identification of thousands of miRNAs in hundreds of plant, animal and protist species (Lin et al. 2009, Szczesniak et al. 2012). These small RNA molecules post-transcriptionally regulate the expression of thousands of genes in plants and animals either by transcript cleavage (Reinhart et al. 2002) or by translational repression (Lai 2002), and are key players in stress-related, developmental and signaling processes (Kedde and Agami 2008, Leung and Sharp 2010, O’Neill 2011). As a result, much hope is placed in untangling mechanisms of miRNA function and harnessing them in a number of applications in biotechnology, medicine or molecular biology. This depends on a satisfactory understanding of their biogenesis and expression regulation, which in turn might require uncovering their exon—intron structures and the confines of a gene. This is true at least for plant miRNAs, as animal miRNAs supposedly lack introns. However, little has been done to determine miRNA gene structures in plants, except for single analyses in A. thaliana (Szarzynska et al. 2009) and Vitis vinifera (Mica et al. 2010).

Keeping in mind the above-mentioned insufficiency of plant data, we have performed large-scale analyses of plant introns and splice sites in eight species: A. thaliana, Chlamydomonas reinhardtii, Glycine max, Oryza sativa, Physcomitrella patens, Selaginella moellendorfii, V. vinifera and Zea mays. These included a search for PPTs, UA-rich tracts, and branch sites, determining expressed sequence tags (ESTs) and RNA-Seq reads that support annotated splice sites, identification of orthologous splice sites, finding novel U12 introns and uncovering miRNA gene structures. In order to accomplish these tasks, we developed novel tools and algorithms, including a highly accurate classifier for U12 intron search and a tool for identification of splicing cis-regulatory elements (both available for download). We also collected some external data from published papers and databases to complement our findings. All data are deposited in a newly created online database with a user-friendly interface. We called the resource ERISdb and made it available at http://lemur.amu.edu.pl/share/ERISdb/.

Methods

Data download

Genome, protein and transcript sequences as well as annotation data and orthologous gene relationships were downloaded from Ensembl Plants release 15 (Vilella et al. 2009, Kersey et al. 2010). EST sequences were downloaded from dbEST release 120701 (Boguski et al. 1993), and miRNA sequences from miRBase 19 (Kozomara and Griffiths-Jones 2011). RNA-Seq data were retrieved from NCBI’s Sequence Read Archive (Leinonen et al. 2011) and included seven libraries: SRP002417 (whole plant, P. patens), SRP002417 (aerial tissue, S. moellendorfii), DRS000668 (tissue pool, O. sativa) and SRP011480 (Z. mays, four libraries: immature tassel, seedling root, seedling shoot and unpollinated ear tip). U12 splice site data were downloaded from U12DB (Alioto 2007). The data comprised U12 splice site and branch site sequences of 17 animal species and A. thaliana. RACE (rapid amplification of cDNA ends) sequencing products for A. thaliana miRNAs were retrieved from NCBI, based on data provided in the corresponding publication (Szarzynska et al. 2009). Additionally, from the study on V. vinifera miRNAs (Mica et al. 2010), we downloaded deep sequencing data that support introns in miRNA genes. Finally, when mapping SREs to splice sites, sequences of exon splicing enhancers identified by Pertea et al. (2007) were used along with putative SREs identified in this research.

Branch site search

For branch site identification, we used a Perl script developed by Schwartz et al. (2008). The script operates in three steps. First, it scans the 100 nucleotides (nt) upstream of the 3’ splice site and identifies the following heptamers: NNYTRAY, NNCTYAC, NNRTAAC and NNCTAAA, which were previously identified in hemiascomycetous yeast (Bon et al. 2003) and Schizosaccharomyces pombe (http://www.sanger.ac.uk/Projects/S_pombe/intron.shtml). Then, it scores each...
heptamer according to the number of mismatches from the optimal consensus of TACTAAC and, finally, discards all introns in which the best-scoring hit is not the most downstream one. Although the last step discards a relatively large fraction of introns, it is believed to reduce the false-positive rate significantly (Schwartz et al. 2008).

Identification of PPTs and UA-rich tracts

PPTs and UA-rich tracts were searched in intronic regions of up to 50 bases upstream of the 3’ splice site. Additionally—in the case of PPTs—we required that they end within the last 10 bases of an intron. The same parameters for the PPT search were previously applied in a large-scale analysis of eukaryotic splice sites (Schwartz et al. 2008) as PPTs beyond these confines are unlikely to be functional (Coolidge et al. 1997, Kol et al. 2005). The algorithm implemented in Python searched for the longest string with the C + U (in the case of PPTs) or A + U (for UA tracts) composition exceeding 85%. Moreover, the tracts were required to be at least five bases long. If a PPT was found downstream of a branch site, it was considered as a ‘putative PPT’, otherwise we called it a ‘CT tract’.

Search for splicing cis-regulatory elements

We wrote a Java program to search for splicing motifs in intronic and exonic sequences. In the first step, by sliding the window of size \( m \) (input parameter) across the input sequences, the program identifies all \( m \)-length substrings and counts the number of their exact occurrences in the input sequences. In the next step, all substrings are connected to each other by similarity. Two substrings \( k \) and \( h \) are connected if and only if the number of similar nucleotides between \( k \) and \( h \) is equal to or greater than the value of the input parameter \( s \). Each distinct substring with all connected substrings forms a motif represented by the position weight matrix (PWM). To identify motifs that are over-represented in the input sequences, the program calculates the number of matches of corresponding PWMs in the input sequences and in the reference sequence. The reference sequence can be provided as an input parameter to the program or is randomly generated with the probability of each residue taken from the input sequences. To achieve satisfactory performance, we used the PWM matching algorithm proposed by Beckstette et al. (2004). Nevertheless, we decided to replace the originally used enhanced suffix arrays with the index based on the wavelet tree data structure (Grossi et al. 2003), which is more efficient than a suffix array and consumes fewer memory resources. Next, for each motif, the program computes the ratio of the number of matches of PWM in the input sequences to the number of matches in the reference sequence. Motifs with a ratio that is greater than the value of the input parameter \( v \) are qualified to the final step, where the program calculates the percentage of input sequences matching the corresponding PWM. If the calculated percentage is greater than the value of the input parameter \( d \), the motif is reported as over-represented in the input sequences.

In this research, we focused on identification of 7- and 8-mers (parameters \( m = 7, s = 6 \) and \( m = 8, s = 7 \)) in all intronic sequences, separately for each species and separately for 3’ and 5’ splice sites as well as independently for short (<120 bases) and long introns. Here, we required that the level of random occurrences of deviation (parameter \( \nu \)) was at least 3 and the fraction of sequences with at least one occurrence of the motif was 0.05 (parameter \( d \)).

Over-representation of the discovered motifs was further verified statistically. Let \( K \) denote the number of occurrences of a particular motif in all analyzed intronic regions (it is a sum of occurrences of all sequences from a corresponding cluster) and \( N \) be the number of all \( m \)-length windows in the data. Additionally, let \( p \) indicate the theoretical probability of finding a particular motif in a randomly generated sequence (the distribution of ACGT symbols is taken from the data). For each potential cis-regulatory element, we calculated the \( P \)-value as the probability of a motif occurring in the analyzed intronic regions by chance \( i \geq K \) times:

\[
p-value = P(i \geq K) = 1 - P(i < K) = 1 - \sum_{i=0}^{K-1} \left( \frac{N}{i} \right) p^i (1 - p)^{N-i}
\]

Obtained \( P \)-values were always <0.0001. Owing to the fact that SREs are supposed to be over-represented, this is strong statistical support for us having found functional SREs. However, we were unable to estimate the false-negative rate as there is no comprehensive and experimentally verified set of plant SREs.

Independently, we searched for cis-regulatory elements within 50 nt of exonic sequences surrounding introns that are retained in at least one splice form. In this case, due to a limited number of sequences, the calculations were performed simultaneously for all species and all introns of this type, and the \( \nu \) parameter was set to 4.

Additionally, we used exonic splicing enhancer predictions in A. thaliana by Pertea et al. (2007), which include 84 hexamers. Thirty-five of these motifs were shown experimentally to affect splicing. We mapped the hexamers to exonic sequences located 50 bases upstream and downstream of the splice sites in all analyzed plant species.

Providing experimental support for splice sites

In order to provide experimental evidence for the analyzed splice sites, we mapped EST sequences to all transcripts in the corresponding plant species and referred the mapping results to annotated splice site coordinates. We used Megablast here and required that the identity is at least 98% and that the alignment length constitutes at least 90% of the EST length. We made sure that the ESTs do not map to transcripts of other genes with higher BLAST scores. We kept only those alignments where ESTs overlapped with known exon–exon junction(s) and a part of the alignment spanned over at least 10 bases of both exons. Additionally, we used RNA-Seq reads from seven libraries in four plant species (Supplementary Table S1).
The reads were filtered using our in-house Python scripts, and only those with a minimum quality score of 20 over 95% of bases were kept. Reads spanning known introns were detected using TopHat2 (Trapnell et al. 2009), with the gene model annotations supplied by the Ensembl Plants release 15 (Kersey et al. 2010).

Identification of orthologous splice sites

We extracted 5’ and 3’ splice sites from orthologous genes; these were truncated to contain up to 100 bases of exonic sequence and up to 120 bases of intronic sequence, depending on the total exon and intron lengths. In order to find out which splice sites are orthologous, we performed a BLAST search and filtered out the cases where the minimum e-value for an alignment was 1 e-5, both query and subject were in the same orientation in BLAST HSPs (high-scoring segment pairs), and query and subject represented splice sites of orthologous genes. The orthologous splice sites were subsequently re-aligned using ClustalW (Thompson et al. 1994) to obtain alignments of the whole splice site sequences, later used for visualization purposes in the database.

U12 intron search

U12 intron data from U12DB served as a positive data set (true U12 introns). The negative data set was comprised of C. elegans splice sites as this nematode species is devoid of the U12 spliceosome and U12 introns (Burge et al. 1998). Using splice site and branch site sequences from the positive data set, we generated PWMs and applied them to calculate a set of features for both positive and negative data sets. The set of 50 features corresponded to 49 nucleotide positions in splice sites and branch sites that were considered, as well as the position of the branch site in relation to the 3’ splice site. The 49 nucleotide positions included 10 nt of exonic and 15 nt of intronic sequence at the 5’ splice site, 6 nt from exonic and 9 nt from the intronic sequence at the 3’ splice site, as well as 9 nt from the branch site. We used implementation of the random forest machine learning algorithm in Weka 3.6.8 (Frank et al. 2004). The settings were default except for the number of trees, set from 10 to 100. We also used an in-house plugin written in Java to balance between sensitivity and specificity by thresholding the class conditional probability function, and we set the plugin to higher specificity at the cost of sensitivity. Using this approach, we generated a classification model that was finally applied to discriminate between U2 and U12 introns in plants. We performed the analysis for all species except for C. reinhardtii which does not contain U12 introns (Bartschat and Samuelsson 2010) and A. thaliana that was used to train the classifier, and U12DB data were used instead. In the 10-fold cross-validation performed on the full data set, the procedure yielded an exceedingly high sensitivity and specificity, 99.517% and 99.997%, respectively. To assess the classification performance further, we ran the classifier for C. reinhardtii splice sites and obtained seven false positives from the set of 102,382 redundant introns.

Splice sites in microRNA genes

In the first step, we identified EST sequences that correspond to known pre-miRNAs from miRBase (Kozomara and Griffiths-Jones 2011). We searched the ESTs from dbEST (Boguski et al. 1993) using Megablast (Altschul et al. 1990) and required that the identity is >97% and that the EST sequence contains at least 90% of known pre-miRNA sequence. The selected ESTs were subsequently mapped to the corresponding genome using Splign (Kapustin et al. 2008) with default settings. The alignments were finally checked manually to remove cases where ESTs came from the antisense strand and to improve the alignment if the splice site was broken because of an imperfection of the EST alignment software. For A. thaliana we additionally downloaded the sequences from RACE experiments (Szarynska et al. 2009) and used a similar approach to that in the case of ESTs. Here, however, it was not required that the RACE product contains 90% of pre-miRNA sequence as the sequences have already been assigned to corresponding pre-miRNAs and often contained <90% of pre-miRNA sequence.

WebLogos and PWMs

WebLogos were generated using the WebLogo (Crooks et al. 2004) command line client with default settings except for -c (for color logos) and -w (logo width, depending on the number of bases in a logo). All splice sites as well as branch sites of both U2 and U12 introns were used. PWMs were generated from the same data sets as WebLogos using in-house Python scripts.

ERISdb: Database Composition and Usage

Browse

The Browse page gives access to all 1,610,648 splice sites in the analyzed plant species. In order to make browsing through this vast amount of data more straightforward, we divided it into three steps. (i) Species selection. Here, one can choose from eight species: A. thaliana, C. reinhardtii, G. max, O. sativa, P. patens, S. moellendorffii, V. vinifera and Z. mays. (ii) Transcript selection. In this page, one can browse through all transcripts in the selected species. Along with transcript names, also the description, genomic coordinates and transcript biotype are shown. One can filter the data by selecting transcript biotype, chromosome or specifying the gene name or a keyword (this searches through the description data). Alternatively, one can input the transcript name and proceed directly to the third step. (iii) Splice site selection. In this view, two integrated parts help select the splice site of interest: the splice site selection panel with numbered 5’ and 3’ splice sites and a graphical representation of the transcript exon–intron structure. When hovering over the splice site selection panel, the green line marker moves towards the corresponding splice site in the gene structure, thus enabling the user to select the desired
splice site precisely. After clicking, the user is redirected to the splice site data page.

**Splice site data page**

This page displays the following data. (i) General information on the splice site, e.g. intron type (U2 or U12) or splice site sequence. (ii) Orthologs: an alignment of the selected splice site and orthologous sequences. Direct links to orthologous splice site data are provided. (iii) ESTs: ESTs that support the splice site are shown in alignment with the genomic sequence. The intronic sequences are truncated to 10 bases at both termini for display purposes and the alignment itself is truncated to 100 bases. One can also view full alignments, where the splice site of interest is marked in yellow as it often happens that the alignment spans over more than one splice site. (iv) RNA-Seq: detected by TopHat2, a block of exonic sequences that spans over a selected splice site and is supported by RNA-Seq data is marked with yellow. (v) PPT, BS, UA tract: these features are displayed together as alignment to the intronic sequence. In the case of a branch site (BS), the score value is provided which details the similarity of the identified branch site to the fungal consensus sequence of TACTAAC. (vi) cis-Regulatory elements: here, alignments of exonic and intronic SREs to the splice site sequence are provided. Putative intronic SREs, exonic SREs identified by Pertea et al. (2007) and exonic SREs identified for retained introns are marked with distinct colors. In the case of intronic SREs, upon clicking a selected sequence, the associated information is displayed: a WebLogo, PWM and sequence score based on the PWM.

**U12 introns**

This page grants direct access to all U12 introns identified in seven plant species. The data can be filtered by species or type of terminal dinucleotides (GT–AG, AT–AC or other). One can also select a gene of interest from a drop-down list. The displayed data include 5′ and 3′ splice site sequences as well as orthologous genes with a U12 intron. The highly conserved intronic sequence at the 5′ end and terminal dinucleotide at the 3′ end are marked with orange, while branch sites are marked with yellow.

**MicroRNA genes**

Here, one can choose from ERISdb predictions (45 miRNAs), Ensembl annotations (eight miRNAs) or miRNA introns supported by RNA-Seq (three miRNAs). Upon selection of an miRNA of interest from ERISdb predictions, experimentally supported gene structures are displayed. The view includes alignments of three sequences: EST, pre-miRNA and the corresponding genomic sequence. The intronic sequences, if longer than 100 bases, are truncated to 50 bases at both ends. In the case of Ensemble Plant miRNAs, the user is redirected to the splice site data page in ERISdb. Finally, for three *V. vinifera* miRNAs with RNA-Seq support for introns, we provide an alignment of reads to the predicted splice sites.

**WebLogos**

WebLogos for 3′ and 5′ splice sites as well as branch sites in corresponding species, separately for U2 and U12 introns, are presented here. Additionally, the corresponding PWMs for splice sites and branch sites are available.

**Other pages**

ERISdb supports download of various types of computed data as well as selected software used during the analyses. The data files along with short description are available from the Download page. The Help page guides the user through the database structure and usage and provides examples for clarity.

**Discussion and Conclusions**

There are three signals that play key roles in correct splicing of introns: (i) 5′ and 3′ splice sites; (ii) PPTs and/or UA-rich tracts in plants; and (iii) branch sites. De novo identification of splice sites was not a subject of this research as we relied on the gene structure annotation from Ensemble Plants. Instead, we provide experimental support for splice sites (see the Methods). As for PPTs, it still appears unclear whether they always play the same roles as in metazoan introns, but growing experimental evidence supports the idea that at least in some genes PPTs are essential for effective splicing and are located in the canonical position between a putative branch site and the 3′ splice site (Brown et al. 2002) though they are less pronounced (Schwartz et al. 2008) and in fact they are lacking in a number of introns. In this research, we classified the CU-rich regions which were found into two categories, keeping in mind that an upstream branch point is required for PPTs to be functional (Simpson et al. 2004): a putative PTT if there was branch site found upstream, or a CU tract if the tract overlapped with a branch site or a branch site was not found.

In plant introns, the picture is even more complex, as they often possess UA tracts, crucial for effective splicing in both U2 and U12 introns (Goodall and Filipowicz 1989, Lewandowska et al. 2004). Although they may be found anywhere throughout the intron, some of them are located between the putative branch site and the 3′ splice site, and it was demonstrated in the potato invertase gene that the same U-rich sequence can function as either a PPT or a UA-rich element, depending on the presence or absence of a functional branch point upstream (Simpson et al. 2004). Keeping this in mind, we present branch sites, PPTs and UA tracts together in an alignment with intronic sequence in the splice site data page to make the data more informative, but we refrain from considering UA-rich tracts as PPTs even if they fall into the canonical PPT location.

This study is the first attempt to search globally for cis-regulatory elements in plant introns. The only other study was focused on exonic splicing enhancers in *A. thaliana* (Pertea et al. 2007). We have identified putative intronic cis-acting elements in 797,423 plant splice sites, mostly at 3′ splice sites (Table 1). As more than half of the analyzed splice sites do
not possess discovered elements, we assume that this might be caused by the usage of exonic splicing elements instead. It is also possible that there are rare and less conserved intronic elements, which are difficult to discover. Additionally, we searched for putative exonic cis-acting elements that might account for intron retention events, and we uncovered two clusters of over-represented 8-mers that possess a consensus sequence of CGSCGCCG, where S stands for C or G. The sequence was found in exonic sequences associated with 2,393 (13.3%) retained introns. At the same time it could be found for as few as 15,404 (0.97%) remaining introns. We decided to focus on intron retention as it is common in plants and accounts for 56% of alternative splicing events in Arabidopsis thaliana (Wang and Brendel 2006). Furthermore, the process might have considerable implications for post-transcriptional regulation of gene expression by affecting target sites for miRNAs, as predicted for humans (Tan et al. 2007), or through the process of nonsense-mediated mRNA decay (Lewis et al. 2003). U12 introns in A. thaliana were first discovered over a decade ago (Shukla and Padgett 1999), yet since then little has been done to identify them in other plant species. Although they are relatively rare (detected in 0.22% of genes in this study), these introns are shown to be implicated in a number of phenomena, and the ability to discriminate between U2- and U12-type introns might be crucial to untangle splicing mechanisms and their regulation in some introns. To this end, using a machine learning approach (random forest algorithm), we generated a highly sensitive and specific classifier that was further used to search for U12 introns in six plant species. Our approach is superior to the widely used PWM-based approach as it is able to learn quite complex relationships between analyzed features based on positive (U12 introns) and negative (U2 introns) input data. In addition, we took advantage of a branch site position in a 3’ splice site, as it was suggested that this feature might possess discriminative power for a search for plant U12 introns (Lewandowska et al. 2004). Altogether we have identified U12 introns in 2,041 plant genes (Table 1; Supplementary Table S2), 82.4% of which were found to have orthologs with U12 introns as well. A fraction of the remaining U12 introns could possess U2 orthologs, as it is widely observed that U12 introns tend to convert to U2 introns (Lin et al. 2010).

In contrast to animal miRNAs, which are often derived from introns or untranslated regions of genes, plant miRNAs are usually transcribed from dedicated MIR genes. The transcription is usually guided by RNA polymerase II, and the resulting primary transcripts, called pri-miRNAs, are capped, polyadenylated and frequently contain introns (Szarzynska et al. 2009, Mica et al. 2010). Despite the fact that it might be important to know the miRNA gene structure for applications in biotechnology or molecular biology, very few studies have been performed to determine the miRNA gene structures in plants: in A. thaliana through direct cloning of pri-miRNAs (Szarzynska et al. 2009) and in V. vinifera using deep sequencing of the whole transcriptome (Mica et al. 2010). These studies generally show that in plant miRNAs there is a wide variation of transcript forms due to frequent events of alternative splicing and usage of alternative transcription starts sites and alternative polyadenylation sites, yet they uncovered only a fraction of miRNA splice sites as cloning is a laborious task and also because pri-miRNAs have relatively short physiological half-lives. However, previously we demonstrated that novel miRNAs can be successfully discovered from EST sequences, both in plants and in animals (Szczesniak et al. 2012). Encouraged by these results, in this study we performed analyses aimed at determination of miRNA gene structures based on alignment of EST sequences to the corresponding genome. We managed to find introns in 45 miRNA precursors in five plant species. Some of the miRNAs contain multiple introns (up to six), there are also several cases of alternative splicing via intron retention (Fig. 1). It is important to point out that when presented in ERISdb, exon–intron structures do

Table 1 Summary of data stored in ERISdb

<table>
<thead>
<tr>
<th>All introns</th>
<th>1,610,648</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supported by EST</td>
<td>704,744 (43.8%)</td>
</tr>
<tr>
<td>Supported by RNA-Seq</td>
<td>425,777 (26.4%)</td>
</tr>
<tr>
<td>CU tract found</td>
<td>856,226 (53.2%)</td>
</tr>
<tr>
<td>UA tract found</td>
<td>1,431,052 (88.8%)</td>
</tr>
<tr>
<td>Branch site found</td>
<td>920,403 (57.1%)</td>
</tr>
<tr>
<td>5’ splice site with intronic SREs</td>
<td>320,110 (19.9%)</td>
</tr>
<tr>
<td>3’ splice site with intronic SREs</td>
<td>477,313 (29.6%)</td>
</tr>
<tr>
<td>5’ splice site with identified ortholog</td>
<td>331,794 (20.6%)</td>
</tr>
<tr>
<td>3’ splice site with identified ortholog</td>
<td>331,498 (20.6%)</td>
</tr>
<tr>
<td>U12 introns</td>
<td>3,039 (0.22%)</td>
</tr>
<tr>
<td>SREs associated with retained introns</td>
<td>15,404 (0.97% of non-retained introns)</td>
</tr>
<tr>
<td></td>
<td>2,393 (13.30% of retained introns)</td>
</tr>
<tr>
<td>Hexamers from Pertea et al. (2007)</td>
<td>1,422,984 (88.35%)</td>
</tr>
<tr>
<td>Intronic SREs</td>
<td>1,055,021 (65.50%)</td>
</tr>
</tbody>
</table>

* Calculated for exonic sequences associated with an intron.

![Fig. 1](https://example.com/schematic.png)

**Fig. 1** Schematic representation of gene structures of three plant miRNAs predicted by us. Pre-miRNA is marked in red. (a) osa-miR156d is produced from a precursor containing six introns and is a record holder in this respect. (b) osa-MIR444c has exceedingly long introns (2,784 and 6,722 bases), and both of them reside in the pre-miRNA sequence. (c) ppt-MIR536c: an example of an alternatively spliced intron in an miRNA gene.
not correspond to the full gene architecture as this would require a search for promoters and/or transcription starts sites as well as polyadenylation sites, while we focused on identification of splice sites.

ERISdb is a comprehensive database of splice sites and splicing signals. In the current version, it provides data for eight plant species. The data stored in ERISdb include (i) general information on splice sites such as WebLogos, branch sites or PPTs; (ii) experimental evidence for annotated splice sites (ESTs and RNA-Seq data); (iii) orthologous splice sites; (iv) putative intronic and exonic cis-regulatory elements; (v) U12 introns; and (vi) introns in miRNA genes. This is the first database to cover such a diverse set of plant splice site-associated data, most of which are not available from any other resource. We also developed novel tools and made them available for download from the Download page. The novelty of database content along with the aesthetic, user-friendly design of the web interface and data download options should make ERISdb a very useful resource for further research in plant molecular biology.

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported by the National Science Centre [2011/01/N/NZ2/01653 to M.W.S. and I.M.]; the European Social Fund [UDA-POKL.04.01.00-106/09 to A.G. and R.P.].

References


