Alternative Splicing at NAGNAG Acceptor Sites Shares Common Properties in Land Plants and Mammals

Kei Iida, Masafumi Shionyu, and Yasuhiro Suso

Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga, Japan

In recent years, several papers have reported that a special type of alternative splicing (AS) event occurs at the tandem 3’ splice site, termed the “NAGNAG acceptor.” This type of AS event (termed AS-NAGNAG) is well studied in both human and mouse. To illustrate the significance of AS-NAGNAG events, we focused on their occurrence in *Arabidopsis thaliana* and *Oryza sativa* (rice). Our study is the first genome-wide approach examining AS-NAGNAG events in land plants. Based on transcripts and genomic sequences, we found 321 and 372 AS-NAGNAG events in *Arabidopsis* and rice, respectively. These events were significantly enriched in genes encoding DNA-binding proteins, and more than half of all AS-NAGNAG events affected polar amino acid residues. The observed properties of AS-NAGNAG events in plants were similar to those seen in mammals. These results showed that AS-NAGNAG events may provide a mechanism for fine-tuning of DNA-binding proteins in both mammals and land plants. We found 7 gene groups of AS-NAGNAG events that were conserved between *Arabidopsis* and rice, including 2 groups for RNA-binding proteins. Conservation of the events for RNA-binding proteins is a property also seen in introns of homologous genes. They included 5 groups of DNA-binding proteins, whose number was larger than expected. We think there is a bias with which AS-NAGNAG events are fixed in genes for DNA-binding proteins. Our analysis showed that AS-NAGNAG events found in land plants share similar properties with those in mammals. Based on our results, we propose that AS-NAGNAG events are likely to be a common mechanism in the fine-tuning of protein functions, especially DNA/RNA-binding proteins, in both mammals and plants. Their role might contribute to the construction of complicated transcriptomes and proteomes in the evolutionary history of mammals and land plants.

Introduction

Alternative splicing (AS) is a mechanism whereby 2 or more types of different mature mRNAs are generated from a single premature mRNA. Recent reports have found that a significant percentage of genes in plants undergo AS. For example, over 20% of pre-mRNAs corresponding to about 4,700 genes are alternatively spliced in *Arabidopsis thaliana* (Wang and Brendel 2006) and also over 20% corresponding to about 6,500 genes in *Oryza sativa* (rice) (Wang and Brendel 2006). On the other hand, more than 50% of genes are subject to AS in human and mouse (Johnson et al. 2003; Carninci et al. 2005). Recently, Hiller et al. (2004) reported that many human genes underwent AS events at splice acceptor sites with special consensus sequences. These sites exhibited tandem repeats of the consensus sequence “NAG” and were thus termed “NAGNAG acceptor” sites (fig. 1). In these AS events (which we will term “AS-NAGNAG events”), both the first AG (the “E” site; Hiller et al. 2004) and the second AG (the “I” site) may function as splice acceptor sites. Such events result in the variable presence or absence of the 3 nt of the latter NAG in the mature mRNAs (fig. 1). Because AS-NAGNAG events never result in frame-shifts, they affect only 1 or 2 amino acids in the translated amino acid sequences unless they generate stop or stop codons. Although the number of affected amino acids is very small, several properties of these events reveal their possible importance in the regulation of protein function. AS-NAGNAG events are reportedly enriched in genes encoding DNA-binding proteins (Akerman and Mandel-Gutfriend 2006) and tend to occur in regions enriched in polar residues (Hiller et al. 2004). Additionally, many cases of AS-NAGNAG events are conserved between human and mouse (Akerman and Mandel-Gutfriend 2006). Based on these results, AS-NAGNAG events are thought to have a regulatory role in the fine-tuning of DNA-binding proteins.

One alternative point of view, however, is that, AS-NAGNAG events may be explained by a simple physical model (Chern et al. 2006). Although Hiller et al. (2006) have argued against this theory, the biological importance of most AS-NAGNAG events remains unclear. Our study will contribute to this discussion using additional data derived from plant genomics. If DNA-binding proteins are also the principle targets of AS-NAGNAG events in plants and if these events primarily affect polar amino acids, the shared importance of AS-NAGNAG events in mammals and plants will be underscored, and these events may well be a shared strategy for regulating protein function. It should be noted that AS-NAGNAG events may be important not only because of their effects on DNA-binding proteins but also due to the influence that these effects then have on the entire transcriptome.

Recently, Campbell et al. (2006) analyzed sequence properties of the 3’ splice acceptor site of AS events and reported that these events occurred in *Arabidopsis* and rice. However, the number of studies of AS-NAGNAG events in plants remains small, and the characterization of such events remains especially poor. In this study, we established algorithms to detect AS-NAGNAG events based on full-cDNA/expressed sequence tags. We attempted to determine how frequently AS-NAGNAG events occur in *Arabidopsis* and in rice using genome-wide analyses.

Characterization of AS-NAGNAG events is also necessary to assess the potential impact of these events on the plant transcriptome. If AS-NAGNAG events found in *Arabidopsis* and rice shared a common role in evolutionary pathways, such events should be conserved. For example,
we found that AS events in genes for Ser/arg-rich splicing factors were highly conserved in Arabidopsis, rice, and probably in moss (Iida and Go 2006). In this study, we surveyed the conservation of AS-NAGNAG events in Arabidopsis and rice. We also focused our attention on a second area, that of AS-NAGNAG events found at noncorresponding introns of homologous genes. If a pair of homologous genes possesses AS-NAGNAG events at different introns, these events should be generated independently, and we term them AS-NAGNAG events at noncorresponding introns (ANNCIs). Although the amino acid position of the events is different, these events may have comparable impact on protein function when they are located in regions concerned with similar or identical molecular functions. In this paper, we provide examples of conserved AS-NAGNAG events or ANNCIs between Arabidopsis and rice as these events seem to play important role in modulating protein function. Also based on the results from these analyses, we discuss the similarities and differences of AS-NAGNAG events found in land plants and mammals.

Materials and Methods

Data Set

We used the entire genomic sequence and the annotated gene sets of Arabidopsis published on the Web site of The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org/). For rice, we used the whole-genomic sequence and the annotated gene sets published by The Institute for Genomic Research (TIGR; http://www.tigr.org/). We also used transcript sequences published by UniGene (Wheeler et al. 2007) and full-length cDNA sequences collected by RIKEN (Seki et al. 2002; Yamada et al. 2003), Ceres Inc. (Haas et al. 2002), and ‘KOME’ (Kikuchi et al. 2003). For Arabidopsis, there were 447,107 transcripts in UniGene and the RIKEN and Ceres data sets had 280,569 and 5,000 sequences, respectively, which were full or partial reads of full-length cDNAs. There were redundancies between UniGene and the RIKEN or Ceres data sets, so the total number of transcripts in Arabidopsis was 501,736. For rice, the UniGene and KOME full-length cDNAs contained 374,632 and 32,127 transcripts, respectively. Because of some redundancy, the number of unique transcripts in rice was 374,954 (fig. 2).

Construction of Transcription Units

We mapped transcript sequences to the genomic sequences to construct transcription units (TUs) (Okazaki et al. 2002; Iida et al. 2004). We used Blast (Altschul et al. 1997) for rough mapping and GeneSequer (Brendel et al. 2004) for detailed mapping. In the first Blast step, we selected transcripts whose conformity with the genomic sequence was greater than 88%. Next, we made a partial sequence of the genome corresponding to each locus of the transcript and used GeneSequer to make pairwise alignments of the genomic sequence and each transcript, taking into account exon–intron boundary rules. Finally, we limited our sample to sequences in which over 90% of the length of the transcript could be mapped to the genomic sequence. We clustered transcripts into 1 TU when they had the same direction and overlapped the same region of the genome.

Identification of AS-NAGNAG Events

We searched for AS-NAGNAG events using multiple alignments of nucleotide sequences in each TU. We created the multiple alignments from pairwise alignments of the genomic sequence and each transcript, taking into account exon–intron boundary rules. Finally, we limited our sample to sequences in which over 90% of the length of the transcript could be mapped to the genomic sequence. We clustered transcripts into 1 TU when they had the same direction and overlapped the same region of the genome.

Gene Ontology Analysis

We performed gene ontology (GO) (Ashburner et al. 2000) analysis to determine which gene types tended to
contain AS-NAGNAG events. For this analysis, we used InterProScan (Mulder et al. 2007) to assign motifs and GOs to all the annotated gene models. We evaluated the number of genes with each GO, comparing the entire gene set with the subset of genes with AS-NAGNAG events, using the fourth class of molecular function of GOs. When 1 gene had a GO whose class was under the fifth class, we converted it into a parent GO of the fourth class. We searched GOs that were enriched in genes undergoing AS-NAGNAG and checked their significance using a chi-square test. When the frequency of a GO in genes with AS-NAGNAG events was greater than that in all genes and the $P$ value computed as a result of the chi-square test was less than 0.05, we considered the GO to be statistically enriched in the AS-NAGNAG group. We used the R package (http://cran.r-project.org/) to calculate $P$ values in chi-square tests.

Analysis of Amino Acids Affected by AS-NAGNAG Events

We analyzed amino acids that were affected by AS-NAGNAG events. Although AS-NAGNAG events were defined by the mRNA sequences, we had to use the open reading frame information from each mRNA for this analysis. We then compared the constructed TUs and annotated gene models. We used AS-NAGNAG events in our analysis when it was possible to map them to annotated gene models. We prepared 6 nt on the NAGNAG site, 2 nt on the 3’ end of the previous exon, and 2 nt following the NAGNAG to determine the differences between amino acid sequences encoded by E and I transcripts (fig. 1) (Hiller et al. 2004). When 1 amino acid residue was exchanged with 2 different amino acid residues by AS-NAGNAG event, we counted all 3 amino acids for the affected amino acids. Besides, when a AS-NAGNAG event had more than 2 different previous exons caused by another AS event, we counted all possible variations. In this analysis, we listed the frequency of each amino acid in the entire protein sequence, in the exon junctions, and in sequences affected by AS-NAGNAG events (table 2).

We also performed a statistical analysis similar to that of Hiller et al. (2004). In this analysis, we listed 10 amino acids on both sides of the exon junctions with and without AS-NAGNAG events and assessed whether these flanking sequences tended to be polar or not.

Analysis of Conserved AS-NAGNAG Events, ANNCIs, and Genomic NAGNAG Sites

We searched for AS-NAGNAG events that were conserved between Arabidopsis and rice. Because it is difficult to align divergent mRNA sequences, we used amino acid sequences translated from mRNAs containing AS-NAGNAG events. We grouped homologous genes utilizing Blast (Altschul et al. 1997), using as a query amino acid sequences translated from mRNA that contained AS-NAGNAG events. For the Blast searches, the parameters were set as follows: the low-complexity filter was on, the $E$ value limit was $10^{-20}$, and the database contained all proteins in Arabidopsis and rice. We next used ClustalW (Thompson et al. 1994) to make multiple alignments of amino acid sequences that included each constructed group. When we detected a pair of AS-NAGNAG events that occurred at the same site on the alignment and were in the same phase, we considered them conserved events. In this analysis, the deviance of 1 amino acid was allowed because we detected cases in which the amino acids encoded near the exon boundaries were more divergent than those not around exon boundaries. In addition, we set another margin when the types of AS-NAGNAG events were different (i.e., the I and E transcripts). We verified conserved event candidates by visual examination of the multiple alignments (supplementary figs. 1 and 2, Supplementary Material online). Furthermore, we examined genes homologous to those with NAGNAG events to verify the corresponding introns or genomic NAGNAG sites.

For the case of AT5G16840.1, we performed homology modeling to characterize AS-NAGNAG events on the tertiary structure. For the modeling, we used SWISS-MODEL (Schwede et al. 2003) with pairwise alignment made by Blast (Altschul et al. 1997).

Results

Construction of TUs

We used 501,736 transcripts of Arabidopsis in our analysis. We mapped these transcripts to the genome using Blast (Altschul et al. 1997) and GeneSeqer (Brendel et al. 2004). As a result, we mapped 450,474 (89.8%) of the transcripts to the genomic sequence with high accuracy (see the Materials and Methods for a detailed description). We constructed TUs based on the mapped transcripts. For Arabidopsis, there were 25,380 TUs, including 21,064 TUs with more than 2 transcripts (fig. 2), and they corresponded to 24,857 loci of annotated genes in the TAIR data set. In the TAIR gene models, the total number of loci was 26,751. Thus, we were able to cover nearly all the Arabidopsis loci in our analysis.

For rice, we used 374,954 full-length cDNA sequences from UniGene and KOME. We were able to map 338,036 sequences to the rice genome and determined 26,508 TUs. In all, 19,080 TUs contained more than 2 transcripts. In all, 26,508 TUs corresponded to 27,510 loci of the annotated genes in the TIGR data set, which accounted for approximately 47% of the 57,916 total loci.

Identification of AS-NAGNAG Events

We described above a method to detect AS-NAGNAG events contained in multiple alignments of TU transcripts. By applying the method to the data sets from Arabidopsis and rice, we identified 321 AS-NAGNAG events in Arabidopsis and 372 events in rice. These events were located in 316 TUs in Arabidopsis (1.5% of all TUs contained more than 2 transcripts) and 363 TUs in rice (1.9% of all TUs contained more than 2 transcripts) (fig. 2, supplementary
Given our knowledge that certain AS-NAGNAG events had been previously annotated in gene models in TAIR or TIGR, we searched for gene models that corresponded to TUs with AS-NAGNAG events from annotated gene sets. We found corresponding gene models for 258 TUs in Arabidopsis and 248 TUs in rice, containing 261 and 253 AS-NAGNAG events, respectively. Of these, 187 events in Arabidopsis and 107 events in rice were already annotated. Therefore, we found 74 and 146 new AS-NAGNAG events in Arabidopsis and rice, respectively.

GO Analysis

To determine what types of genes tended to contain AS-NAGNAG events, we performed GO analysis. For this analysis, we used InterProScan (Mulder et al. 2007) to assign GOs to the 26,751 annotated Arabidopsis genes and 43,720 annotated rice genes. We excluded 14,196 transposable element–related genes because they account for a large fraction of the rice genes that negatively impact GO analysis. Using InterProScan, we identified 47,595 GOs (types of GOs: 1,428) in 15,482 Arabidopsis genes. In rice, we determined 54,603 GOs (types of GOs: 1,406) in 18,120 genes. We then used a data set composed of 258 genes from Arabidopsis and 248 genes from rice that corresponded to TUs with AS-NAGNAG events (see the previous sections for a detailed description). We searched this data set for GOs that were statistically enriched in AS-NAGNAG gene groups. We found 28 genes in Arabidopsis and 29 genes in rice with “GO:0003677:DNA binding” (table 1). The GOs were enriched in a group of genes with AS-NAGNAG events, and the differences were statistically significant (P values <0.05). In addition, “GO:0008026: ATP-dependent helicase activity” and “GO:0016746: transferase activity, transferring acyl groups” were also enriched in AS-NAGNAG groups in Arabidopsis. In rice, 4 other GOs, “GO:0019887:protein kinase regulator activity,” “GO:0015485:cholesterol binding,” “GO:0031072:heat shock protein binding,” and “GO:0019888:protein phosphatase regulator activity” were enriched in AS-NAGNAG groups.

Analysis of Amino Acid Residues Affected by AS-NAGNAG Events

We analyzed amino acid residues affected by AS-NAGNAG events. For this analysis, we used 506 genes (258 for Arabidopsis and 248 for rice) corresponding to TUs with AS-NAGNAG events. The results showed that in both Arabidopsis and rice, AS-NAGNAG events predominantly affected polar amino acid residues (table 2). Glutamine, serine, alanine, glutamic acid, and lysine each comprised more than 5% of the total number of amino acids in samples from both species. Excluding alanine, each of these is a polar residue, and together they accounted for more than 50% of amino acids affected by AS-NAGNAG events.

| Table 1 GO Analyses of Genes with AS at the NAGNAG Acceptor Site |
| --- | --- | --- | --- | --- | --- |
| GO | Number in All Genes | Fractions in All Genes | Number in Genes with AS-NAGNAG | Fractions in genes with AS-NAGNAG | P Value in Chi-Square Test | Description of GO |
| GO:0003677 | 59 | 0.117 | 28 | 0.023 | 0.018 | DNA binding |
| GO:0008026 | 4 | 0.008 | 4 | 0.023 | 0.024 | ATP-dependent helicase activity |
| GO:0016746 | 122 | 0.067 | 4 | 0.023 | 0.024 | Transferase activity and transferring acyl groups |

| B) Oryza sativa |
| --- | --- | --- | --- | --- | --- |
| GO | Number in All Genes | Fractions in All Genes | Number in Genes with AS-NAGNAG | Fractions in genes with AS-NAGNAG | P Value in Chi-Square Test | Description of GO |
| GO:0019887 | 4 | 0.002 | 2 | 0.013 | 1.25 × 10^-10 | Protein kinase regulator activity |
| GO:0015485 | 1 | 0.001 | 1 | 0.006 | 1.95 × 10^-04 | Cholesterol binding |
| GO:0031072 | 126 | 0.068 | 4 | 0.023 | 0.023 | Heat shock protein binding |
| GO:0019888 | 5 | 0.003 | 1 | 0.006 | 0.045 | Protein phosphatase regulator activity |

---

*These GOs had “warning” in chi-square test, which indicates possibility of incorrect approximating of chi-square values.*
We also compared the number of polar amino acids contained within the 10 flanking amino acid residues from exon junctions with and without AS-NAGNAG events. Flanking residues around exon junctions with AS-NAGNAG events were more polar than those without AS-NAGNAG events (P < 0.00001 in t-test), both in Arabidopsis and rice.

Analysis of Conserved AS-NAGNAG Events, ANNCIs, and Genomic NAGNAG Sites

We searched for AS-NAGNAG events that were conserved between Arabidopsis and rice. In this analysis, we used 506 genes (258 for Arabidopsis and 248 for rice) corresponding to TUs with AS-NAGNAG events. We found 7 homologous gene groups with conserved AS-NAGNAG events: small nuclear ribonucleoprotein D2 (Sm-D2), urease accessory protein D, Similar to surfeit locus protein 2 (SURF2) family protein, RNA recognition motif (RRM)-containing protein, putative O-acetyltransferase, auxin-induced gene IAA13, and an unknown protein (fig. 2 and table 3). For 4 groups (Sm-D2, urease accessory protein D, RRM-containing protein, and IAA13), the influence on the encoded amino acids was also conserved (table 3). For example, we could assign the tertiary structure of the RRM domain of serine/

Table 2
Frequency Analysis of Amino Acids Affected AS-NAGNAG Events

| Amino Acid | Raw Count | Frequency | Arabidopsis thaliana | Oryza sativa |
|------------|-----------|-----------|----------------------|-------------|
|            | All Proteins | Exon Junctions | NAGNAG | All Proteins | Exon Junctions | NAGNAG |
| D          | 687,322 | 5.43 | 9,055 | 6.73 | 0 | 0.00 |
| E          | 854,021 | 6.75 | 10,025 | 7.45 | 31 | 11.36 |
| K          | 803,868 | 6.35 | 4,854 | 3.61 | 33 | 12.09 |
| R          | 682,642 | 5.40 | 12,076 | 8.98 | 15 | 5.49 |
| H          | 286,937 | 2.27 | 1,100 | 0.82 | 0 | 0.00 |
| N          | 554,845 | 4.39 | 3,081 | 2.29 | 1 | 0.37 |
| Q          | 442,548 | 3.50 | 1,416 | 1.05 | 84 | 30.77 |
| S          | 1,147,656 | 9.07 | 8,506 | 6.32 | 35 | 12.82 |
| T          | 804,710 | 6.36 | 10,634 | 7.90 | 35 | 12.82 |
| V          | 847,389 | 6.70 | 21,496 | 15.98 | 20 | 7.33 |
| L          | 1,199,656 | 9.48 | 6,888 | 4.90 | 31 | 1.37 |
| I          | 671,735 | 5.31 | 7,703 | 5.73 | 1 | 0.37 |
| P          | 612,022 | 4.84 | 2,178 | 1.62 | 1 | 0.37 |
| F          | 539,483 | 4.26 | 2,303 | 1.71 | 3 | 1.10 |
| M          | 516,051 | 4.08 | 2,329 | 1.74 | 0 | 0.00 |
| W          | 158,424 | 1.25 | 2,747 | 2.04 | 1 | 0.37 |
| C          | 231,416 | 1.83 | 1,705 | 1.27 | 1 | 0.37 |
| 12,653,152 | 100 | 134,548 | 100 | 273 | 100 |

* Data for amino acids affected by AS events at NAGNAG site.

506 genes (258 for Arabidopsis and 248 for rice) corresponding to TUs with AS-NAGNAG events. We found 7 homologous gene groups with conserved AS-NAGNAG events: small nuclear ribonucleoprotein D2 (Sm-D2), urease accessory protein D, Similar to surfeit locus protein 2 (SURF2) family protein, RNA recognition motif (RRM)-containing protein, putative O-acetyltransferase, auxin-induced gene IAA13, and an unknown protein (fig. 2 and table 3). For 4 groups (Sm-D2, urease accessory protein D, RRM-containing protein, and IAA13), the influence on the encoded amino acids was also conserved (table 3). For example, we could assign the tertiary structure of the RRM domain of serine/

Table 3
AS Events at NAGNAG Acceptors Conserved between Arabidopsis and Rice

| Name of Group | Included Genes | Descriptions of Genes | Number of Intron Where AS-NAGNAG Found | Phase of the Intron | NAGNAG Sequence | Amino Acid Encoded in E Transcript | Amino Acid Encoded in I Transcript |
|---------------|----------------|-----------------------|----------------------------------------|---------------------|----------------|-------------------------------|-------------------------------|
| CSV_1         | AT2G47640.1     | Sm-D2, putative       | 2                                      | 0                   | UAGCAG         | Q                            | —                            |
|               | AT3G62840.1     |                       | 2                                      | 0                   | UAGCAG         | Q                            | —                            |
|               | Os05g24970.1    |                       | 1                                      | 0                   | UAGCAG         | Q                            | —                            |
| CSV_2         | AT3G14440.1     | SURF2 family protein  | 4                                      | 2                   | UAGCAG         | SS                           | —                            |
|               | Os04g40430.1    |                       | 4                                      | 2                   | UAGCAG         | SS                           | —                            |
| CSV_3         | AT5G06840.1     | RRM-containing protein| 2                                      | 2                   | UAGCAG         | SR                           | R                            |
|               | Os04g50790.1    |                       | 2                                      | 2                   | UAGCAG         | SS                           | —                            |
| CSV_4         | AT5G09710.1     | Unknown protein       | 1                                      | 0                   | UAGAAG         | ED                           | D                            |
|               | Os02g54080.1    |                       | 2                                      | 1                   | UAGAAG         | ED                           | D                            |
| CSV_5         | AT3G03053.1     | Urease accessory protein (URED) | 4                                      | 0                   | CAGUAG         | *                            | —                            |
|               | Os02g76760.1    |                       | 5                                      | 0                   | CAGUAG         | *                            | —                            |
| CSV_6         | AT3G06650.1     | O-acetyltransferase, putative | 2                                      | 2                   | UAGUAG         | TR                           | T                            |
|               | Os03g19970.1    |                       | 2                                      | 2                   | UAGUAG         | AR                           | A                            |
| CSV_7         | AT2G33310.1     | Auxin-induced gene, IAA13 | 1                                      | 2                   | UAGCAG         | SS                           | S                            |
|               | Os02g57250.1    |                       | 1                                      | 2                   | UAGCAG         | SS                           | S                            |

* We counted number of introns only for those in coding sequences.
arginine (SR)-rich factor 9G8 (Hargous et al. 2006) to the RRM domain encoded by AT5G16480. The sequence identity between these RRM domains was approximately 35%.

We next analyzed the impact of AS-NAGNAG events on the tertiary structure by homology modeling (fig. 3). The sequence containing the AS-NAGNAG events was positioned on the loop structure situated between 2 beta strands. These beta strands are known to be responsible for RNA binding in the SR-rich factor 9G8 (Hargous et al. 2006). The RRM domain encoded by the E transcript had a longer loop than that encoded by the I transcript. Therefore, the position of Glu42, which might affect RNA-binding properties, was different between E and I transcripts. A second example is the gene Sm-D2 encoded by AT2G47640 and AT3G62840, to which we could assign the tertiary structure of human Sm-D2 (Kambach et al. 1999). Although the position of the AS-NAGNAG event was disordered in the tertiary structure of human Sm-D2, the AS-NAGNAG event did lie on the N-terminal region that may serve as an interaction face for a heptameric ring or for mRNA strands. For urease accessory protein D, AS-NAGNAG events that led to creation of a premature termination codon (PTC) were conserved; we will return to this topic in the Discussion.

We searched for genes homologous to those containing AS-NAGNAG events and analyzed whether or not they had corresponding introns at the same locations as the AS-NAGNAG events. When they did, we also checked for the existence of genomic NAGNAG sites in the homologous genes. Out of all 506 genes with AS-NAGNAG events, 442 genes had at least 1 homologous gene in the alternate species (rice homologues for *Arabidopsis* genes and *Arabidopsis* homologues for rice genes). For 286 of the AS-NAGNAG events, at least 1 homologue contained corresponding introns. We found that 88 events (46 in *Arabidopsis* and 42 in rice) had homologues with genomic NAGNAG sites at corresponding sites (supplementary table 2, Supplementary Material online). Because they included 15 events consisting 7 groups conserved AS-NAGNAG

\[ \text{Fig. 3.} - \text{Tertiary structure model of RRM-containing protein, AT5G16840.1. (A) and (B) are models for protein-encoded I transcript, and (C) and (D) are models for protein-encoded E transcript. (A) and (C) show the secondary structures. We showed serine residues affected by AS-NAGNAG events and charged residues which positions were also affected by the events. (B) and (D) show the surface of the proteins. Positively charged residues (Arg and Lys) are showed with white, and negatively charged ones (Asp and Glu) showed with black. The front side of these figures is used for RNA interaction surface in the original proteins. These models were built by homology modeling based on the RRM domain of human SR-rich factor 9G8 (PDB ID 2HVZ) (Hargous et al. 2006). (A) and (C) were drawn using MOLSCRIPT (Kraulis 1991) and Raster3D (Merritt and Murphy 1994). (B) and (D) were drawn using UCSF Chimera (Pettersen et al. 2004).} \]
events, 73 AS-NAGNAG events (38 in *Arabidopsis* and 35 in rice) had homologues with genomic NAGNAG sites but no observed AS-NAGNAG event. We also found 23 cases where several homologous genes had AS-NAGNAG events in different introns (supplementary table 3 and fig. 2, Supplementary Material online). Figure 4 is one example, NAC (petunia NAM and *Arabidopsis* ATAF1,2 and CUC2) domain containing transcription factors. In this case, *Arabidopsis* gene had an AS-NAGNAG event on the DNA-binding region and rice had one on the activator domain.

**Discussion**

This study is the first genome-wide study analyzing AS-NAGNAG events in *A. thaliana* and *O. sativa*. One of our goals was to characterize AS-NAGNAG events in plants and then compare our findings with existing data derived from studies on mammals. In the present study, we found about 2% of all TUs having AS-NAGNAG events in *Arabidopsis* and rice, respectively. On the other hand, Hiller et al. (2004) reported that at least 5% of all human genes contain AS-NAGNAG events. Using these numbers only, it would appear that the prevalence of AS-NAGNAG events in plants is smaller than that in human. When we consider all AS events and not solely those associated with NAGNAG, the fraction of genes with AS is reportedly ~20% in *Arabidopsis* and rice (Wang and Brendel 2006). In comparison, the prevalence in mammals is reportedly at least 50% (Johnson et al. 2003; Carninci et al. 2005). The difference between the rate of AS-NAGNAG events in plants and human may in fact reflect a different background frequency of more general AS events between the 2 species.

We found that certain features of AS-NAGNAG events in plants highly resemble those in mammals. For example, AS-NAGNAG events in plants are enriched in genes encoding DNA-binding proteins and mainly affect polar amino acids such as lysine, glutamine, glutamic acid, and serine. These findings are highly consistent with those found in mammals (Hiller et al. 2004; Akerman and Mandel-Gutfreund 2006). This similarity suggests that AS-NAGNAG events have a common role in mammals and plants. Even though the absolute number of affected amino acids is small, polar amino acids can play an important regulatory role in DNA-binding proteins. Hiller et al. (2006) used the mouse Pax3 gene as an example. In this case, the presence or absence of glutamine caused by AS-NAGNAG events can change DNA-binding properties. Human EGR-1 protein, a C2H2-type zinc finger protein, follows a similar pattern. In this case, 3 amino acids, lysine–threonine–serine, were changed by AS, and these events can change the DNA-binding properties of the protein (Larsson et al. 1995; Stetefeld and Ruegg 2005).

**Fig. 4**.—(A) A part of pairwise alignment of AT3G10480.1 and Os08g06140.1. Both of them are members of NAC family transcription factors. In this pair, both of genes had AS-NAGNAG events but the site quite different. These events were thought to be generated independently, so they are termed ANNCIs. (B) A scheme for domain structure of NAC family transcription factors. AT3G10480.1 from *Arabidopsis* has AS-NAGNAG event on DNA-binding domain that account for C terminal part of NAC domain. Os08g06140.1 from rice has the event on activation domain. This scheme is based on a tertiary structure of abscisic-acid-responsive NAC (PDB ID 1UT4) (Ernst et al. 2004) and a figure drawn by Ooka et al. (2003).
We hypothesize that AS-NAGNAG events found in plant DNA-binding proteins have at least a minor regulatory role of a similar kind. We also suggest that both mammals and plants use the same strategy for regulating protein functions through AS-NAGNAG events.

Our study also compared the presence of AS-NAGNAG events in Arabidopsis and rice and found 7 homologous gene groups where these events were conserved. The number of the conserved AS-NAGNAG events looked large, when we compared that with known conserved AS events at splice acceptor site, which was only 5 and did not include current AS-NAGNAG events (Wang and Brendel 2006). However, when we compared the number of conserved AS-NAGNAG events in land plants and mammals, there appeared to be a significant difference between the 2. Akerman and Mandel-Gutfreund (2006) reported high conservation of AS-NAGNAG events between human and mouse, based on findings from 215 conserved events between human and mouse. The possibility certainly exists that a larger number of cases show the conservation of the AS-NAGNAG events. Thirty-eight AS-NAGNAG events in Arabidopsis and 35 events in rice had corresponding introns in homologous genes with genomic NAGNAG sites but without confirmed AS events. Further accumulation of transcripts might provide more confirmation of AS-NAGNAG events. However, even in the case of human, for whom more transcripts have been accumulated than for Arabidopsis and rice, only 13% of all genomic NAGNAG sites were associated with transcript-confirmed AS events (Hiller et al. 2004). We believe that a certain number of genomic NAGNAG sites cannot provide AS variants or can provide AS variants at very low rates. We also cannot expect that all AS-NAGNAG events with homologous genes having genomic NAGNAG sites will become conserved cases. Although the number of conserved events in plants is very small, the result shows interesting consistency with mammals. Hiller et al. (2004) reported that several splicing-related genes, such as PRPF3, PRPF8, U2AF1, and U2AF2, had AS-NAGNAG events conserved between human and mouse. In our results, Sm-D2, a component of snRNP, and one protein containing an RRM that can act in RNA metabolism were included in the conserved cases. Akerman and Mandel-Gutfreund (2006) reported that RNA-binding proteins were also common sites for AS-NAGNAG events. Although we did not find the relationship between AS-NAGNAG events and RNA-binding proteins to be statistically significant, AS-NAGNAG events may well have an important role in regulation of these RNA-binding proteins in plants because of the conservation of these events between Arabidopsis and rice. As described in the Results, these AS-NAGNAG events lie on protein–RNA or protein–protein interaction surfaces, suggesting their ability to modify protein function.

The presence of AS-NAGNAG events in RRM-containing proteins demonstrates the potential importance of AS-NAGNAG events on changing the positions of functional residues in tertiary structures (fig. 3). In such cases, AS-NAGNAG events occurring in positions neighboring the functional sites can modify the functions of the proteins, even if the sequences of the functional residues themselves are not changed by AS-NAGNAG events. Conservation of AS events in RNA-binding proteins resembles our previous results on SR proteins (Iida and Go 2006). These types of conserved AS events found on RNA-binding proteins should have important role in both Arabidopsis and rice, along with the presence of AS-NAGNAG events.

We found the cooccurrence of AS-NAGNAG events and genes with specific GO term DNA binding to be statistically significant, however, no genes for DNA-binding proteins reveal conservation of the events between Arabidopsis and rice. On the other hand, we found that DNA-binding proteins were enriched in genes with ANNCIs. We are unable to suggest a mechanism by which an AS-NAGNAG event moves from its original site to another intron. Such events should be generated independently. Out of 23 groups with ANNCIs, 5 groups could be considered DNA-binding proteins (supplementary table 3, Supplementary Material online). In fact, many DNA-binding proteins contain AS-NAGNAG events. However, the expected number of ANNCIs, assuming no bias in the distribution of AS-NAGNAG events, is 0.9 (if 28 and 29 genes encoding DNA binding in Arabidopsis and rice, respectively, are clustered into the homologous gene group). This difference is statistically significant (P value < 0.01, with binomial test). We then considered that some bias might impact the AS-NAGNAG events fixed in mRNAs encoding DNA-binding proteins in evolutionary path. It is likely that certain AS-NAGNAG events are useful in fine-tuning the function of DNA-binding proteins. One example is the NAC family of transcription factors (fig. 4); the Arabidopsis gene has an AS-NAGNAG event on the DNA-binding domain and rice has an event on the activating domain (Ooka et al. 2003). Even if the location was different, both events could modulate the function of the transcription factors. In this case, the Arabidopsis AS-NAGNAG event is situated on the loop between 2 beta strands, whose high mobility precluded precise observation of their structure (Ernst et al. 2004). Such flexibility, however, likely facilitates the occurrence of AS-NAGNAG events within the protein.

In this study, we found several AS-NAGNAG events that could generate premature stop codons (PTC). The most interesting example was found in the mRNA-encoding urease accessory protein D because this AS-NAGNAG event was conserved between Arabidopsis and rice (table 3). A second example occurred in the mRNA for geranyl diphosphate synthase, which we listed as containing conserved AS-NAGNAG events (ANNCI 22 in supplementary table 3, Supplementary material online). In this case, 2 independent AS-NAGNAG events resulted in PTCs. Conserved AS-NAGNAG events and ANNCIs were expected to have potentially important roles, one of which is likely to be creation of PTCs. This indicates that AS-NAGNAG may use a similar mechanism to “regulated unproductive splicing and translation” (RUST; Lewis et al. 2003). We previously found that AS events occurring in SR proteins in land plants can regulate mRNAs with this type of mechanism (Iida and Go 2006). Wang and Brendel (2006) also reported the positive importance of RUST in land plants. RUST seems to be a widespread regulatory mechanism in land plants, and AS-NAGNAG is likely the most economic way to generate PTCs.
This is the first genome-wide study of AS-NAGNAG events in Arabidopsis and rice. We have shown that in important ways, AS-NAGNAG events are very similar to those in mammals: these events are enriched in genes encoding DNA-binding proteins, and they tend to affect the existence of polar amino acid residues. We also found that several potentially important AS-NAGNAG events were conserved in genes for RNA-binding proteins, a situation that holds true for both mammals and plants. Based on these similarities, we believe AS-NAGNAG events have a definite biological role and are not merely accidental events. In contrast, Chern et al. (2006) suggested the opposite that AS-NAGNAG events occur by chance. We can agree with this hypothesis, but only insofar as it pertains to the initial generation of AS-NAGNAG events. Although many DNA-binding proteins contain AS-NAGNAG events, none were conserved between Arabidopsis and rice. This suggests that AS-NAGNAG events may not only arise frequently but also disappear frequently, which would follow the reasoning of Chern’s hypothesis. Besides, our result for usages of amino acid residues affected by AS-NAGNAG events also supported this hypothesis. Randomly generated AS-NAGNAG data set that satisfied observed phase usage of the events, showed similar properties with observed results (data not shown). On the other hand, we found several cases of ANNCIs occurring in DNA-binding proteins. Based on these findings, we suggest that there is a bias with which AS-NAGNAG events are fixed in DNA-binding proteins. This potential bias can result in a high frequency of DNA-binding proteins in genes with AS-NAGNAG events and can also lead to ANNCIs consisting of several genes for DNA-binding proteins. This potential bias also suggests the importance of AS-NAGNAG events with regards to genes coding for DNA-binding proteins. The above facts lead us to believe that AS-NAGNAG events are biologically consequential, even if their original appearance is due solely to chance.

Although AS-NAGNAG events affect only a very small fraction of proteins, we have shown in this study that they commonly regulate protein functions in both mammals and plants and as such are likely to play an important biological role. Mammals and plants share a basic intron splicing process. However, it is known that initial recognition mechanism for exons or introns is different between them (Lorkovic et al. 2000), and it is therefore unclear whether or not the common properties we found between mammals and plants in this study have the same origin. Despite this uncertainty, the common properties themselves show that AS-NAGNAG events may contribute to gene regulation and to the diversity of transcriptomes and proteomes in the evolutionary history of the eukaryote.

**Supplementary Material**

Supplementary tables 1–3 and figures 1 and 2 are available at [Molecular Biology and Evolution online](http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We are grateful to Dr Mitiko Go at Ochanomizu University for her invaluable advice in this work. We also appreciate Mr Hideaki Haruna for his contributions on predicting DNA-binding proteins and transcription factors. This work was supported by a grant of the Genome Network Project from Ministry of Education, Culture, Sports, Science, and Technology of Japan. Also this work was supported in part by a grant of the Human Frontier Science Program to K.I.

**Literature Cited**

Akerman M, Mandel-Gutfreund Y. 2006. Alternative splicing regulation at tandem 3‘ splice sites. Nucleic Acids Res. 34: 23–31.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Ashburner M, Ball CA, Blake JA, et al. 20 co-authors. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Brendel V, Xing L, Zhu W. 2004. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics. 20:1157–1169.

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.

Carninci P, Kasukawa T, Katayama S, et al. (194 co-authors). 2005. The transcriptional landscape of the mammalian genome. Science. 309:1559–1563.

Chern TM, van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, Zavolan M. 2006. A simple physical model predicts small exon length variations. PLoS Genet. 2:e45.

Ernst HA, Olsen AN, Larsen S, Lo Leggio L. 2004. Structure of the conserved domain of ANAC, a member of the NAC family of transcription factors. EMBO Rep. 5:297–303.

Haas BJ, Volfovsky N, Town CD, Troukhin M, Alexandrow N, Feldmann KA, Flavell RB, White O, Salzberg SL. 2002. Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol. 3:research0029.1–research0029.12.

Hargous Y, Hautbergue GM, Tintaru AM, Skrivoska L, Golovanov AP, Stevenin J, Lian LY, Wilson SA, Allain FH. 2006. Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8. EMBO J. 25:5126–5137.

Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M. 2004. Widespread occurrence of alternative splicing at NAGNAG acceptors: simply noise or noise and more? PLoS Genet. 2:1944–1946.

Iida K, Go M. 2006. Survey of conserved alternative splicing events of mRNAs encoding SR proteins in land plants. Mol Biol Evol. 23:1085–1094.

Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K. 2004. Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. Nucleic Acids Res. 32:5096–5103.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction micro-arrays. Science. 302:2141–2144.
Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, Luhrmann R, Li J, Nagai K. 1999. Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. Cell. 96: 375–387.

Kikuchi S, Satoh K, Nagata T, et al. (74 co-authors). 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science. 303:376–379.

Kraulis PJ. 1991. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr. 24:946–950.

Larsson SH, Charlier JP, Miyagawa K, Engelkamp D, Rassoulzadegan M, Ross A, Cuzin F, van Heyningen V, Hastie ND. 1995. Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing. Cell. 81:391–401.

Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci USA. 100:189–192.

Lorkovic ZJ, Wieczorek Kirk DA, Lambermon MH, Filipowicz W. 2000. Pre-mRNA splicing in higher plants. Trends Plant Sci. 5:160–167.

Merritt EA, Murphy ME. 1994. Raster3D Version 2.0. A program for photorealistic molecular graphics. Acta Crystallogr D Biol Crystallogr. 50:869–873.

Mulder NJ, Apweiler R, Attwood TK, et al. (45 co-authors). 2007. New developments in the InterPro database. Nucleic Acids Res. 35:D224–D228.

Okazaki Y, Furuno M, Kasukawa T, et al. (137 co-authors). 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 420:563–573.

Ooka H, Satoh K, Doi K, et al. (16 co-authors). 2003. Comprehensive analysis of NAC family genes in Oryza sativa and Arabidopsis thaliana. DNA Res. 10: 239–247.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem. 25:1605–1612.

Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 31:3381–3385.

Seki M, Narusaka M, Kamiya A, et al. (20 co-authors). 2002. Functional annotation of a full-length Arabidopsis cDNA collection. Science. 296:141–145.

Stetefeld J, Ruegg MA. 2005. Structural and functional diversity generated by alternative mRNA splicing. Trends Biochem Sci. 30:515–521.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Wang BB, Brendel V. 2006. Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci USA. 103:7175–7180.

Wheeler DL, Barrett T, Benson DA, et al. (30 co-authors). 2007. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 35:D5–D12.

Yamada K, Lim J, Dale JM, et al. (70 co-authors). 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. Science. 302:842–846.

Takashi Gojobori, Associate Editor

Accepted January 9, 2008