Splicing signals and factors in plant intron removal

Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, U.K.

Abstract
Constitutive splicing of the potato invertase mini-exon 2 (9 nt long) requires a branchpoint sequence positioned around 50 nt upstream of the 5’ splice site of the adjacent intron and a U11 element found just downstream of the branchpoint in the upstream intron [Simpson, Hedley, Watters, Clark, McQuade, Machray and Brown (2000) RNA 6, 422–433]. The sensitivity of this in vivo plant splicing system has been used to demonstrate exon scanning in plants, and to characterize plant intronic elements, such as branchpoint and polypyrimidine tract sequences. Plant introns differ from their vertebrate and yeast counterparts in being UA- or U-rich (up to 85% UA). One of the key differences in splicing between plants and other eukaryotes lies in early intron recognition, which is thought to be mediated by UA-binding proteins. We are adopting three approaches to studying the RNA–protein interactions in plant splicing. First, overexpression of plant splicing factors and, in particular, UA-binding proteins, in conjunction with a range of mini-exon mutants. Secondly, the sequences of around 65% of vertebrate and yeast splicing factors have high-quality matches to Arabidopsis proteins, opening the door to identification and analysis of gene knockouts. Finally, to discover plant-specific proteins involved in splicing and in, for example, rRNA or small nuclear RNA processing, green fluorescent protein-cDNA fusion libraries in viral vectors are being screened.

Introduction
Precursor mRNA (pre-mRNA) splicing requires the early recognition of pre-mRNA splicing signals. In metazoan organisms, the 5’ and 3’ splice sites, and an internal branchpoint associated with a downstream polypyrimidine tract, are essential intron signals. Early in spliceosome assembly, U1 small nuclear ribonucleoprotein particle (snRNP) is recruited to the 5′ splice site where U1 small nuclear RNA (snRNA) base pairs with the pre-mRNA. Branchpoint bridging protein (BBP/SF1) and U2 auxiliary factor of 65 kDa (U2AF65) bind co-operatively to the branchpoint and polypyrimidine tract, after which U2snRNP is recruited and U2snRNA base pairs with to the branchpoint sequence. Interactions between protein factors at the 5′ splice site and branchpoint/poly-pyrimidine tract/3′ splice site lead to formation of the commitment complex and spliceosome [1–3]. In addition, splice-site selection is modulated by other factors interacting with intron and exon splicing enhancer and silencer elements. Thus, splice-site selection in multiple intron pre-mRNA transcripts is co-ordinated through the activity of a number of splicing regulatory factors, including serine-arginine-rich (SR) proteins, that permit co-operative interaction between splice sites along the pre-mRNA transcript and lead to highly regulated splice-site selection [1–6].

In the absence of plant nuclear in vitro splicing extracts, characterization of plant intron splicing signals such as branchpoint, polypyrimidine tract and UA-rich sequences, and the determination of the roles of putative splicing factors and RNA-binding proteins, has been extremely difficult. The requirement for strong splicing signals in the invertase mini-exon system makes it highly sensitive to mutation and provides a system to investigate, in detail, the sequence and spatial requirements of intron signals, and the effects of overexpression of splicing factors. This paper describes the utility of the mini-exon system in such characterizations and more general approaches to identifying proteins involved in RNA processing in plants.

The potato invertase mini-exon system
The potato invertase mini-exon is constitutively included in invertase transcripts and is dependent on strong constitutive splicing signals [7]. The key elements which determine mini-exon inclusion are a branchpoint sequence, an adjacent downstream U-rich region and the distance between...
these signals and the 5′ splice site downstream of the mini-exon. The model for mini-exon splicing is that exon-bridging interactions, between factors at the branchpoint/polypyrimeidin tract in intron 1 and the 5′ splice site of intron 2, promote splicing of intron 2. This is then followed by removal of intron 1 (Figure 1a) [7]. The increased distance of the branchpoint/polypyrimeidin tract from the mini-exon is needed to allow the exon-bridging interactions to occur. Mutation of the splice sites, branchpoint or polypyrimeidin tract, or reduction of the distance between the branchpoint/polypyrimeidin tract and the mini-exon, leads to increased or complete skipping of the mini-exon. The skipping phenotype is readily identifiable and quantifiable by reverse transcriptase PCR.

**Branchpoint and polypyrimeidin tract sequences**

Various lines of evidence suggest that branchpoint sequences that share similarity with those of vertebrate introns may be important in splicing and splice-site selection in plants [8,9]. However, the contribution of sequence context to the strength of plant branchpoints has not been determined. A complete mutational analysis of the branchpoint sequence in intron 1 of the mini-exon system identified the relative requirements for particular nucleotides in every position [10]. This is summarized by the sequences CURAY or YURAY, where the underlined U and A are essential nucleotides, and purines and pyrimidines are preferred nucleotides at positions 3 and 5 respectively. In position 1, pyrimidines are preferred but also C is preferred to U (Figure 1b). These results represent the first detailed characterization of a branchpoint in a plant intron. Furthermore, the preferred sequences match consensus branchpoint sequences of both vertebrate and plant introns [8,11].

Plants do not, in general, contain strong polypyrimeidin tracts as found in vertebrate introns, but tend to have U-rich sequences between putative branchpoints and the 3′ splice site. U- or UA- richness is a general feature of plant introns, essential for efficient splicing [12], and mutations to U-rich regions in a number of plant introns alter splicing efficiency and can activate cryptic splice sites ([13–17] for reviews; [7,18]). Although the position of U-rich sequences, adjacent and downstream of branchpoints, in plant invertase genes suggests that they may be polypyrimeidin tracts, it was necessary to distinguish between a polypyrimeidin tract function and that of a UA- or U-rich intronic element required for efficient plant intron splicing. Mutational series of the U11 sequence in potato invertase intron 1, where increasing numbers of Us were replaced by Cs or As (‘C’ and ‘A’ series respectively), were tested.

---

**Figure 1**

Structure of potato invertase mini-exon and characterization of branchpoint and polypyrimeidin tract

(a) Model of the exon-definition process that promotes splicing of the 9 nt invertase mini-exon [7]. Exons are indicated as boxes and introns (IVS) by lines. Splice sites are shown as GU and AG. The branchpoint (A) and the associated polypyrimeidin tract (pY; a U11 sequence) is over 50 nt from the downstream 5′ splice site in all dicotyledon invertase genes isolated so far.

(b) Mutational analysis of the branchpoint and polypyrimeidin tract generates the CURAY branchpoint sequence, where the size of the letters represents the contribution of particular nucleotides to splicing. The minimum polypyrimeidin sequence for efficient splicing is two groups of Us separated by three to six pyrimidines.
for splicing of the mini-exon. Inclusion of > 2As drastically reduced splicing efficiency, while Cs could compensate to a large extent for Us [10]. These results suggested that the U-rich region functioned as a polypyrimidine tract rather than a U-rich intronic element. To function efficiently, the poly-pyrimidine tract had to contain a minimum of two groups of three or four Us up to 6 nt apart (Figure 1b). Thus at least in some plant introns, polypyrimidine tracts are required for splicing, as is generally the case in vertebrate introns.

**U-rich-binding proteins (UBPs)**

Since the initial realization that UA-rich sequences were required for efficient splicing in plants [12], UA- or U-rich-binding proteins have been postulated to interact with and identify intron sequences within a pre-mRNA transcript [13–17]. RNA-binding proteins with affinity for U-rich sequences have been identified in *Arabidopsis* [19–21]. All have been shown to have high affinity for U-rich sequences *in vitro* and therefore represent candidates for proteins involved in early intron recognition. The most studied of these is UBP-1, which can enhance splicing of poorly spliced introns when overexpressed in protoplasts, but also appears to have a role in mRNA stability [20]. We have made use of the range of mutants in the mini-exon splicing system to begin to investigate differential interactions of such proteins. The A and C series of mutants, designed to distinguish between the polypyrimidine tract and U-rich intronic elements, were used in conjunction with overexpression of UBP-1 and other proteins, following co-transfection of tobacco protoplasts. Despite their similar affinities for U-rich sequences *in vitro*, differential splicing efficiencies were observed for each of the proteins with the C and A series of mutants (C. G. Simpson and J. W. S. Brown, unpublished work). This suggests that, although quite similar in structure and in some cases sequence, subtle differences exist among the proteins in terms of specificity of sequence binding and their effects on splicing of the mini-exon.

**Splicing factor orthologues in *Arabidopsis***

A number of splicing factors have been isolated from plants. These include snRNP, SR and heterogeneous ribonucleoprotein particle (‘hnRNP’) proteins (see [17] for review). The *Arabidopsis* genome sequence has greatly aided the identification of splicing proteins that are conserved among eukaryotes. In an analysis of such proteins, splicing factors with significant homology to counterparts in yeast and vertebrates were identified for around 65% of the eukaryotic proteins (C. G. Simpson and J. W. S. Brown, unpublished work). In some cases sequence similarity was striking, while in others similarity resided in only portions of the protein, such that identification was tentative. In particular, the majority of protein factors involved in mammalian branchpoint, polypyrimidine-tract and 3′-splice-site selection have well-conserved orthologues in

**Figure 2**

Proteins identified to interact at the branchpoint/polypyrimidine tract/3′ splice site during early spliceosomal complex assembly in mammalian introns

<table>
<thead>
<tr>
<th>Complex E</th>
<th>Human</th>
<th>Arabidopsis</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBBP</td>
<td>639aa</td>
<td>32% (668aa)</td>
</tr>
<tr>
<td>U2AF35</td>
<td>475aa</td>
<td>40% (491aa)</td>
</tr>
<tr>
<td>U2AF35</td>
<td>240aa</td>
<td>56% (247aa)</td>
</tr>
<tr>
<td>p14</td>
<td>125aa</td>
<td>64% (121aa)</td>
</tr>
<tr>
<td>SAP155</td>
<td>1304aa</td>
<td>68% (1305aa)</td>
</tr>
<tr>
<td>SAP145</td>
<td>872aa</td>
<td>42% (645aa)</td>
</tr>
<tr>
<td>SAP130</td>
<td>1217aa</td>
<td>58% (1219aa)</td>
</tr>
<tr>
<td>SAP114</td>
<td>793aa</td>
<td>37% (773aa)</td>
</tr>
<tr>
<td>SAP62</td>
<td>464aa</td>
<td>62% (265aa)</td>
</tr>
<tr>
<td>SAP61</td>
<td>501aa</td>
<td>45% (512aa)</td>
</tr>
<tr>
<td>SAP49</td>
<td>424aa</td>
<td>58% (378aa)</td>
</tr>
</tbody>
</table>
Arabidopsis (Figure 2). That many yeast and vertebrate splicing factors failed to identify similar proteins in Arabidopsis suggests that these genes and proteins have diversified in plants or that their functions have been taken over by other plant-specific proteins. To begin to address the functions of identified genes, various tools are available: overexpression of proteins with mini-exon mutants, transfer-DNA or transposon-gene knockouts in Arabidopsis, or virus-induced or RNA-induced gene silencing.

Identification of nuclear proteins

Plant viral vectors have been developed for ectopic expression of RNA and proteins in plant cells. In particular, tobacco mosaic virus-based vectors have been designed for efficient, high-throughput expression in tobacco hosts. cDNA libraries have been constructed from Nicotiana benthamiana mRNA in tobacco mosaic virus vectors, where partial cDNAs are expressed as 3′- or 5′-green fluorescent protein (GFP) fusion proteins. Infection of leaves of tobacco plants produced many hundreds of individual infection sites per leaf, each of which represents a different fusion product. Lesions are rapidly screened by fluorescence microscopy and those showing nuclear or nucleolar localization of GFP are selected. Following isolation of the lesion and passing through another infection cycle, the cDNA is cloned and sequenced. This allows large-scale screening of cDNAs for particular cellular localization. Thus far, GFP localization has been observed throughout the whole nucleoplasm, to putative nuclear pore structures, strands within the nucleus, nuclear speckles and the nucleolus (K. J. Oparka, unpublished work). The lack of an in vitro splicing extract from plants has hampered progress in understanding the nuances and subtleties of plant intron splicing in comparison with other eukaryotic systems. The identification of splicing factors conserved in plants and of novel plant proteins, the technologies for examining gene function in plant cells and sensitive in vivo splicing assay systems, such as the potato mini-exon, provide new and exciting opportunities to study constitutive and alternative splicing in plants.

References


Received 21 November 2001