VERSITA

Cent. Eur. J. Biol. • 7(4) • 2012 • 620-625
DOI: 10.2478/s11535-012-0060-1

Central European Journal of **Biology**

Natural rules for *Arabidopsis thaliana* pre-mRNA splicing site selection

Research Article

Ning Wu[1,*], Kanyand Matand[1,2,#], Huijuan Wu[3], Baoming Li[3], Kayla Love[4], Brittany Stoutermire[3], Yanfeng Wu[2]

[1]*Center for Biotechnology Research and Education, Langston University, Langston, Oklahoma 73050, USA*

[2]*Department of Biology, School of Arts and Science, Langston University, Langston, Oklahoma 73050, USA*

[3]*Department of Biology, Beijing Center for Physical and Chemical Analysis, 100089 Beijing, China*

[4]*Department of Chemistry, School of Arts and Science, Langston University, Langston, Oklahoma 73050, USA*

Received 13 February 2012; Accepted 08 May 2012

Abstract: The accurate prediction of plant pre-mRNA splicing sites has been studied extensively. The rules for plant pre-mRNA splicing still remain unknown. This study, based on confirmed sequence data, systematically analyzed all expressed genes on *Arabidopsis thaliana* chromosome IV to quantitatively explore the natural splicing rules. The results indicated that defining *Arabidopsis thaliana* pre-mRNA splicing sites required a combination of multiple factors including (1) relative conserved consensus sequence at splicing site; (2) individual nucleotide distribution pattern in 50 nucleotides up- and down-stream regions of splicing site; (3) quantitative analysis of individual nucleotide distribution by using the formulations concluded from this study. The combination of all these factors together can bring the accuracy of *Arabidopsis thaliana* splicing site recognition over 99%. The results provide additional information to the future of plant pre-mRNA splicing research.

Keywords: *Arabidopsis thaliana* • *Pre-mRNA splicing*

1. Introduction

The accurate removal of introns from precursor messenger RNA (pre-mRNA) is an essential procedure for gene expression. Several key factors involved in intron removal have been suggested by previous studies. These include: the relative consensus sequence at the splicing sites, the individual nucleotide distribution pattern in flank regions of splicing sites, and a conserved branch point sequence [1,2]. According to past studies, some conserved short sequences within introns have been identified to function in intron splicing across all species, which include dinucleotide guanine (G) and uridine (U) at 5' splicing site and adenine (A) and guanine at 3' splicing site, and a short conserved "branch point" sequence located within 50 nucleotides upstream of the 3' splicing site [3]. However, particular structural and sequence features distinguish plant introns from the other species. Although plant introns share a high level sequence similarity with other species, there is a lack of a conserved branch point sequence when comparing vertebrate and yeast introns [4,5]. Former studies demonstrated that alternation of the sequences around splicing sites might lead to the change of intron/exon recognition mechanisms [6] and lead to different transcripts and functions [7], which indicated that splicing of a particular intron depends on a fine balance of multiple splicing signals of varying strengths in the sequence context of an intron/exon organization [5,8]. Based on current available information, several plant splicing site prediction methods have been developed [9-11]. However, because of the lack of an *in vitro* system capable of efficiently splicing plant introns experimentally and

*E-mail: *nwu@luresext.edu, #kmatand@luresext.edu*

the complicacy of plant species, there are many uncertainties of plant pre-mRNA splicing mechanisms [12]. The signals that specifically define the borders of splicing sites in plant are still not fully understood due to lack of a complete and accurate description of the gene structure on the basis of sequence [13]. Recently, upon the genome research proceeding into the post-genome era, the comprehensive data collection of genomics and functional genome over many major species provides scientists the opportunity to review and analyze those confirmed sequence data by using different bioinformatical means and allows them to illustrate the potential mechanism of pre-mRNA splicing. Some researchers have applied large confirmed human gene samples in their studies to improve the results' accuracy and sensitivity [14]. However, to date, there have been no reports of such a study in plant genomics research. As a representative of plant, *Arabidopsis thaliana* has been well studied for years. Both the complete genome and the most expressed gene sequences are available for analysis. This study employed current available genomic data to explore all confirmed, natural expressed genes on *Arabidopsis thaliana* chromosome IV to illustrate the potential intrinsic splicing site strength and the natural rules for splicing site selection.

2. Experimental Procedures

2.1 Data source

Through the "National Center for Biotechnology Information" (NCBI) Plant Genomes Central, *Arabidopsis* chromosome IV genome sequence has been targeted for this study (http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?taxid=3702&chr=4). All 18.6 million bp nucleotide sequences with 5,122 genes in the region were focused on. The complete sense sequence structure of each individual gene was retrieved from *Arabidopsis thaliana* database (http://mips.gsf.de/proj/thal/db/index.html) located at the "Munich Information Center for Protein Sequences" (MIPS, Germany) through the NCBI on site linkage. All the sequences were downloaded to the local computers.

2.2 Sequence data manipulation

For each downloaded gene sequence, all exon and intron nucleotide sequences were marked separately according to the current available gene structure information. Additionally, the sequences of 50 nucleotides upstream and downstream of each splicing site were pulled out for the analysis (Figure 1). For the sequences that were shorter than 50 bp on one side of splicing site, manual sequence trimming on the other side was performed to bring equal length of sequences in both exon and intron flank regions. The data was divided into four groups including (1) 3' end of exon sequence, (2) 5' end of intron sequence, (3) 3' end of intron sequence, and (4) 5' end exon sequence. All the data was integrated into Microsoft Excel data sheet.

2.3 Sequence data analysis

Microsoft Excel program was applied for this study. The present frequency of each individual nucleotide on each position through the range of 100 nucleotides centered by the splicing site was calculated. The relative conservative nucleotide sequences across 5' and 3' splicing sites were analyzed (Table 1). The splicing sites that showed complete 50 bp up- and down-stream of splicing site were defined as "full sequence" sites and were used as the experimental data for statistical analysis. The occurrence of each nucleotide in 50 bp regions up- and down-stream of each individual splicing site was calculated. A series paired group t-tests were performed to inspect distribution difference of each type of nucleotide between up- and down-stream regions of 5' and 3' splicing sites. The exact number differences of A, G, C (cytosine), U nucleotides across each splicing site was also calculated and analyzed. The splicing sites that showed shorter sequences (less than 50 bp) in either exon or intron flank region were defined as "shorter sequence" sites and were used as the test group for results examination.

Figure 1. Individual gene exon/intron sequence data mining.

A: 5' splicing site

Position	E-2	E-1	I-1	I-2	I-3
A	8986	1512	17	26	9352
G	1248	10744	14281	24	1843
C	1529	547	8	99	701
U	2557	1517	14	14171	2424
Consensus sequence	A	G	G	U	A

B: 3' splicing sites

Position	I-5	I-4	I-3	I-2	I-1	E-1	E-2
A	2388	3801	1070	14397	20	3679	3513
G	1558	5321	138	17	14408	7465	2567
C	1609	1255	9078	10	7	1604	2097
U	8892	4070	4161	23	12	1699	6270
Consensus sequence	U	N	C	A	G	G	U

Table 1. (A) Nucleotide distribution around 14,320 of 5' splicing sites. (B) Nucleotide distribution around 14,447 of 3' splicing sites. E: exon region; I: intron region. N: any nucleotide.

3. Results and Discussion

Total 18.6 million nucleotides with 5,122 genes on *Arabidopsis thaliana* chromosome IV were downloaded from the Munich Information Center for Protein Sequences (MIPS, Germany) (http://mips.gsf.de/proj/thal/db/index.html). The flank sequences with 50 nucleotides up- and down-stream of each splicing site were defined as "full sequence" and grouped by their locations for analysis. For sequences shorter than 50 nucleotides - "shorter sequence" on one side of splicing site, manual sequence trimming on the other side was performed to bring equal length of sequences in both exon and intron flank regions. Totally 30,870 splicing sites (15,410 of 5' sites and 15,460 of 3' sites) were included in this study with 28,767 "full sequence" sites and 2,103 "shorter sequence" sites, which covered 3,038,104 nucleotides.

The consensus sequences on both 5' and 3' splicing sites were determined by calculating the distribution probability of each nucleotide (Table 1). The relatively consensus sequences on 5' and 3' splicing sites were observed as AG|GUA and UNCAG|GU (N could be any nucleotide) respectively, which is consistent with previous studies [3,12,15]. The results indicated that consensus sequences at splicing sites are important for site recognition but are not enough to define the sites alone.

The distributions of A, G, C, U among all 28,767 "full sequence" were calculated respectively. Paired group sample t-tests had been performed to examine the appearance difference of each type of nucleotide between up- and down-stream flank regions. The analysis showed the same results in both 5' and 3' splicing sites with the number of U in intron flank region larger than in exon flank region. This result was reversed for A, G, and C (Figure 2). Statistical analysis showed that there were significant differences in all four types of nucleotides with all the P values equal to or approximating zero. The results provided strong evidence to support the theory that the distribution patterns of nucleotides in flank regions of splicing sites are the splicing signals which define the 5' and 3' splicing sites [10,16]. In addition to previous reported U, G, and C, this study also proved that A also played an important role in splicing site determination.

Based on statistical results, the differences of each nucleotide between up- and down-stream regions were calculated. Among 14,320 of 5' splicing sites, 13,061 sites showed the numbers of U in exon flank regions (U) less or equal to it in intron flank regions (u) as U-u≤0 (formulation-1). The remaining 1,259 sites (U-u>0), about 91.58%, showed the numbers of A in exon flank region (A) less or equal to it in intron flank regions (a), which was defined as "if U-u>0, then A-a≤0" (formulation-2). Combining the results of formulations 1 and 2, it covered 14,214 out of 14,320 of 5' splicing sites (Table 2). On 3' splicing site, 13,496 out of 14,447 splicing sites showed the numbers of U in intron flank regions (u) larger than it in exon flank regions (U) as u-U≥0 (formulation-3). The left over 951 of u-U<0 sites showed the numbers

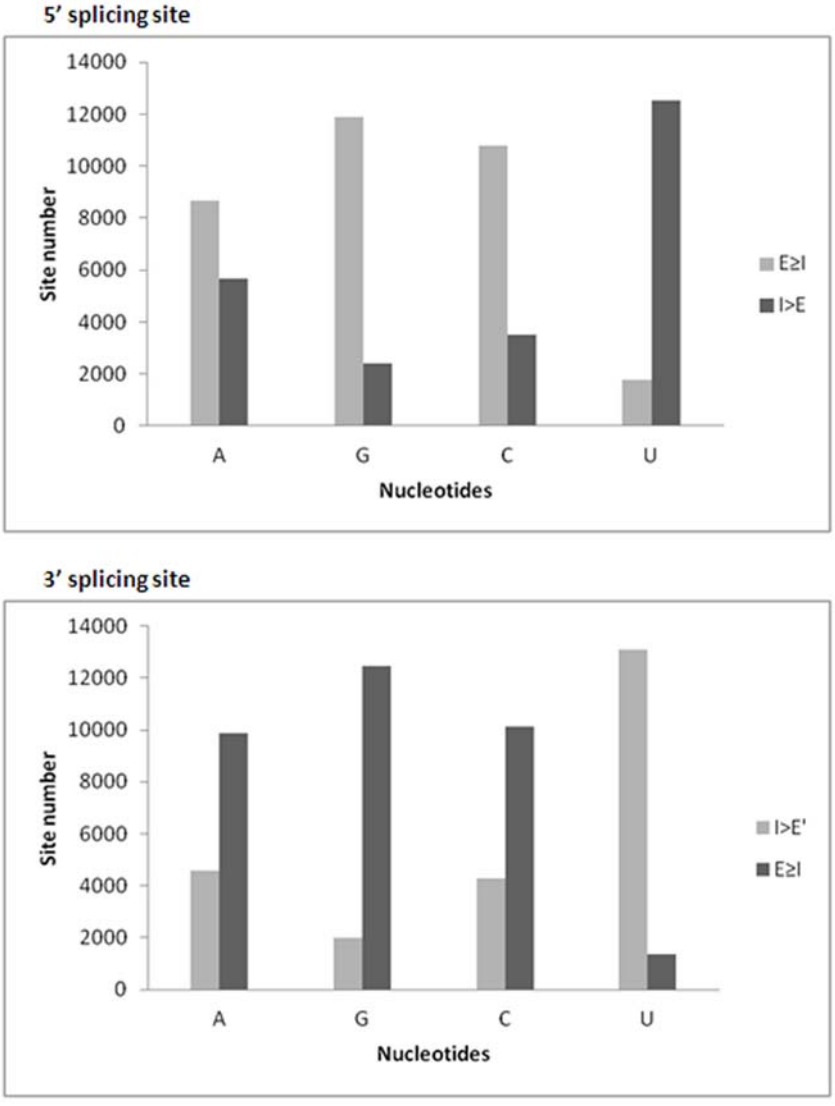

Figure 2. Nucleotide distribution in both 5' and 3' splicing site flank regions. I>E: the number of nucleotide in intron larger than it in exon. E≥I: the number of nucleotide in exon larger or equal to it in intron.

of A in intron flank region (a) larger or equal to it in exon flank regions (A), which was expressed as "If u-U<0, then a-A≥0" (formulation-4). Combining the results of formulations 3 and 4, it covered 14,346 out of 14,447 3' splicing sites (Table 3). The results indicated that, in addition to the consensus sequence, if applying intronic U-rich sequences only (formulation-1 or 3), there were 91.21% of 5' splicing sites and 93.42% of 3' splicing sites fulfilled the criteria, respectively. The left over splicing sites had to be defined by additional features. A previous study has shown that the relative contrast in U and G+C content between introns and their flanking exons correlates with splicing efficiency [17]. However, our results show that, after applying formulation-1 or 3, neither G nor C could provide sufficient evidence to define the left over splicing sites when compared to A (Table 2 and 3). By combining the applications of formulations (1 and 2, 3 and 4), the accuracies for defining 5' and 3' splicing site reached 99.26% and 99.30% respectively. It was interesting that, if the general rule of intronic U-rich sequence was broken, A became the next appropriate nucleotide for splicing site defining with contrast distribution pattern to its general format.

To test the efficiency and accuracy of formulation 1 to 4, we applied them to examine 2,103 of splicing sites with "shorter sequence". In total 1,090 of 5' splicing sites and 1,013 of 3' splicing sites were tested. The shortest exon sequences were the starting codon (AUG) only and always showed the same sequence length in

Formulation-1: U-u≤0

	A-a	G-g	C-c	U-u
≤0	6686	3420	4722	13061
>0	7634	10900	9598	1259
Subtotal	14320	14320	14320	14320

Formulation-2: If U-u>0, then A-a≤0

	A-a	G-g	C-c	U-u
≤0	1153	486	559	0
>0	106	773	700	1259
Subtotal	1259	1259	1259	1259

Table 2. The number differences of each nucleotide around the 5' splicing site flank regions. (Uppercase: exon region; Lowercase: intron region).

Formulation-3: u-U≥0

	a-A	g-G	c-C	u-U
<0	8761	11553	8734	951
≥0	5686	2894	5713	13496
Subtotal	14447	14447	14447	14447

Formulation-4: If u-U<0, then a-A≥0

	a-A	g-G	c-C	u-U
<0	101	571	427	951
≥0	850	380	524	0
Subtotal	951	951	951	951

Table 3. The number differences of each nucleotide around the 3' splicing site flank regions. (Uppercase: exon region; Lowercase: intron region).

the beginning of intron sequences as GUA. The results of other shorter sequence sites demonstrated the accuracies of 98.44% on 5' splicing sites (formulation-1 and 2) and 98.42% on 3' splicing sites (formulation-3 and 4) respectively. In some rare cases that both U and A distributions were unable to define splicing site, the distribution of C in exons (less than what it would be in introns) would be the next factor to determine the splicing sites, which brought the total accuracies of "shorter sequence" site to 99.63% on 5' splicing site and 99.90% on 3' splicing site (data not shown).

The accurate prediction of plant mRNA splicing sites has been extensively studied. Several different computer-aided splicing site selection models that were believed to mimic the *in vivo* splicing process mathematically have been developed based on several key factors. These factors include: the consensus splicing site sequence, intronic U-rich sequence, and the branch point confirmed through the different species (vertebrate, yeast) [9,10,15,18]. However, the rules and mechanisms for plant pre-mRNA splicing remain unknown. Our study, based on the confirmed sequence data, systemically analyzed all expressed gene structures on *Arabidopsis thaliana* chromosome IV to quantitatively explore the natural splicing rules. In conclusion, defining *Arabidopsis thaliana* pre-mRNA splicing sites requires the combination of multiple factors, which includes (1) relative conserved consensus sequence at splicing site; (2) individual nucleotide distribution pattern in 50 nucleotides up- and down-stream regions of splicing site with intronic U-rich sequence and exonic G, C, A-rich sequences; (3) quantitative analysis of individual nucleotide distribution pattern. The calculation preference is as follows (upper case represents exon sequence, lower case represents intron sequence): On 5' splicing site: (i) U-u≤0, (ii) if U-u>0, then A-a≤0, (iii) if U-u>0 and A-a>0, then C-c≤0. On 3' splicing site: (i) u-U≥0, (ii) if

u-U<0, then a-A≥0, (iii) if u-U<0 and a-A<0, then c-C≥0. According to statistical analysis results, the combination of all these factors together can bring the accuracy of splicing site recognition over 99%. The results of this study provide new data and extend previous research findings which will aid in the future of plant pre-mRNA splicing research.

Acknowledgements

Dr. Ning Wu and Dr. Kanyand Matand have the equal contribution to this study and both are listed as corresponding authors. All listed authors have agreed this submission without conflict of interest.

References

[1] Green M.R., Biochemical mechanisms of constitutive and regulated pre-mRNA splicing, Annu Rev Cell Biol.,1991, 7, 559-599

[2] Wang Z., Burge C., Splicing regulation: from a parts list of regulatory elements to an integrated splicing code, RNA, 2008, 14, 802-813

[3] Lal S., Choi J.H., Shaw J.R., Hannah L.C., A splice site mutant of maize activates cryptic splice sites, elicits intron inclusion and exon exclusion, and permits branch point elucidation, Plant Physiol., 1999, 121, 411-418

[4] Goodall G.J., Filipowicz W., Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants, EMBO J., 1991, 10, 2635-2644

[5] Brown J.W., Simpson C.G., Splice site selection in plant pre-mRNA splicing, Annu Rev Plant Physiol Plant Mol Biol., 1998, 49, 77-95

[6] Simpson C.G., McQuade C., Lyon J., Brown J.W., Characterization of exon skipping mutants of the COP1 gene from Arabidopsis, Plant J., 1998, 15, 125-131

[7] Lazar G., Goodman H.M., The Arabidopsis splicing factor SR1 is regulated by alternative splicing, Plant Mol Biol., 2000, 42, 571-581

[8] Sarmah B., Chakraborty N., Chakraborty S., Datta A., Plant pre-mRNA splicing in fission yeast, Schizosaccharomyces pombe, Biochem Biophys Res Commun., 2002, 293, 1209-1216

[9] Lou H., McCullough A.J., Schuler M.A., 3' splice site selection in dicot plant nuclei is position dependent, Mol Cell Biol, 1993, 13, 4485-4493

[10] Brendel V., Kleffe J., Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA, Nucleic Acids Res, 1998, 26, 4748-4757

[11] Luehrsen K.R., Taha S., Walbot V., Nuclear pre-mRNA processing in higher plants, Prog Nucleic Acid Res Mol Biol, 1994, 47, 149-193

[12] Hebsgaard S.M., Korning P.G., Tolstrup N., Engelbrecht J., Rouze P., Brunak S., Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information, Nucl. Ac. Res., 1996, 24, 3439-3452

[13] Guigó R., Flicek P., Abril J.F., Reymond A., Lagarde J., Denoeud F., et al., EGASP: the human ENCODE Genome Annotation Assessment Project, Genome Biol, 2006, 7, 1-31

[14] Dogan R.I., Getoor L., Wilbur J., Mount S., Features generated for computational splice-site prediction correspond to functional elements, BMC Bioinformatics, 2007, 8, 410-424

[15] White O., Soderlund C., Shanmugan P., Fields C., Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin, Plant Mol Biol, 1992, 19, 1057-1064

[16] Kleffe J., Hermann K., Vahrson W., Wittig B., Brendel V., Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences, Nucl. Ac. Res., 1996, 24, 4709-4718

[17] Carle-Urioste J.C., Brendel V., Walbot V., A combinatorial role for exon, intron and splice site sequences in splicing in maize, Plant J, 1997, 11, 1253-1263

[18] Brendel V., Kleffe J., Carle-Urioste J.C., Walbot V., Prediction of splice sites in plant pre-mRNA from sequence properties, J Mol Biol, 1998, 276, 85-104