Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes

HITOMI ITOH,^{1,2} TAKANORI WASHIO,^{1,3} and MASARU TOMITA^{1,2}

¹Laboratory for Bioinformatics, Institute for Advanced Biosciences, Keio University, Fujisawa, Kanagawa 252-8520, Japan

²Department of Environmental Information, Keio University, Fujisawa, Kanagawa 252-8520, Japan

ABSTRACT

We previously reported a computational approach to infer alternative splicing patterns from *Mus musculus* full-length cDNA clones and microarray data. Although we predicted a large number of unreported splice variants, the general mechanisms regulating alternative splicing were yet unknown. In the present study, we compared alternative exons and constitutive exons in terms of splice-site strength and frequency of potential regulatory sequences. These regulatory features were further compared among five different species: *Homo sapiens*, *M. musculus*, *Arabidopsis thaliana*, *Oryza sativa*, and *Drosophila melanogaster*. Solid statistical validations of our comparative analyses indicated that alternative exons have (1) weaker splice sites and (2) more potential regulatory sequences than constitutive exons. Based on our observations, we propose a combinatorial model of alternative splicing mechanisms, which suggests that alternative exons contain weak splice sites regulated alternatively by potential regulatory sequences on the exons.

Keywords: alternative splicing; exonic splicing enhancers; ESEs; splice-site strength

INTRODUCTION

One of the most striking discoveries of the genomic era is the unexpectedly small number of genes in the human genome, now predicted to contain roughly 35,000 genes, tens of thousands less than initially expected (Claverie 2001). Alternative splicing plays an extremely important role in expanding protein diversity and might therefore partially explain the apparent discrepancy between gene number and organism complexity (Szathmary et al. 2001). Approximately 40%–60% of human genes are estimated to have distinct splice variants (Modrek and Lee 2002). The regulation of alternative splicing can involve on/off regulation of the products of particular genes and the production of alternative products with clearly separable functions, often in a cell-type-specific manner. Despite the importance of alternative splicing, the general mechanisms that regulate those alternative splicing are unknown.

Numerous studies have been reported on the conse-

quences of single nucleotide mutations in donor and acceptor site motifs in pre-mRNA splicing. In the human *ATP7A* gene, for example, mutation in the invariant donor splice site of a constitutive exon causes complete skipping of the exon, which leads to severe Menke disease (Moller et al. 2000). Similarly, information-theory-based analysis of splice sites in the human *XPC* gene has indicated that a strong acceptor site hosts few alternative splicing events whereas a weak acceptor site has frequent exon skipping (Khan et al. 2002). Another recently reported study, however, has indicated that donor sites and acceptor sites themselves contain insufficient information for accurate splicing in higher eukaryotes (Lim and Burge 2001). Therefore, in addition to splice-site strength, it is necessary to consider additional auxiliary factors for an understanding of alternative splicing regulation.

SR proteins and exonic splicing enhancers (ESEs) play an important role in the regulation of alternative splicing. An SR protein binds to an ESE through its RNA-recognition motifs (RRM) and contacts the components of a spliceosome through its RS domain. In the human *SMN2* gene, for example, point mutations to ESE motifs on exon 7 can disrupt an ASF/SF2 (alternative splicing factor/splicing factor 2)-dependent regulation of alternative splicing and lead to inefficient inclusion of exon 7 (Cartegni and Krainer 2002). Although many SR-protein-binding site motifs have been identified by selective evolution of ligands by expo-

Reprint requests to: Takanori Washio, Laboratory for Bioinformatics, Institute for Advanced Biosciences, Keio University, Fujisawa, Kanagawa 252-8520, Japan; e-mail: washio@sfc.keio.ac.jp; fax: +81 (466) 47-5099.

³**Present address:** Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, 630-0192, Japan.

Article and publication are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.5221604>.

nential enrichment (SELEX), the distribution of those regulatory sequences in alternative and constitutive exons has remained uncertain.

Although both splice-site strength and exonic regulatory sequences have been reported to be involved in regulation of alternative splicing, most of these studies have been performed on individual genes. Thus, it has been impossible to determine whether the reported feature is (1) specific to the reported gene, (2) specific to the reported species, or (3) universal to all species. On the other hand, large-scale bioinformatics studies have contributed greatly to a general understanding of alternative splicing. Computational and statistical analyses of human genes have identified 10 ESE motifs that all showed enhancer activity in humans (Fairbrother et al. 2002). Another computational analysis of mouse transcriptome has identified sequence motifs enriched around the donor and acceptor sites in both constitutive and alternative exons (Zavolan et al. 2002). Thus, large-scale bioinformatics analyses of more than 1000 genes can present a broad view of alternative splicing regulatory mechanisms.

In the present study, we have compared alternative and constitutive exons in terms of splice-site strength and frequency of potential regulatory sequences. We have applied an information-theory-based approach for computing splice-site strength, and a phylogenetic footprinting approach for predicting potential regulatory sequences. In addition, we have compared these features among mammals (*Homo sapiens* and *Mus musculus*), plants (*Arabidopsis thaliana* and *Oryza sativa*), and *Drosophila melanogaster*. Based on the results from this comparative analysis of alternative and constitutive exons, we have delineated a potential model for alternative splicing regulatory mechanisms.

RESULTS

Detection of putative splice variants

We detected putative splice variants by mapping full-length cDNA sequences to genomic sequences by the use of BLAST

(Altschul et al. 1990), followed by SIM4 (Florea et al. 1998; Table 1). *H. sapiens* had the largest number of introns per cDNA (6.38 introns/cDNA) whereas *D. melanogaster* had only 3.54 introns/cDNA. Numbers of putative splice variants detected from *A. thaliana* and *D. melanogaster* were comparatively small because redundant sequences were excluded from the data sets in the process of constructing the full-length cDNA databases for public use (Seki et al. 2002; Stapleton et al. 2002). Because each of the five species had a sufficient number of data sets for comparative analyses, *A. thaliana* and *D. melanogaster* were included in further comparative analyses as well.

As illustrated in Figure 1, variation in gene structure due to alternative splicing takes many different forms (McKeown 1992). The putative alternatively spliced transcripts were classified according to the various alternative splicing patterns (Table 2). Also, exons that were conserved across the entire transcript in splice variant clusters were detected as constitutive exons. Cassette exons were detected two- to fivefold more frequently than other patterns, including internal donor site, internal acceptor site, and retained intron. It was difficult to identify where potential regulatory sequences might exist in internal exons with different donor/acceptor sites. Because they therefore could not be compared with constitutive exons, we focused only on cassette exons for further analyses.

Comparison of splice-site strength in constitutive exons and alternative exons

First, we computed information contents on splice sites in all internal exons (constitutive and alternative exons) for all five species by use of methods from information theory (Schneider 1997). Information content is defined as the number of choices needed to describe a sequence pattern, using a logarithmic scale in bits and can be computed from probability of each nucleotide at each position. The sequence of the splice site dictates interaction strength between spliceosome and splice site, and thus splice site

TABLE 1. Mapping and clustering results

Species	Successfully mapped	Spliced cDNA	Intron	Intron/cDNA	Clusters	Splice variants
<i>H. sapiens</i>	15,506	9,146	58,349	6.38	2,818	985
<i>M. musculus</i>	49,162	19,200	87,183	4.54	8,879	1,424
<i>A. thaliana</i>	13,071	11,172	57,002	5.10	2,529	162
<i>O. sativa</i>	27,176	20,213	100,234	4.96	5,914	975
<i>D. melanogaster</i>	9,583	8,125	28,795	3.54	442	118

Full-length cDNA and genomic DNA were used to construct data sets of putative splice variants. "Successfully mapped" column represents number of full-length cDNA that were successfully mapped to genomic DNA through the use of BLAST (Altschul et al. 1990) and SIM4 (Florea et al. 1998). "Splice cDNA" column represents number of full-length cDNA that contained one or more introns. "Intron" column represents number of introns that were detected from inferred gene structure of full-length cDNA, and "Intron/cDNA" column is number of introns divided by total number of spliced cDNA. Full-length cDNA sequences were grouped into clusters, as summarized in "Clusters" column. The cluster was thought to be a "splice variant," if introns and exons had different boundaries in different transcripts.

FIGURE 1. Patterns of alternative splicing variation. Putative splice variants were classified according to the basic patterns of alternative splicing (McKeown 1992). (A) Cassette exon: an exon is spliced out with neighboring introns or included in the mature mRNA. (B) Internal donor site: competing donor (5') splice sites exist for one acceptor site within an intron. (C) Internal acceptor site: competing acceptor (3') splice sites exist for one donor site within an intron. (D) Retained intron: An intron is spliced out or included in RNA.

use; information content has been suggested to be related to this interaction strength (Stephens and Schneider 1992).

Information contents of positions -13 to -8 upstream of the acceptor site were higher in mammals than in plants and *D. melanogaster* (Supplemental Fig. 1, available online at <http://www.bioinfo.sfc.keio.ac.jp/research/intron/>). Information contents between positions $+3$ and $+6$ downstream from the donor site were higher in mammals and *D. melanogaster* than in plants. The chi-squared ($p < 0.05$, $df = 8$) test indicated that plants had significantly higher information contents between positions $+7$ and $+15$ downstream from the donor site than other species.

To compare splice-site strength in constitutive and alternative exons, we applied the individual information (R_i) technique. Individual information contents (R_i , bits) can be computed by adding up information content of given nucleotides from individual positions, using the weight matrix generated from the frequency of each nucleotide at each position. We computed the individual information content of each individual splice site by summing information contents of positions -3 to $+7$ for donor sites and positions -13 to $+1$ for acceptor sites. All of the weight matrices and individual information contents of all splice sites are available as supplementary material online at <http://www.bioinfo.sfc.keio.ac.jp/research/intron/>.

Based on the individual in-

formation content of all splice sites, we then computed the average individual information contents (R_i) of splice sites in constitutive and alternative exons (Fig. 2; Table 3). Student's unpaired t test ($p < 0.05$, two-sided) indicated that average individual information contents of alternative splice sites were significantly lower than those of constitutive splice sites in *H. sapiens*, *M. musculus*, *O. sativa*, and *D. melanogaster*. Although we observed comparatively low information contents of *A. thaliana* alternative exons, the number of alternative exons (25 exons) was insufficient to draw a statistically significant conclusion.

The average individual information contents of splice sites were further compared among five species. For the donor site, *D. melanogaster* had the highest average individual information

content, with mammals next, followed by plants. For acceptor sites, mammals again had higher average individual information contents than plants, but *D. melanogaster* had the lowest average individual information content. Student's unpaired t test ($p < 0.05$, two-sided) indicated that *O. sativa* had significantly lower average individual information contents than *M. musculus* for both donor and acceptor sites.

These differences in average individual information contents were further assessed with a method developed by Shapiro and Senapathy; they have developed a method to score splice-site strength, based on percentage of each nucleotide at each position (Shapiro and Senapathy 1987). We applied Shapiro's method to positions -3 to $+7$ for donor sites and positions -13 to $+1$ for acceptor sites, computed Shapiro's score of individual splice sites, and compared the average score of splice sites in constitutive and alternative exons (Table 3). Differences between these two

TABLE 2. Classification of splice variants and detection of constitutive exon

Species	Cassette exon	Internal donor site	Internal acceptor site	Retained intron	Constitutive exon
<i>H. sapiens</i>	1,757	281	243	264	4,316
<i>M. musculus</i>	1,054	459	452	346	6,659
<i>A. thaliana</i>	25	51	88	45	7,138
<i>O. sativa</i>	558	258	256	481	1,251
<i>D. melanogaster</i>	145	35	28	22	286

Putative splice variants were classified according to the basic patterns of alternative splicing (see Figure 1). We focused only on cassette exons, defined as "an exon that is spliced out with neighboring introns or included in the mature mRNA," for our analysis. An exon was detected as "constitutive exon," if the exon is an "internal exon" (neither first exon nor last exon) and conserved across the entire transcript in a splice variant cluster.

FIGURE 2. Average individual information content of constitutive splice sites and alternative splice sites. Bars represent average individual information contents (R_i) of constitutive (black bars) and alternative (light bars) exons. Error bar represents standard error of the averages. Differences between these two averages were assessed with Student's unpaired t test ($p < 0.05$, two-sided). These differences were further assessed with Shapiro's method (Shapiro and Senapathy 1987), as summarized in Table 3. Statistical verification of our results supported our two main observations that alternative exons have significantly weaker splice sites than constitutive exons and that plants have significantly weaker splice sites than mammals.

averages were assessed with Student's unpaired t test ($p < 0.05$, two-sided). Because we obtained exactly the same results as with the individual information technique, we further supported our two main observations that alternative exons have significantly weaker splice sites than constitutive exons and that plants have significantly weaker splice sites than mammals.

Distribution of nucleotide substitution rates in the 5' end/3' end of evolutionarily conserved exons (constitutive + alternative exons)

We applied a phylogenetic footprinting approach for analyzing potential regulatory sequences in exons. Phylogenetic

footprinting is one of the well-known approaches to predicting regulatory sequence, by which unusually well-conserved sequences among a set of orthologous genes are extracted as candidates for functional regulatory elements (Blanchette and Tompa 2002). To collect orthologous genes, we mapped *M. musculus* and *O. sativa* full-length cDNA clones to *H. sapiens* and *Triticum aestivum* TIGR gene indices and full-length cDNAs, respectively. *A. thaliana* and *D. melanogaster* were excluded from further analyses, because of the small sample size of alternative exons. As a result, we extracted 1008 constitutive exons and 413 alternative exons from the *M. musculus*–*H. sapiens* comparison, and 275 constitutive exons and 96 alternative exons from the *O. sativa*–*T. aestivum* comparison. There were originally 6659 constitutive exons and 1054 alternative exons in *M. musculus* and 1241 constitutive exons and 558 alternative exons in *O. sativa*. Numbers of constitutive and alternative exons were substantially reduced, because orthologous genes were not found in corresponding species or the exons that were not evolutionarily conserved, and those that had the different splice-site boundaries were excluded from data sets. All of the evolutionarily conserved constitutive and alternative exons are available as supplemental material online.

Evolutionarily conserved alternative exons and constitutive exons in orthologous genes were aligned using CLUSTAL-W (Thompson et al. 1994). The results of comparisons between *M. musculus* and *H. sapiens*, and *O. sativa* and *T. aestivum* (wheat) are shown in Figure 3 as distributions of nucleotide substitution rates at the 5' ends and 3' ends of evolutionarily conserved exons (constitutive and alternative). Note that data sets of constitutive and alternative exons were combined in order to see the general feature of *M. musculus* and *O. sativa* evolutionarily conserved exons. Figure 3 shows a "spike" pattern, in which nucleotide substitution rates were considerably higher at $3n$ bp from acceptor sites and at $3n + 1$ bp from donor sites than others; this might be due to synonymous substitutions and non-synonymous substitutions, because the length of an exon is a multiple of three in most cases (Tomita et al. 1996). Based on this observation, we additionally drew the average nucleotide substitution rate of the position $3n$, $3n + 1$, and $3n + 2$ of the "spikes" in Figure 3.

In the *M. musculus*–*H. sapiens* comparison (Fig. 3A), we observed a gradual increase of nucleotide substitution rates as the distance from boundaries of exons and introns increased. On the other hand, no such feature was seen in the distribution of nucleotide substitution rates in the *O. sativa*–*T. aestivum* comparison (Fig. 3B). We conducted the chi-squared test ($p < 0.05$, $df = 7$) and Student's unpaired t test ($p < 0.05$, two-sided) to assess this observation. Both statistical tests indicated that nucleotide substitution rates in positions -25 to -5 from the donor site are significantly lower than those in positions -50 to -26 from the donor site. Although we observed a gradual in-

TABLE 3. Information contents and Shapiro's score of constitutive and alternative splice sites

Species	Exon type	Data set size	Donor site (R_i , bits)	Donor site (Shapiro's method)	Acceptor site (R_i , bits)	Acceptor site (Shapiro's method)
<i>H. sapiens</i>	Constitutive	4,316	7.80 ± 3.30	79.97 ± 12.54	7.71 ± 3.98	81.67 ± 11.86
	Alternative	1,757	7.31 ± 3.65	78.40 ± 13.03	7.15 ± 4.02	80.41 ± 11.32
<i>M. musculus</i>	Constitutive	6,659	8.05 ± 2.51	81.39 ± 8.18	8.23 ± 2.87	83.26 ± 6.88
	Alternative	1,054	7.49 ± 2.96	79.76 ± 9.14	7.58 ± 3.18	82.28 ± 7.19
<i>A. thaliana</i>	Constitutive	7,138	7.34 ± 2.04	79.41 ± 6.65	7.70 ± 2.73	79.74 ± 8.28
	Alternative	25	6.76 ± 2.11	77.28 ± 6.90	6.76 ± 3.28	77.31 ± 9.37
<i>O. sativa</i>	Constitutive	1,241	7.09 ± 1.89	79.34 ± 6.69	7.63 ± 2.72	78.83 ± 9.32
	Alternative	558	6.79 ± 2.34	78.33 ± 7.85	7.09 ± 3.84	75.68 ± 11.04
<i>D. melanogaster</i>	Constitutive	286	8.44 ± 2.61	80.96 ± 7.79	7.29 ± 2.95	82.54 ± 8.60
	Alternative	62	6.53 ± 2.94	75.04 ± 8.97	6.27 ± 3.74	80.28 ± 9.80

The individual information (R_i) technique and Shapiro's method were applied to compare the splice-sites strength in constitutive and alternative exons. Average individual information contents, average Shapiro's score, and their standard deviation are shown for each exon type and for each of five species. Differences in average information contents and average Shapiro's score were assessed with Student's unpaired t -test ($P < 0.05$, two-sided). Both methods obtained exactly the same results: Alternative exons have significantly weaker splice-sites than constitutive exons, and plants have significantly weaker splice-sites than mammals. All of the weight matrices and individual information contents of all splice-sites are available as supplemental online material (<http://www.bioinfo.sfc.keio.ac.jp/research/intron/>).

crease of nucleotide substitution rates with distance from the acceptor site, Student's unpaired t test indicated that no statistically significant difference was observed in the nucleotide substitution rate in positions +3 to +25 from the acceptor site, in comparison with those in positions +26 to +50. Interestingly, no statistical significant differences were observed in the *O. sativa*–*T. aestivum* comparison. Exactly the same results were obtained from comparisons with cattle, pigs, and dogs for *M. musculus* and from comparisons with barley, maize, and rye for *O. sativa* (data not shown).

Comparison of potential regulatory sequences on alternative exons and constitutive exons

Constitutive exons and alternative exons were then compared in terms of nucleotide substitution rates. A Kolmogorov–Smirnov test indicated that the nucleotide substitution rates of alternative exons were significantly lower than those of constitutive exons in *M. musculus* (Fig. 4). Student's unpaired t test ($p < 0.05$, two-sided) verified the significantly low nucleotide substitution rates of alternative exons in *M. musculus*. Exactly the same feature was observed in the *O. sativa*–*T. aestivum* comparison, but the difference was not statistically significant because of the small sample size (96 data sets) of alternative exons (data not shown). Based on these results, we observed that alternative exons contain more evolutionarily conserved sequences than do constitutive exons.

Finally, we compared constitutive exons and alternative exons in terms of potential regulatory sequences on the exons. We extracted the top five candidates of potential regulatory sequences, which were observed more frequently in exons than expected (Table 4). In *M. musculus* alternative

exon, some of the potential regulatory sequences overlapped each other. To further assess this observation, we combined those candidates and counted the number of longer sequences. As a result, we found five AGAAGAAG (combination of GAAGAAG and AGAAGAA), seven AAGAAGAA (combination of AGAAGAA and AAGAAGA), three AAGAAGAAG (combination of GAAGAAG, AGAAGAA, and AAGAAGA), and one GATGAAGAAG (combination of GAAGAAG and GATGAAG).

Out of 62 motifs found in *M. musculus* alternative exon, 21 heptamers were found between –25 and –5 from the donor site, 14 heptamers between +3 and +25 from the acceptor site, and 27 heptamers in other positions. This result might indicate that those heptamers tend to concentrate in positions –25 to –5 from the donor site, which supports our previous observation that nucleotide substitution rates in positions –25 to –5 from the donor site are significantly lower than those in positions –50 to –26 from the donor site. However, the statistical significance of this result cannot be verified, because the length of alternative exons greatly varies.

In *M. musculus*, purine-rich motifs were frequently observed to exist only in alternative exons, whereas only a few such motifs occurred in constitutive exons (Table 4). In *M. musculus*, constitutive exons had more than twice as many data sets (1008 exons) as alternative exons (413 exons), but a larger number of potential regulatory sequences was extracted from alternative than from constitutive exons. To compute the level of conservation of these potential regulatory sequences, we divided the number of evolutionarily conserved heptamers by the total number of heptamers observed in the exons. Notably, the level of conservation of these heptamers was larger in alternative exons than in constitutive exons, both in *M. musculus* and *O. sativa*.

A *M. musculus* and *H. sapiens* comparison **B** *O. sativa* and *T. aestivum* comparison

FIGURE 3. Distribution of nucleotide substitution rates in the 5' end/3' end of evolutionarily conserved exons (constitutive + alternative exons). (A) Distributions of nucleotide substitution rates in 1421 evolutionarily conserved exons (1008 constitutive + 413 alternative) are shown as far as 50 bp from the acceptor site (*left*) and the donor site (*right*) of *M. musculus*. (B) Distributions of nucleotide substitution rates in 371 evolutionarily conserved exons (275 constitutive + 96 alternative) are shown as far as 50 bp from the acceptor site (*left*) and the donor site (*right*) of *O. sativa*. "Spike" patterns are probably due to synonymous substitutions and nonsynonymous substitutions, because the length of most exons is a multiple of three (Tomita et al. 1996). The average nucleotide substitution rates of the *bottom*, *middle*, and *upper* parts of the "spikes" are drawn as slopes. Statistical tests indicated that nucleotide substitution rates in position -25 to -5 from the *M. musculus* donor site are significantly lower than those in positions -50 to -26 from the donor site.

DISCUSSION**Alternative exons have weaker splice sites than constitutive exons**

We observed lower splice-site strength in alternative exons than in constitutive exons (Fig. 2). Khan et al. (2002) have reported the consequence of single nucleotide polymorphism (SNP) in the acceptor site of the human *XPC* gene. The polymorphism resulted in lower information content

in the acceptor site, indicating that a strong acceptor site has a few alternative splicing events whereas a weak acceptor site has frequent exon skipping (Khan et al. 2002). However, this reported characteristic of splice sites has been obtained from experimental validation of only one gene; thus, there has been no way to distinguish if this characteristic is (1) specific to the *XPC* gene, (2) specific to humans, or (3) universal. Here, we compared thousands of alternative exons and constitutive exons from five different species,

FIGURE 4. Histograms of nucleotide substitution rates of alternative and constitutive exons in *M. musculus*. Light bars constitute a histogram of nucleotide substitution rates in alternative exon (413 data sets); dark bars, a histogram of these rates in constitutive exons (1008 data sets). Statistical tests indicated that the nucleotide substitution rates of alternative exons were significantly lower than those of constitutive exons; this result implies that alternative exons contain more evolutionarily conserved sequences than do constitutive exons.

and statistically verified that alternative exons have significantly weaker splice sites than constitutive exons (Fig. 2; Table 3).

Stamm et al. (1994) have created a database of alternatively spliced exons in neurons and indicated that donor site consensus (CAG|GTAAGT) is present 40% less often in neuron-specific alternative exons than in constitutive exons. This study seems to suggest that alternative exons possess weaker splice sites than constitutive exons do. However, a later study by Rogan and Schneider (1995) stated that splice-site sequences that deviate from the consensus do not necessarily produce significantly lower amounts of spliced mRNA. In contrast, the information-theory-based method does not destroy subtle details of splice-site sequences. Furthermore, the total information at the splice sites can be easily determined by use of the individual information (R_i) technique, which adds up the information from individual positions.

The individual information technique and Shapiro's method were applied to compare the splice-site strength in constitutive and alternative exons. Both methods obtained exactly the same results, that alternative exons have significantly weaker

TABLE 4. Predicted potential regulatory sequences in *M. musculus* and *O. sativa*

<i>M. musculus</i>							
Alternative exon				Constitutive exon			
Sequence	Observed	O/E value	Level of conservation	Sequence	Observed	O/E value	Level of conservation
GAAGAAG	13	1.78	43.3%	TGGAGAA	10	2.17	22.2%
CTGGAGC	13	2.03	56.5%	GAGCTGG	10	2.07	25.6%
AGAAGAA	12	1.60	44.4%	TGCTGCT	9	2.56	18.4%
AAGAAGA	12	1.65	50.0%	GCTGCTG	9	1.52	14.1%
GATGAAG	12	3.09	57.1%	CTGCTGC	9	1.52	15.3%
<i>O. sativa</i>							
Alternative exon				Constitutive exon			
Sequence	Observed	O/E value	Level of conservation	Sequence	Observed	O/E value	Level of conservation
GACAAGA	4	2.40	44.4%	TGAAGAA	8	2.33	25.8%
ACAAGCT	4	1.75	40.0%	CCAAAGA	6	2.10	33.3%
AGGAAGA	4	1.72	33.3%	AACAAGC	6	2.19	46.2%
CAAGTAC	3	4.14	33.3%	GTGGTGG	6	2.33	27.3%
GAGGACG	3	3.61	50.0%	AGAAGCA	6	1.75	25.0%

Top five candidates of potential regulatory sequence, which were extracted from 1008 *M. musculus* constitutive exons, 413 *M. musculus* alternative exons, 275 *O. sativa* constitutive exons, and 96 *O. sativa* alternative exons. Because some of the motifs seemed to overlap to each other, we combined those candidates and counted the number of longer sequences. We found five AGAAGAAG (combination of GAAGAAG and AGAAGAA), seven AAGAAGAA (combination of AGAAGAA and AAGAAGA), three AAGAAGAAG (combination of GAAGAAG, AGAAGAA, and AAGAAGA), and one GATGAAGAAG (combination of GAAGAAG and GATGAAG). The level of conservation of these potential regulatory sequences was computed by dividing the number of evolutionary conserved heptamers by the total number of heptamers observed in the exons.

splice sites than constitutive exons. Because alternative exons contain weaker splice sites than constitutive exons, weak splice sites must be one of the fundamental factors of alternative splicing regulation and a universal characteristic of alternative exons in *H. sapiens*, *M. musculus*, *A. thaliana*, *O. sativa*, and *D. melanogaster*.

3' ends of exons are more evolutionarily conserved

We observed a gradual increase of nucleotide substitution rates with increasing distance from splice sites in both *M. musculus* and *H. sapiens* but not in *O. sativa* or *T. aestivum* (Fig. 3). As indicated by Blanchette and Tompa (2002), it is highly possible that those evolutionarily conserved sequences among a set of orthologous genes are potential regulatory sequences. A number of experimental studies have identified potential regulatory sequences of pre-mRNA splicing, mostly in individual genes. Although these studies may have provided an excellent novel feature of those potential regulatory sequences, most of them have not shown where those regulatory sequences are most likely to be found.

Several statistical analyses of our results validated that nucleotide substitution rates at positions -25 to -5 from the donor site are significantly lower than at other regions in the *M. musculus*–*H. sapiens* comparison (Fig. 3A). It is reasonable to suggest that evolutionarily conserved sequences, possibly some regulatory sequences, are more likely to be found at the 3' end of an exon. Although we adopted the same approach and statistical validation methods in the *O. sativa*–*T. aestivum* comparison, no remarkable feature was observed in the distribution of nucleotide substitution rates (Fig. 3B). Here, we have observed potential differences between mammals (*M. musculus*) and plants (*O. sativa*) in the distribution of evolutionarily conserved sequences, possibly including some regulatory sequences; potential regulatory sequences are more likely to be found at the 3' end of exons in mammals, but not in plants.

Alternative exons have more evolutionarily conserved sequences than constitutive exons

In the *M. musculus*–*H. sapiens* comparison, both the KS test and *t* test indicated that alternative exons contain more evolutionarily conserved sequences than constitutive exons (Fig. 4). This observation has led to our hypothesis that alternative exons have more potential regulatory sequences than constitutive exons, because selective pressure causes functional elements to evolve at a slower rate than that of nonfunctional sequences (Blanchette and Tompa 2002). A recent interesting finding on the evolution of alternative exons can further support our hypothesis.

Modrek and Lee (2003) have indicated that the inclusion level of an exon in *H. sapiens* ESTs is highly correlated with that in *M. musculus*, and implied that the evolutionarily

conserved alternative exons are similarly regulated in both organisms. This observation may further support our hypothesis that alternative exons have more potential regulatory sequences than constitutive exons; if the evolutionarily conserved exons are similarly regulated, potential regulatory sequences on the exons are more likely to be conserved as well.

Modrek and Lee (2003) have also suggested that when an exon is present in the ortholog of one genome but not the other, there could be exon creation or exon loss during the evolution of that genome; thus alternative splicing is greatly associated with the exon creation and/or exon loss during the evolution. Another study by Sorek et al. (2004) has suggested that the conservation of alternative exons in more than one species suggests the functional importance of the alternative exons, in comparison with nonconserved alternative exons. We extracted evolutionarily conserved alternative exons from our *M. musculus*–*H. sapiens* comparison and our *O. sativa*–*T. aestivum* comparison; as such, those evolutionarily conserved alternative exons must have been created before the branching of *M. musculus* and *H. sapiens* or *O. sativa* and *T. aestivum* and most likely have some functional importance.

Alternative exons have more potential regulatory sequences than constitutive exons

We compiled large numbers of both constitutive and alternative exons and discovered that alternative exons have more evolutionarily conserved sequences than constitutive exons (Fig. 4), and that alternative exons contain more purine-rich potential regulatory sequences than constitutive exons (Table 4). It must be admitted that 13 cases out of a data set of 413 *M. musculus* alternative exons might be rather few. However, we believe that this small number is due to our strict requirements, that we only included the heptamers with exact matches. It is well known that although SR protein have distinct RNA-binding specificities, the consensus sequences that they recognize are rather degenerate (Graveley 2000). As such, if we included weaker matches, the number may substantially increase. In addition, it is unlikely that those heptamers were extracted by a probabilistic chance, because we used a second-order Markov model to compute the expected value of the heptamers, and extracted the heptamers whose O/E value is greater than 1.5. Although a number of potential SR-protein-binding sites have been compiled in literature, there has been no knowledge of how those sequences are distributed between alternative and constitutive exons. Thus, it has been difficult to present a broad view of potential regulatory sequences in the regulation of alternative splicing.

In the *M. musculus*–*H. sapiens* comparison, both the KS test and *t* test indicated that alternative exons contain more evolutionarily conserved sequences than constitutive exons (Fig. 4). This observation has led to our hypothesis that

alternative exons have more potential regulatory sequences than constitutive exons. To assess this hypothesis, we extracted potential regulatory sequences in both alternative and constitutive exons (Table 4). In *M. musculus*, constitutive exons had more than twice as many data sets (1008 exons) as alternative exons (413 exons), but a larger number of potential regulatory sequences was extracted from alternative exons. It is also worth noting that the level of conservation of those heptamers was greater in alternative exons than in constitutive exons. Alternative exons in *M. musculus* contained a number of purine-rich motifs; with a few exceptions, most ESE motifs are reported to be purine rich. Notably, some of these predicted potential regulatory sequences on alternative exons have a high sequence similarity with the binding site for ASF/SF2, which is RGAAGAAC (Tacke and Manley 1995), with the binding site for Tra2, which is GAA repeats (Tacke et al. 1998), and with predicted RESCUE-ESE motifs, which is GAAGAA (Fairbrother et al. 2002). Only a few purine motifs were found in constitutive exons. In *O. sativa*, despite the small sample size (96 alternative exons), we also observed a lower nucleotide substitution rate in alternative than in constitutive exons and a higher level of conservation of the heptamers in alternative exons than in constitutive exons (Table 4). Based on these results, we concluded that another important factor of alternative splicing is to have more purine-rich regulatory sequences than are present in constitutive exons.

Plant–mammal comparison of various splicing regulatory factors

In our comparative analyses of plants and mammals, our two main observations were that plants have significantly higher information contents between positions +7 and +15 downstream from the donor site than other species (Supplemental Fig. 1, available online at <http://www.bioinfo.sfc.keio.ac.jp/research/intron/>), and that plants have significantly weaker splice sites than mammals (Fig. 2). McCullough et al. (1993) have suggested that the donor sites located at transition regions from GC- to AT-rich sequences are preferentially selected in the pea *RBCS3A* gene. Again, this reported characteristic on plant introns was obtained from experimental validation of only one pea gene; thus, there has been no way to distinguish if this characteristic is (1) specific to the *RBCS3A* gene, (2) specific to peas, or (3) universal in plants. Furthermore, the fundamental role of this compositional bias in plant splicing has remained unclear.

Our comparative analysis of plants and mammals may allow us to present a model for plant splicing. The higher information contents in positions +7 to +15 downstream from the donor site possibly represents the compositional bias (i.e., AT richness) of plant introns (Supplemental Fig. 1, available online at <http://www.bioinfo.sfc.keio.ac.jp/research/intron/>). Also, the lower information contents

of positions +3 to +6 downstream from the donor site and those of positions –13 to –8 upstream of the acceptor site possibly have resulted in weak donor and acceptor sites, respectively (Fig. 2; Supplemental Fig. 1, available online at <http://www.bioinfo.sfc.keio.ac.jp/research/intron/>). Those observations have been obtained from comparative analyses of more than 1000 splice sites from five different species, and are verified by solid statistical tests. Thus, we can suggest a model for plant splicing: Plants have a strong compositional bias in their introns to support the relatively weaker splice sites.

In addition to the strong compositional bias and weaker splice sites, we have observed potential differences between mammals (*M. musculus*) and plants (*O. sativa*) in the distribution of evolutionarily conserved sequences; we observed a gradual increase of nucleotide substitution rates with increasing distance from splice sites in *M. musculus* but not in *O. sativa* (Fig. 3). This result might be due to a compositional bias of plant introns, that the splice sites located at transition regions from GC- to AT-rich sequences are preferentially selected in plant splicing. Also, although purine-rich motifs were found only in the alternative exons as potential regulatory sequences in *M. musculus*, several purine-rich motifs were found in both the constitutive and alternative exons in *O. sativa* (Table 4). This result is most likely due to the rather small sample size (96 alternative exons). However, these features of potential regulatory sequences might be potential differences between mammals and plants, because there are several plant-specific requirements for pre-mRNA splicing, such as weaker splice sites and a compositional bias in their introns.

A “weaker/more” combinatorial model of alternative splicing regulatory mechanisms

Our main discoveries have been that alternative exons have weaker splice sites than constitutive exons, and that alternative exons have more potential regulatory sequences than constitutive exons. These observations raise one important question: Is the weakness of the splice sites connected to the greater potentiality for regulatory sequences? A recent experimental study reported a fascinating observation regarding this issue. Cystic fibrosis transmembrane regulator (CFTR) has an alternative exon (exon 12) whose acceptor site has a relatively weak consensus (AAG|GTATGA). Pagani et al. (2003) have shown that a point mutation at position +4 downstream from the consensus can strengthen the acceptor site, which leads the alternative exon to express constitutively. In addition, they have confirmed that a point mutation in the exonic regulatory sequence (GGATAC) of the alternative exon results in a severe splicing defect, which, surprisingly, increases the exclusion of the alternative exons from the mRNA transcript. Pagani et al.’s study indicates that both weak splice sites and exonic regulatory sequences on the alternative exon are indispensable to the

alternative regulation of exon 12 in the *CFTR* gene. Once again, this excellent observation is obtained from just one individual gene and does not present a broad view of alternative splicing regulation.

We have indicated that alternative exons have (1) weaker splice sites and (2) more potential exonic regulatory sequences than constitutive exons. It is reasonable to suggest that the fundamental role of weak splice sites in alternative exons is to have the flexibility to be included or excluded from mature mRNA transcripts; if the splice sites were stronger, the exon would lose this flexibility and express constitutively. Also, such alternative exons may have more potential regulatory sequences to be regulated in a cell-type specific manner because SR proteins, which are known to bind to such regulatory sequences, also express in a cell-type-specific manner. Taking together all of these observations, we propose a “weaker/more combinatorial model” as a potential model of alternative splicing regulatory mechanisms: This model suggests that alternative exons contain weaker splice sites to be regulated alternatively by more potential regulatory sequences on the exons.

We applied our weaker/more combinatorial model of alternative splicing regulatory mechanisms to *M. musculus* and *O. sativa* genes. Figure 5A illustrates the gene structure and multiple sequence alignment of the alternative exon for *M. musculus* immunoglobulin, conserved among *M. musculus* (mouse), *H. sapiens* (human), chicken, and cattle. Immunoglobulin is well known to function as an antibody, and has five major classes with distinct functions in immune response. We observed potential regulatory sequences of alternative splicing, AAGAAGA and AGAAGAA, both of which have been evolutionarily conserved more frequently than expected in other alternative exons of *M. musculus* (boxed in Fig. 5A). As we have described above, the purine-rich motif AGAAGAA has remarkably high sequence similarity with the binding site of ASF/SF2. In addition, information contents of the donor site (9.02 bits) and acceptor site (−1.30 bits) in the alternative exon are lower than those of the donor site (10.42 bits) and acceptor site (9.98 bits) in the adjacent constitutive exons.

Figure 5B illustrates another example, the *O. sativa* *atp6* gene for plasma membrane H⁺ ATPase, which is known to be involved in many physiological functions, and acts as a primary transporter by pumping protons out of the cell and creating pH and electrical potential differences across plasmalemma (Michelet and Boutry 1995). As illustrated in Figure 5B, the alternative exon is conserved among *O. sativa*, *T. aestivum* (wheat), and sorghum. Likewise, the evolutionarily conserved region contains ACAAGCT, which has been conserved more than expected in other alternative exons of *O. sativa*. Additionally, the alternative exon has weaker splice sites, 2.53 bits for the donor site and 8.19 bits for the acceptor site, than the adjacent constitutive exons, which have 5.73 bits for the donor site and 9.96 bits for the acceptor site.

Our large-scale comparative analyses and statistical validation of more than 1000 alternative exons have provided substantial evidence for our weaker/more combinatorial model of alternative splicing regulation. Taking together all the observations above and solid statistical validations of our results, we can conclude that alternative exons contain weaker splice sites in order to be regulated alternatively by potential regulatory sequences, which are found more frequently in alternative exons than in constitutive exons.

For the past several years, the regulatory functions of many splicing regulatory sequences in individual genes have only been anecdotally reported. We have applied, for the first time, comparative analyses of various transcriptomes to delineate a potential model of alternative splicing regulatory mechanisms. Our bioinformatics approach may thus represent the best model for transcriptome analysis of alternative splicing.

MATERIALS AND METHODS

Detection of putative splice variants

All of the acquired data sets and database URLs are listed in Table 5. We used full-length cDNA and genomic DNA to construct data sets of putative splice variants and infer complete gene structures. For the past several years, expressed sequence tags (ESTs) have been the most effective materials for estimating the frequency of alternative splicing and detecting novel candidates for alternatively spliced genes (Mironov et al. 1999). However, a recent study by Sorek and Safer (2003) showed how a highly contaminated EST library causes incorrect inference of an enormous number of new splice variants; hence, we used much more desirable material, full-length cDNA sequences, for detecting putative splice variants (Kochiwa et al. 2002).

Mapping the full-length cDNA of complete genomic sequences was performed through the use of BLAST (Altschul et al. 1990), followed by SIM4 (Florea et al. 1998). Clones were successfully mapped if the e-value was lower than E-100. Based on the roughly determined location of each of the clones on the genomic sequences, SIM4 was used to find the exact location and complete gene structure of each clone. Clones were excluded from further analysis if any of the exons had lower than a 90% match or if internal regions were not correctly mapped. Full-length cDNA sequences were grouped into the same cluster if they were mapped in the same orientation and if their genomic mappings overlapped. Of these clusters, single-exon-derived transcripts, alternative promoters, alternative terminal exons, and simple redundant clusters (clusters with no variation of any patterns) were excluded from further analysis. If introns and exons had different boundaries in different transcripts, the cluster was thought to be a “splice variant”. Mapping and clustering results are shown in Table 1.

Because it is difficult to identify where potential regulatory sequences exist in exons with different donor/acceptor sites and thus impossible to compare them with constitutive exons, we focused only on cassette exons for our analysis (Table 2); we defined an *alternative exon* as a “cassette exon” and a *constitutive exon* as “an internal exon that is conserved across the entire transcript in a splice variant cluster”.

FIGURE 5. Examples of the “weaker/more combinatorial model” in *M. musculus* and *O. sativa* genes. (A) Gene structure and multiple sequence alignment of an alternative exon, conserved among mouse, chicken, cattle, and human, in the *M. musculus* immunoglobulin gene. (B) Gene structure and multiple sequence alignment of an alternative exon, conserved among *O. sativa*, sorghum, and wheat, in the *O. sativa* a1 gene for plasma membrane H⁺ ATPase.

Information content of constitutive and alternative splice sites

Information content on splice sites was calculated based on Shannon’s information theory (Stephens and Schneider 1992). We computed the “uncertainty” at a position by the following equation:

$$R_{\text{sequence}}(l) = 2 + \sum_{b \in \{a, c, g, t\}} f(b, l) \times \log_2 f(b, l)$$

where $f(b, l)$ is the probability of base b at position l .

To compare the average information contents of constitutive and alternative splice sites, we applied the individual information

TABLE 5. Acquired data

Species	Sequence	Data sets	Database
<i>H. sapiens</i>	fl-cDNA	21,243	New Energy and Industrial Technology Development Organization http://www.nedo.go.jp/bio-e/ Working Draft Sequence. As of April 2001 http://genome.cse.ucsc.edu/goldenPath/01apr2001/bigZips/
	genomic DNA		
<i>M. musculus</i>	fl-cDNA	60,770	The Institute of Physical and Chemical Research (RIKEN) http://fantom2.gsc.riken.go.jp/ Mouse Genome Sequencing Consortium ftp://ftp.ensembl.org/pub/assembly/mouse/mgsc_assembly_3
	genomic DNA		
<i>A. thaliana</i>	fl-cDNA	8,095	The Institute of Physical and Chemical Research (RIKEN) http://pfgweb.gsc.riken.go.jp/pub_data/ The Institute for Genomic Research ftp://ftp.tiger.org/pub/data/a_thaliana/ceres/
	genomic DNA		
<i>O. sativa</i>	fl-cDNA	28,469	National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/Genomes/ National Institute Agrobiological Sciences http://cdna01.dna.affrc.go.jp/cDNA/ Rice Genome Research Program at National Institute Agrobiological Sciences http://rgp.dna.affrc.go.jp TMRI
	genomic DNA		
<i>D. melanogaster</i>	fl-cDNA	9,872	Berkeley Drosophila Genome Project resources http://www.fruitfly.org/sequence/download.html Berkeley Drosophila Genome Project resources http://www.fruitfly.org/sequence/download.html
	genomic DNA		

In addition to these data sets, we constructed online supplemental data at: <http://www.bioinfo.sfc.keio.ac.jp/research/intron/>.

(R_i) technique (Schneider 1997). We first generated an individual information weight matrix from the frequencies of each nucleotide at each position for each of five species. All of the weight matrices are available as supplemental online material. The individual information weight matrix can be calculated by the following equation:

$$R_{iw}(b,l) = 2 + \log_2 f(b,l).$$

The information content of each individual splice site was calculated by summing $R_{iw}(b,l)$ of the specified positions. We added $R_{iw}(b,l)$ of positions -3 to $+7$ for donor sites, because information contents were observed to be saturated downstream from position -3 and upstream of position $+7$ (Supplemental Fig. 1, available online at <http://www.bioinfo.sfc.keio.ac.jp/research/intron/>). Similarly, we summed $R_{iw}(b,l)$ of positions -13 to $+1$ for acceptor sites to observe the potential differences between plants and mammals downstream from position -13 , and because information contents seemed to be saturated downstream from position $+1$ (Supplemental Fig. 1, available online at <http://www.bioinfo.sfc.keio.ac.jp/research/intron/>). All individual information contents of all splice sites are available as supplemental online material. Using the information contents of each individual splice site, we then computed the average individual information content of alternative and constitutive splice sites (Fig. 2). Differences between these two averages were assessed with Student's unpaired t test ($p < 0.05$, two-sided).

Shapiro's score of constitutive and alternative splice sites

Shapiro and Senapathy (1987) have developed a method to score the strength of a splice site based on percentages of each nucleotide at each position. Shapiro's score of donor site is $100 * (t - \min) / (\max - \min)$, where t is the sum of percentages at positions -3 to

$+7$, $\min$ is the sum of the lowest percentages at positions -3 to $+7$, and $\max$ is the sum of the highest percentages at positions -3 to $+7$. On the other hand, Shapiro's score of acceptor site is $100 * ((t1 - l1)/(h1 - l1) + (t2 - l2)/(h2 - l2))/2$, where $t1$ is the sum of the best 8 of 10 percentages at positions -13 to -4 , $l1$ is the sum of the lowest 8 of 10 percentages at position -13 to -4 , $h1$ is the sum of the highest 8 of 10 percentages at positions -13 to -4 , $t2$ is the sum of percentages at positions -3 to $+1$, $l2$ is the sum of the lowest percentages at positions -3 to $+1$, and $h2$ is the sum of the highest percentages at positions -3 to $+1$.

All of the weight matrices and individual information contents of all splice sites are available as supplemental online material.

Prediction of potential regulatory sequences in alternative exons

Phylogenetic footprinting is one of the well-known approaches to predicting regulatory sequence, by which unusually well-conserved sequences among a set of orthologous genes are extracted as candidates for functional regulatory elements (Blanchette and Tompa 2002). This method has been broadly applied to predict potential regulatory sequences, including novel functional sequence motifs in the promoter region (Cliften et al. 2003) and transcription factor binding sites (McCue et al. 2002). To collect orthologous genes, we retrieved cattle, dog, human, and pig genes for *M. musculus* and barley, maize, and wheat for *O. sativa*. (Note from Fig. 5 that we used an additional species, chicken for *M. musculus* and sorghum for *O. sativa*, for multiple sequence alignment in our examples.) We mapped all *M. musculus* and *O. sativa* full-length cDNA clones to TIGR gene indices (<http://www.tgi.org/tdb/tgi/>) by use of BLAST (E-value: E-50 or less). We then extracted evolutionarily conserved alternative exons and constitutive exons in orthologous genes and aligned the extracted

exons using CLUSTAL-W (Thompson et al. 1994). Using the alignment data, we computed the nucleotide substitution rate at each position for all exons (constitutive + alternative).

Both a reasonable amount of evolutionary distance and a sufficient number of data sets are necessary to apply a phylogenetic footprinting approach to prediction of functional regulatory sequences; hence, we chose *M. musculus*–*H. sapiens* and *O. sativa*–*T. aestivum* (wheat) comparisons, which had the largest number of evolutionarily conserved exons available for both alternative and constitutive exons. To access the differences in the distribution of nucleotide substitution rates (Fig. 3), we first divided an exon into four regions: positions +3 to +25 from the acceptor site, positions +26 to +50 from the acceptor site, positions –50 to –26 from the donor site, and positions –25 to –5 from the donor site. We excluded positions +1 and +2 from the acceptor site and –4 to –1 from the donor site to avoid possible bias of splice-site consensus sequences. The expected rate of nucleotide substitutions was computed by calculating the average nucleotide substitution rates at all regions. In addition, Student's unpaired *t* test was conducted to average the nucleotide substitution rates in given regions to further validate our observations.

We then compared the nucleotide substitution rates per exon and performed a Kolmogorov–Smirnov test and Student's unpaired *t* test to see if any differences existed between the nucleotide substitution rate histograms for constitutive and alternative exons (Fig. 4); the level of significance was $p < 0.05$ (two-sided). Noting that alternative exons contain more conserved sequences, we extracted the evolutionarily conserved sequences whose lengths ranged from 7 bp to 20 bp from the alignment results of alternative exons and constitutive exons. Maximum length was set to 20 bp to avoid extracting false positives from unusually long conserved sequences. The minimum length was set to 7 bp because known exonic splicing enhancer motifs, identified experimentally, have an average length of approximately 7 bp.

We then computed the expected value of all possible combinations of 7-bp motifs in the extracted sequences. To consider codon bias in the coding region, we used a second-order Markov model and computed the expected value by the following equation: expected number of GATCATC was $n(G) * p(T|GA) * p(C|AT) * p(A|TC) * p(T|CA) * p(C|AT)$, where $n(G)$ is the number of nucleotide Gs and $p(C|AT)$ is the probability that nucleotide C comes after dinucleotide AT. The five most frequently observed sequence motifs were extracted as candidates for potential regulatory sequences for alternative splicing. To compute the level of conservation of these potential regulatory sequences, we divided the number of evolutionarily conserved heptamers by the total number of heptamers observed in the exons. For example, heptamer CTGGAGC was observed in 23 *M. musculus* alternative exons, 13 of which are perfectly conserved in *H. sapiens*; the level of conservation of the heptamers was 13/25, which is 56.5%.

ACKNOWLEDGMENTS

We thank Hiromi Kochiwa, Ryosuke Suzuki, Atsushi Sakurai, Rintaro Saito, Sumie Kitamura-Abe, Noriyuki Kitagawa, Haruo Suzuki, and members of the Institute for Advanced Biosciences for their helpful suggestions and discussions during the course of this work. We also thank members of the CASYS (cDNA Analysis System) consortium in the G-language project for their advice/comments on the mapping procedure. Finally, we thank Professor Ko

Shimamoto and Assistant Professor Masayuki Isshiki of the Laboratory of Plant Molecular Genetics, Nara Institute of Science and Technology, for many useful discussions, especially of *O. sativa* alternative splicing. This work was supported by Japan's Ministry of Agriculture, Forestry, and Fisheries (Rice Genome Project SY-1104), and by the 21st Century COE Program of Japan's Ministry of Education, Culture, Sport, Science and Technology.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

Received October 31, 2003; accepted April 21, 2004.

REFERENCES

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.
- Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12**: 739–748.
- Cartegni, L. and Krainer, A.R. 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat. Genet.* **30**: 377–384.
- Claverie, J.M. 2001. Gene number. What if there are only 30,000 human genes? *Science* **291**: 1255–1257.
- Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**: 71–76.
- Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**: 1007–1013.
- Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**: 967–974.
- Graveley, B.R. 2000. Sorting out the complexity of SR protein functions. *RNA* **6**: 1197–1211.
- Khan, S.G., Muniz-Medina, V., Shahnavi, T., Baker, C.C., Inui, H., Ueda, T., Emmert, S., Schneider, T.D., and Kraemer, K.H. 2002. The human XPC DNA repair gene: Arrangement, splice site information content and influence of a single nucleotide polymorphism in a splice acceptor site on alternative splicing and function. *Nucleic Acids Res.* **30**: 3624–3631.
- Kochiwa, H., Suzuki, R., Washio, T., Saito, R., Bono, H., Carninci, P., Okazaki, Y., Miki, R., Hayashizaki, Y., and Tomita, M. 2002. Inferring alternative splicing patterns in mouse from a full-length cDNA library and microarray data. *Genome Res.* **12**: 1286–1293.
- Lim, L.P. and Burge, C.B. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci.* **98**: 11193–11198.
- McCue, L.A., Thompson, W., Carmack, C.S., and Lawrence, C.E. 2002. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.* **12**: 1523–1532.
- McCullough, A.J., Lou, H., and Schuler, M.A. 1993. Factors affecting authentic 5' splice site selection in plant nuclei. *Mol. Cell. Biol.* **13**: 1323–1331.
- McKeown, M. 1992. Alternative mRNA splicing. *Annu. Rev. Cell Biol.* **8**: 133–155.
- Michelet, B. and Boutry, M. 1995. The plasma membrane H⁺-ATPase (a highly regulated enzyme with multiple physiological functions). *Plant Physiol.* **108**: 1–6.
- Mironov, A.A., Fickett, J.W., and Gelfand, M.S. 1999. Frequent alternative splicing of human genes. *Genome Res.* **9**: 1288–1293.

- Modrek, B. and Lee, C. 2002. A genomic view of alternative splicing. *Nat. Genet.* **30**: 13–19.
- . 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* **34**: 177–180.
- Moller, L.B., Tumer, Z., Lund, C., Petersen, C., Cole, T., Hanusch, R., Seidel, J., Jensen, L.R., and Horn, N. 2000. Similar splice-site mutations of the ATP7A gene lead to different phenotypes: Classical Menkes disease or occipital horn syndrome. *Am. J. Hum. Genet.* **66**: 1211–1220.
- Pagani, F., Stuani, C., Tzetis, M., Kanavakis, E., Efthymiadou, A., Doudounakis, S., Casals, T., and Baralle, F.E. 2003. New type of disease causing mutations: The example of the composite exonic regulatory elements of splicing in CFTR exon 12. *Hum. Mol. Genet.* **12**: 1111–1120.
- Rogan, P.K. and Schneider, T.D. 1995. Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum. Mutat.* **6**: 74–76.
- Schneider, T.D. 1997. Information content of individual genetic sequences. *J. Theor. Biol.* **189**: 427–441.
- Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., et al. 2002. Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* **296**: 141–145.
- Shapiro, M.B. and Senapathy, P. 1987. RNA splice junctions of different classes of eukaryotes: Sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **15**: 7155–7174.
- Sorek, R. and Safer, H.M. 2003. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.* **31**: 1067–1074.
- Sorek, R., Shamir, R., and Ast, G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20**: 68–71.
- Stamm, S., Zhang, M.Q., Marr, T.G., and Helfman, D.M. 1994. A sequence compilation and comparison of exons that are alternatively spliced in neurons. *Nucleic Acids Res.* **22**: 1515–1526.
- Stapleton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S., et al. 2002. A *Drosophila* full-length cDNA resource. *Genome Biol.* **3**: research0080.0081–0080.0088.
- Stephens, R.M. and Schneider, T.D. 1992. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* **228**: 1124–1136.
- Szathmary, E., Jordan, F., and Pal, C. 2001. Molecular biology and evolution. Can genes explain biological complexity? *Science* **292**: 1315–1316.
- Tacke, R. and Manley, J.L. 1995. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J.* **14**: 3540–3551.
- Tacke, R., Tohyama, M., Ogawa, S., and Manley, J. 1998. Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. *Cell* **93**: 139–148.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. 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.
- Tomita, M., Shimizu, N., and Brutlag, D.L. 1996. Introns and reading frames: Correlation between splicing sites and their codon positions. *Mol. Biol. Evol.* **13**: 1219–1223.
- Zavolan, M., van Nimwegen, E., and Gaasterland, T. 2002. Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Res.* **12**: 1377–1385.

RNA

A PUBLICATION OF THE RNA SOCIETY

Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes

HITOMI ITOH, TAKANORI WASHIO and MASARU TOMITA

RNA 2004 10: 1005-1018

References

This article cites 34 articles, 13 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/10/7/1005.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:
<http://rnajournal.cshlp.org/subscriptions>