DNA AND CELL BIOLOGY
Volume 19, Number 12, 2000
Mary Ann Liebert, Inc.
Pp. 739–756

An Alternative-Exon Database and Its Statistical Analysis

STEFAN STAMM,[1,*] JIAN ZHU,[2] KENTA NAKAI,[3] PETER STOILOV,[1] OLIVER STOSS,[1] and MICHAEL Q. ZHANG[2]

ABSTRACT

We compiled a comprehensive database of alternative exons from the literature and analyzed them statistically. Most alternative exons are cassette exons and are expressed in more than two tissues. Of all exons whose expression was reported to be specific for a certain tissue, the majority were expressed in the brain. Whereas the length of constitutive exons follows a normal distribution, the distribution of alternative exons is skewed toward smaller ones. Furthermore, alternative-exon splice sites deviate more from the consensus: their 3′ splice sites are characterized by a higher purine content in the polypyrimidine stretch, and their 5′ splice sites deviate from the consensus sequence mostly at the +4 and +5 positions. Furthermore, for exons expressed in a single tissue, adenosine is more frequently used at the −3 position of the 3′ splice site. In addition to the known AC-rich and purine-rich exonic sequence elements, sequence comparison using a Gibbs algorithm identified several motifs in exons surrounded by weak splice sites and in tissue-specific exons. Together, these data indicate a combinatorial effect of weak splice sites, atypical nucleotide usage at certain positions, and functional enhancers as an important contribution to alternative-exon regulation.

INTRODUCTION

ALTERNATIVE SPLICING IS A VERSATILE MECHANISM to regulate gene expression at the level of pre-mRNA processing. The regulation of splicing was probably crucial for the evolution of eukaryotes (Herbert and Rich, 1999). Proper splicing regulation is important for an organism, as it has been estimated that as many as 15% of genetic defects caused by point mutations in humans manifest themselves as pre-mRNA splicing defects caused by changed splice site sequences (Krawczak *et al.*, 1992; Nakai and Sakamoto, 1994; Philips and Cooper, 2000). In addition, it became apparent that point mutations in exons can cause missplicing by changing exonic sequence elements (Cooper and Mattox, 1997), for example in tauopathies (Gao *et al.*, 2000) or spinal muscular atrophy (Lorson *et al.*, 1999). A recent survey of disease-associated genes suggested that as much as a third might be alternatively spliced, suggesting that more diseases might be associated with splicing defects (Hanke *et al.*, 1999).

Significant progress has been made in understanding the mechanism of constitutive splicing. Three major *cis* elements on RNA have been defined; the 5′ and 3′ splice sites and the branch point. These elements are recognized by the spliceosome, a 60S complex containing small nuclear RNAs (U1, U2, U4, U5, U6) and more than 50 different proteins (Neubauer *et al.*, 1998). In the spliceosome, U1 snRNP, U2AF, SF1, U6 snRNP, and U2 snRNP are the *trans*-acting factors that ultimately recognize the 5′ and 3′ splice sites and the branch point (reviewed in Green, 1991; Krämer, 1996; Elliot, 2000; Moore, 2000). Proper recognition of these sites is facilitated by serine-arginine (SR) proteins (Manley and Tacke, 1996) and hnRNPs (Weighardt *et al.*, 1996) binding to exonic or intronic elements of the pre-mRNA. The characterization of RNA sequences binding to these *trans*-acting factors by systematic evolution of ligands by exponential enrichment (SELEX) revealed several consensus sequences that are usually degenerate (e.g.; Watakabe *et al.*, 1993; Tian and Kole, 1995; Coulter *et al.*, 1997; Perez *et al.*, 1997; Liu *et al.*, 1998, 2000; Tacke and Manley, 1999). The splice site recognition medated by this complex most likely occurs concomitantly with the transcription of the RNA (Cramer *et al.*, 1999) in a large complex termed "RNA factory" or "transcriptosome" (Du and Warren, 1996; McCracken *et al.*, 1997; Nayler *et al.*, 1998).

In contrast to constitutive splicing, the regulation of which is largely known, the mechanisms regulating alternative exon

[1]Max-Planck Institute of Neurobiology, Planegg, Germany.
[*]*Present address*: Institute of Biochemistry, University of Erlangen-Nuremberg, Erlangen, Germany.
[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
[3]Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan.

usage are just beginning to emerge. It is clear that a fine-tuned balance of interactions between *cis* and *trans* factors is responsible for alternative splice site selection (Black, 1995; Grabowski, 1998; Varani and Nagai, 1998). Alternative-exon usage is frequently associated with a certain tissue type or developmental stage. It is currently debated whether different concentrations of *trans*-acting factors in different cell types and developmental stages cause alternative processing of pre-mRNAs. Evidence for this mechanism is provided by the observed antagonistic effects of hnRNPA1 and several SR proteins such as SF2/ASF, SRp40, and SRp55 on splice site selection *in vivo* (Cáceres *et al.*, 1994; Screaton *et al.*, 1995; Wang and Manley, 1995) and *in vitro* (Mayeda and Krainer, 1992). In addition, the expression levels of various SR proteins are variable among tissues (Ayane *et al.*, 1991; Hanamura *et al.*, 1998; Zahler *et al.*, 1993). Another possibility for cell type-specific splicing is the existence of cell- or developmental stage-specific splicing factors modulating splice site selection. The best studied example of this mechanism is the female-specific expression of *Drosophila* transformer (Boggs *et al.*, 1987). There, a functional transformer protein determines sexual fate by directing alternative splicing decisions. Furthermore, tissue-specific factors involved in splice site selection exist, such as a male germ line-specific transformer-2 variant in *Drosophila melanogaster* (Mattox *et al.*, 1990) and *D. virilis* (Chandler *et al.*, 1997), its mammalian isoform htra2-beta3 that is expressed only in some tissues (Nayler *et al.*, 1998a), the muscle-specific form Nop30 (Stoss *et al.*, 1999a), the neuron-specific factor NOVA-1 (Jenson *et al.*, 2000), and a testis- and brain-enriched factor rSLM-2 (Stoss *et al.*, submitted). This multifactorial recognition of alternative splice sites makes their prediction on the basis of sequence data difficult. Currently, prediction of splice sites us-

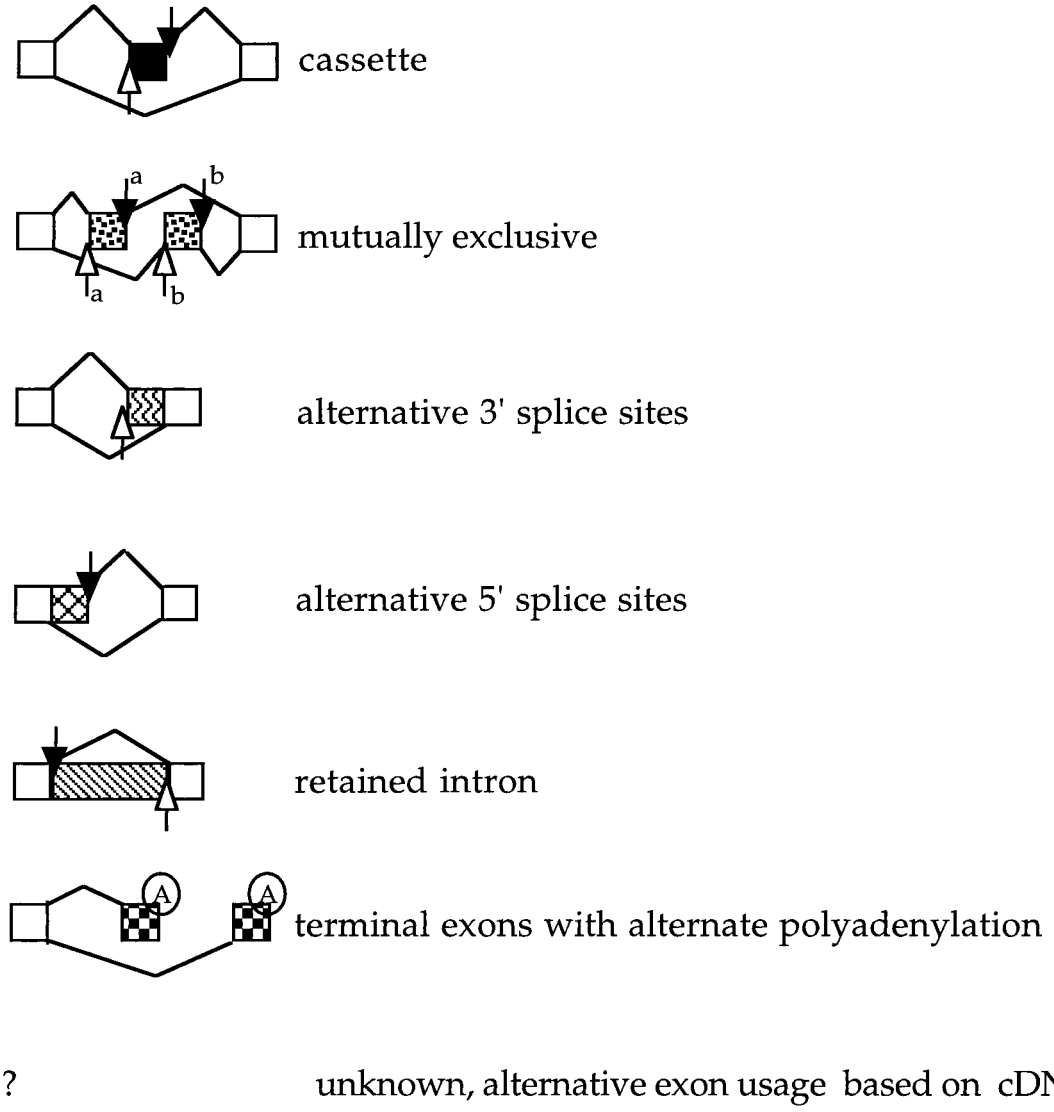


FIG. 1. Alternative splicing modes. Type of alternative exons that were used to classify alternative splicing events. Alternative exons are shown as boxes with different shading. Flanking constitutive exons are shown as white boxes. Open arrows indicate the position of the alternative 3′ splice site; closed arrows indicate the position of the 5′ splice sites analyzed. "a" and "b" in mutually exclusive exons refer to their correlation analysis in Figure 8.

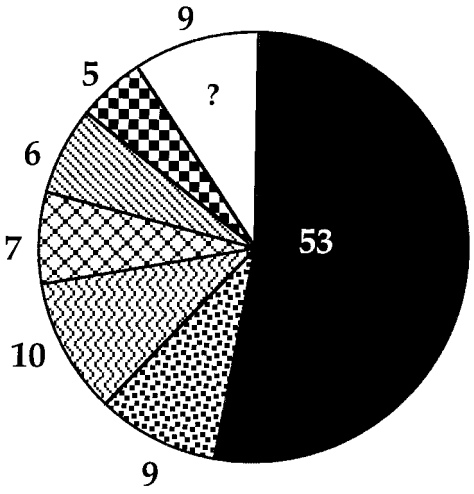


FIG. 2. Distribution of the splicing events in the alternative-exon database. The shading corresponds to the alternative splice modes shown in Figure 1. Numbers indicate percent usage of these splice modes in the dataset. ? = exons with unknown splicing mode.

ing computer programs is still not accurate (Usuka and Brendel, 2000; Thanaraj, 2000).

Most of our knowledge of alternative splicing regulation comes from biochemical studies that identified *trans*-acting factors mainly by using synthetic RNA substrates (Krämer, 1996). This biochemical analysis is limited, as it does not take into account the complexity of natural substrates, the coupling between pre-mRNA processing and transcription (McCracken *et al.*, 1997; Nayler *et al.*, 1998b; Cramer *et al.*, 1999), and the spatial organization of active genes in the nucleus (Schul *et al.*, 1998; Smith *et al.*, 1999). Furthermore, the amount of alternative exons described is larger than the number of alternative splicing events that have been analyzed *in vivo* (Stoss *et al.*, 1999b) and *in vitro.*

As a first step to define naturally occurring *cis* elements regulating pre-mRNA processing, we collected alternative exons from the literature and developed a program that facilitates their analysis. We previously compiled alternative exons that are expressed in neurons (Stamm *et al.*, 1994) and analyzed the statistical features of human constitutive exons (Zhang, 1998). This work suggested that alternative exons differ from constitutive ones by use of certain nucleotide tuples and an unusually high frequency of adenosine at the −3 position of the 3′ splice site. However, this analysis was restricted to mainly one cell type, neurons. To get a broader view, we substantially expanded the compilation of alternative exons from the literature and analyzed their features. The compilation of published data allows connection of the sequence information with biologic features determined experimentally, which distinguishes this dataset from an earlier compilation of exons based on PIR or GenBank data (Dralyuk *et al.*, 2000).

RESULTS

Data collection and data structure

To compile alternative exons and their biologic features, we used "alternative splicing" as a keyword in a MEDLINE search. The following features of alternative exons were collected: species, splicing mechanism, tissue specificity, developmental regulation, regulatory features, and the sequence of the alternatively spliced exon, as well as its flanking constitutive exons. In addition, known regulatory splicing elements were compiled. We did not collect branchpoint sequences, as there is insufficient experimental evidence. Similarly, no data were collected from viruses or plants. The dataset can be viewed at our home page (www.stamms-lab.net/AEDB).

We then developed Java/TCL-based programs to allow the data to be searched, aligned into subgroups, and analyzed by various methods. These methods are calculation of splice site strength (Zhang and Marr, 1994), splice site composition, nucleotide tuple composition, exon length comparison, and a Gibbs algorithm (Neuwald *et al.*, 1995) to find common locally conserved regions that could represent regulatory motifs. In addition, a connected program allows the analysis of newly discovered alternative exons by performing BLAST (Altschul *et al.*, 1990) searches against the database, calculation of splice site scores, and a search function for known exon enhancer motifs.

Type and classification of alternatively spliced exons

Depending on their splicing mode, alternative exons can be classified as cassettes, mutually exclusive, retained intron, and alternative 3′ and 5′ splice sites (Breitbart *et al.*, 1987) (Fig. 1). A comparison of these splicing modes from our dataset reveals that most described alternative exons (53%) are cassette exons (Fig. 2), which is in agreement with an EST-based study, where roughly 61% of the alternative exons were found to be either cassette or mutually exclusive exons (Mironov *et al.*, 1999).

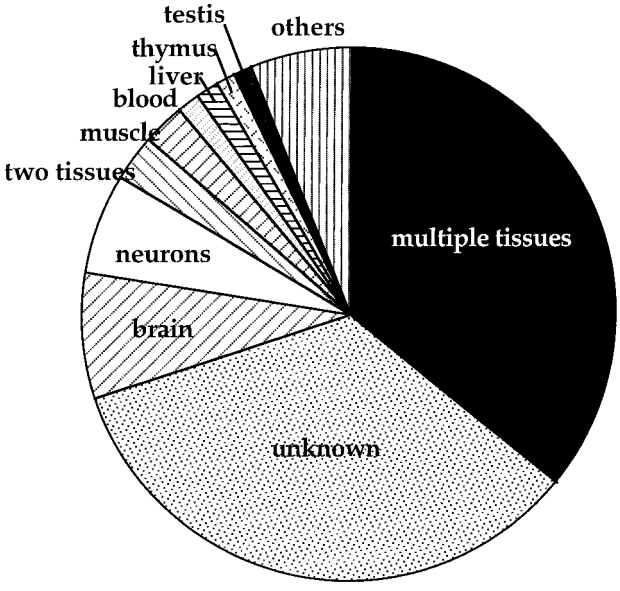
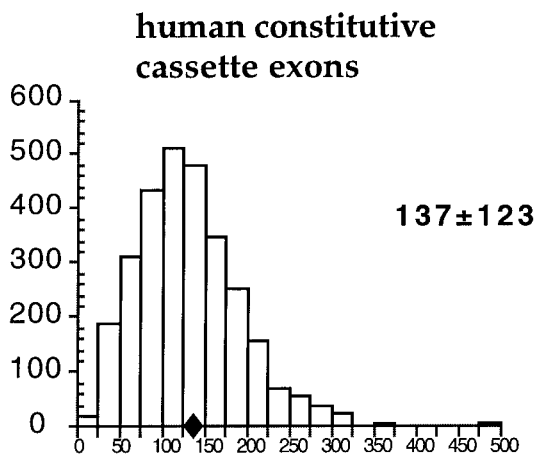
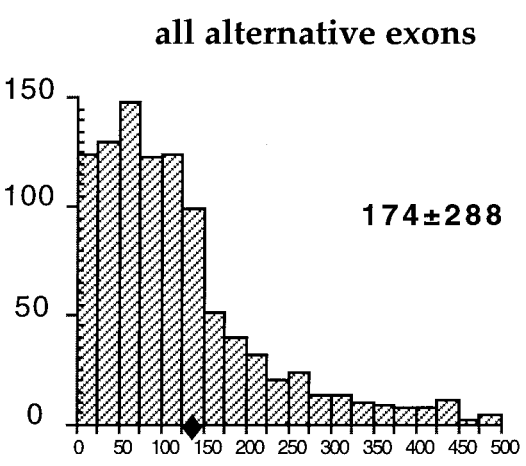
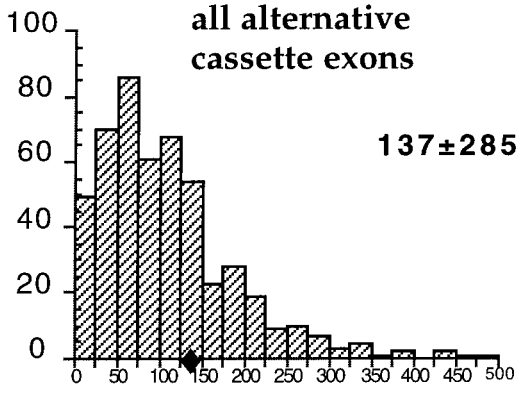
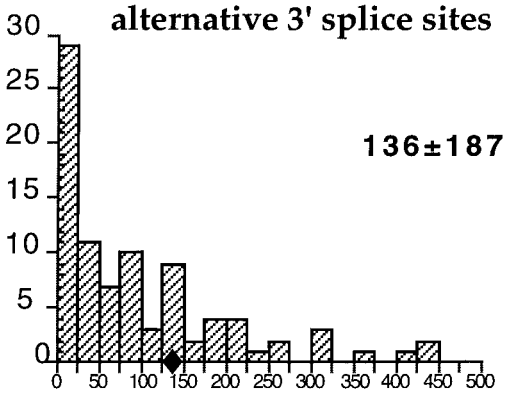


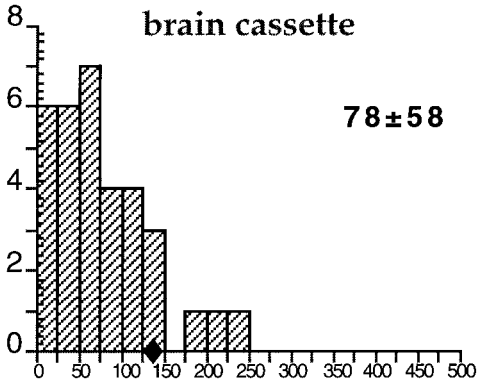
FIG. 3. Tissue distribution of alternative spliced exons. The predominant way of detecting alternative splice variants is PCR. Using this technique, alternatively spliced exons are frequently detected in multiple (at least two) tissues.

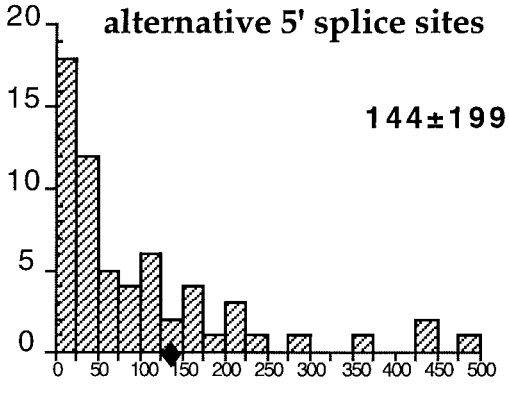

human constitutive cassette exons
137±123
600
500
400
300
200
100
0
0 50 100 150 200 250 300 350 400 450 500

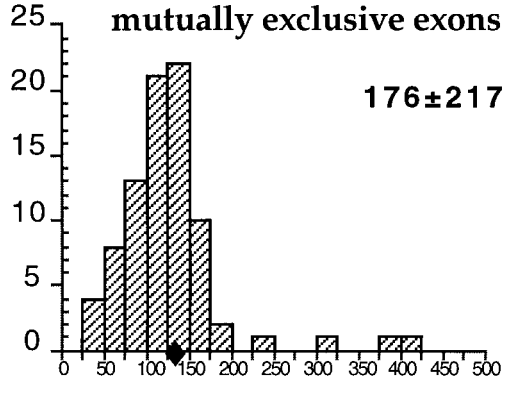

all alternative exons
174±288
150
100
50
0
0 50 100 150 200 250 300 350 400 450 500

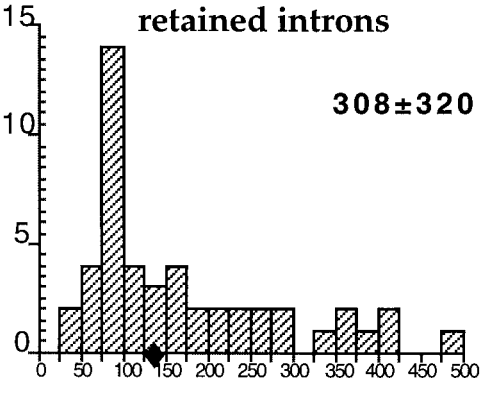

all alternative cassette exons
137±285
100
80
60
40
20
0
0 50 100 150 200 250 300 350 400 450 500


alternative 3' splice sites
136±187
30
25
20
15
10
5
0
0 50 100 150 200 250 300 350 400 450 500


brain cassette
78±58
8
6
4
2
0
0 50 100 150 200 250 300 350 400 450 500


alternative 5' splice sites
144±199
20
15
10
5
0
0 50 100 150 200 250 300 350 400 450 500


mutually exclusive exons
176±217
25
20
15
10
5
0
0 50 100 150 200 250 300 350 400 450 500


retained introns
308±320
15
10
5
0
0 50 100 150 200 250 300 350 400 450 500

We then asked in which tissues alternative exons are known to be expressed (Fig. 3). Most alternative exons can be detected in multiple; that is, more than two, tissues. The most common single tissue and cell type in which expression has been determined experimentally is brain and neurons, followed by muscle, blood, liver, thymus, and testis. Most of the expression analysis is performed by RT-PCR, and it is therefore semiquantitative at best. It remains to be seen whether the high usage of alternative exons in the brain reflects the current focus of research or indicates a high usage of alternative exons in this tissue. This comparison of alternative exon usage in different tissues is an approximation, as not all investigators examined the same tissues. Furthermore, only in rare cases was RNA *in situ* hybridization used to detect alternative exon usage at a single-cell level. However, it is clear from this analysis that expression of an alternative exon is usually not restricted to a certain tissue or cell type, indicating that an intrinsic balance of factors expressed in all tissues (Hanamura *et al.*, 1998) could be responsible for alternative splice site usage for most exons.

Length of alternative exons

Experimental evidence suggests that the length of alternative exons (Black, 1991), as well as the lengths of its flanking introns (Bell *et al.*, 1998), is involved in alternative splicing regulation. We therefore calculated the length of alternative exons that are indicated in Figure 1 as shaded boxes and determined their length distribution. All types of alternative exons have a mean length of 174 ± 288 (SD) nt. For comparison, we used internal human cassette exons (Zhang, 1998) that show a similar length distribution of 137 ± 123 nt (Fig. 4). These exons were also used for other comparisons (see below). The statistical significance of the differences between various alternative exons and the constitutive control exons was determined by performing the Wilcoxon rank test (Conover, 1980). In this test, t and P values, as well as degrees of freedom (df), can be calculated for two samples, x and y, of the size nx and ny, where t = (mean (x) − mean (y) − mean (x-y))/std (x-y) and df = nx + ny − 2 and P of z≥t. We considered P values <0.05 as significant.

We then determined the length distribution of alternative exons according to their splicing mechanism (see Fig. 1). Both a direct comparison of human alternatively spliced cassette exons (123 ± 99; data not shown) and alternative cassette exons from all tissues (137 ± 285) with constitutive human exons revealed a similarity in the mean of the distribution. However, alternative exons are characterized by a smaller mode. Most of these exons are expressed in more than two tissues. Interestingly, when we analyzed cassette exons that are expressed in a single tissue, namely brain, neurons, or muscle, we observed that these exons are shorter on average (78 ± 58, 101 ± 125, and 58 ± 55 nt; Fig. 4 and data not shown) than constitutive exons. These differences are statistically significant; e.g., for brain cassette exons, t = 2.7; df = 2954, and P = 0.0062. Furthermore, alternative exons that are created by a splice site duplication (alternative 3′ and 5′ splice sites) still have means of 136 ± 187 and 144 ± 199 nt, respectively, which is comparable to constitutive exons. The length of those exons is the distance between the upstream and downstream 5′ and 3′ splice sites, respectively (see Fig. 1). Inspection of the length distribution shows that the mode, indicating the highest frequency of exons, is smaller in alternative exons than in constitutive ones. This is apparent with alternative 3′ and 5′ splice sites, for which the mode of the distribution is within 1–25 nt and therefore significantly smaller than the mode for constitutive exons, which is 100–125 nt. We counted the intronic part from exons arising from intron retention (retained introns; Fig. 1) and found that they are on average larger than constitutive cassette exons (308 ± 320 nt) and smaller than constitutive introns, which average around 1127 nt (Hawkins, 1988). However, the mode of their distribution is in the range of 75–110 nt. Common to all distributions is that alternative exon length does not follow a normal distribtion but is skewed toward the smaller exons, which underlines a bias toward smaller exons in alternative splicing.

The decreased size of alternative exons expressed in only a single tissue could indicate the necessity of a certain amount of RNA binding factors to be assembled on constitutive exons prior to their recognition (Fig. 4).

Composition of splice sites

Exons are flanked by splice sites whose nucleotide usage follows a consensus sequence that reflects binding of the U1 and U6 snRNA at the 5′ splice site (Zhuang and Weiner, 1986; Wise 1993) and binding of U2AF to the 3′ splice site (Moore, 2000). We therefore asked whether the nucleotide composition of splice sites flanking alternative exons deviates from the composition of constitutive exons.

3′ Splice sites. First, we calculated the nucleotide frequency at the 3′ splice site for classes of alternative exons and expressed the deviation of this per cent nucleotide usage from the per cent nucleotide usage of constitutive exons (Breitbart *et al.*, 1987). The location of the alternative 3′ splice sites is indicated in Figure 1 (open arrows). In some exons, a deviation from the AG consensus at the −1 and −2 position was reported. As these exons do not follow the consensus for the newly discovered class of AT-AC exons (Tarn and Steitz, 1997), we believe that many of them represent experimental or annotational errors. As can be seen in Figure 5A, when all alternative exons are considered, 3′ splice sites have a more purine-rich polypyrimidine tract. The purine content is strongest in retained introns. In contrast, polypyrimidine tracts of mutually exclusive exons adhere well to the consensus sequence.

We then examined subgroups of cassette exons with a different biologic regulation. On inspection of exons that are ex-

FIG. 4. Length distribution of various types of alternative exons. The length distribution of human constitutive cassette exons is shown as a reference (white columns). The ◆ sign indicates the mean of the distribution in human constitutive cassette exons. Numbers in each histogram indicate mean and SD of the distribution of each class of alternative exons (striped columns). The Y axis of each histogram represents the numbers of exons; the X axis represents the nucleotide sequence length. To allow comparison, only exons up to 500 nt are shown, although in some classes, larger exons exist.

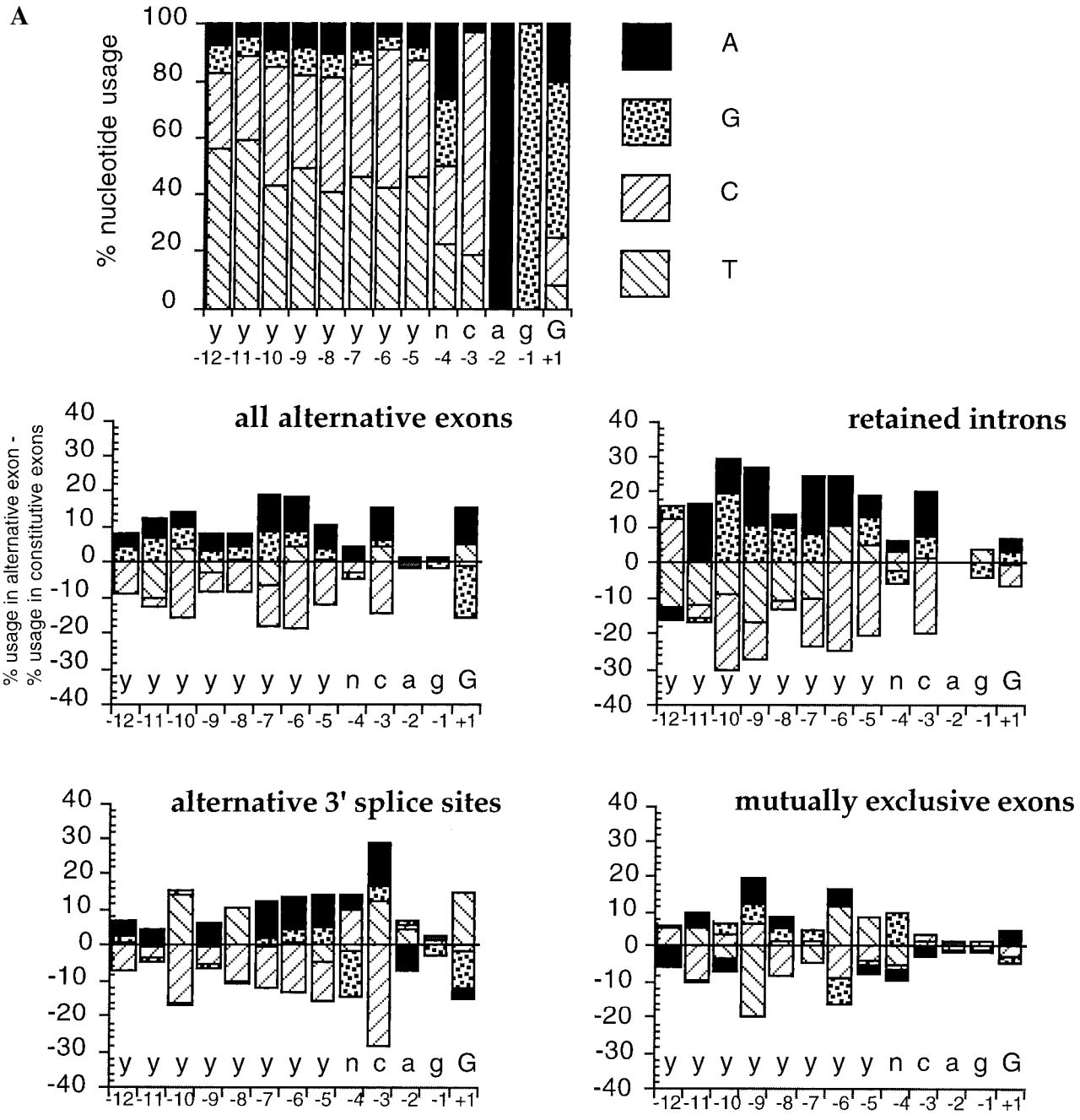


FIG. 5. (*Continued on next page.*) Nucleotide usage at the 3′ splice site. (**A**) Percent nucleotide usage of constitutive exons (top). The 3′ splice site consensus sequence is indicated below the X axis; the corresponding positions are shown underneath. The deviation in nucleotide usage is plotted. The Y axis is (% usage in alternative exons—% usage in constitutive exons). The splicing type is shown for each class of exons. For simplicity, X and Y axes are labeled only in the first graph. (**B**) Deviation of nucleotide usage in subgroups of cassette exons. The cassette exons analyzed are included only in the tissues indicated.

pressed only in brain, neurons, or muscle, as well as all classes that are developmentally regulated, it becomes apparent that these splice sites deviate more from the consensus than do other classes of exons. This deviation appears not to be random, as it is most pronounced at the −3 and −10 position of developmentally regulated exons; at the +1, −3, −9, and −11 position of brain-specific exons; at the +1, −3, and −9 position of neuron-specific exons; and at the +1, −7, and −12 position of muscle-specific exons. In all these classes, the single most divergent nucleotide usage is the presence of an adenosine at the −3 position of the 3′ splice site, (Fig. 5B). Interestingly, exons that are regulated during development show the same deviation at the −3 position. In the Holliday-like structure proposed for the spliceosome (Steitz, 1992), the

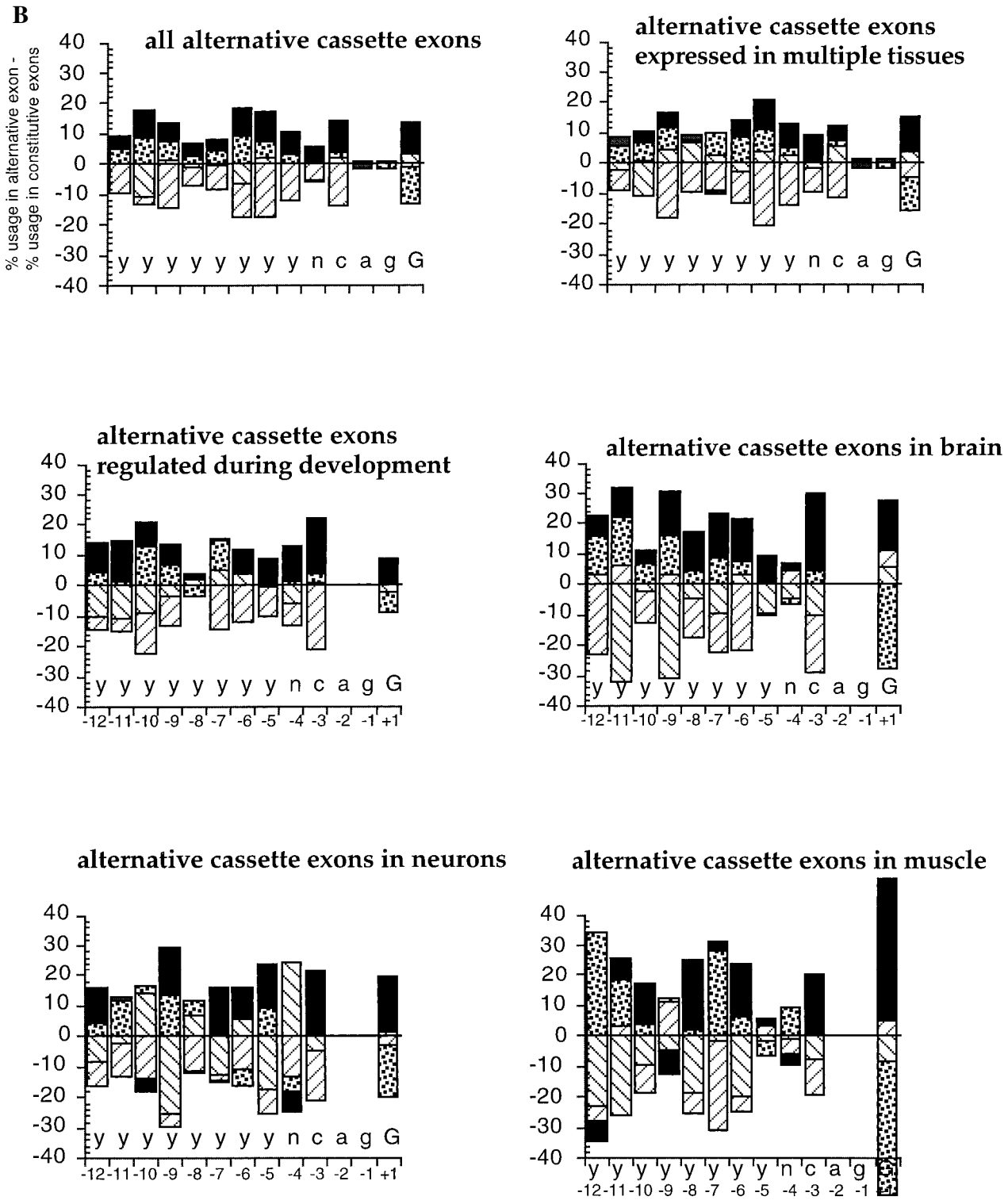


FIG. 5. (*Continued*)

nucleotide at the −3 position would base pair with the G9 of U1 snRNA, which would favor a C at this position. Mutations of this nucleotide at the −3 position influence exon usage (Epstein *et al.*, 1994), and our statistical data suggest that an A at this position could be involved in tissue-specific regulation. Interestingly, the SELEX consensus sequence of U2AF (Wu *et al.*, 1999) has a T at this position and does not reflect the mammalian splice site consensus. Furthermore, the reduced pyrimidine content of the polypyrimidine tract of alternative exons will most likely reduce its affinity for U2AF, which was

shown to bind to the consensus TTTYYYYTNTAGG (Wu *et al.*, 1999).

5′ Splice sites. Next, we asked how the 5′ splice site nucleotide composition of alternative exons differs from the consensus sequence and calculated the difference in nucleotide frequency at each position (Fig. 6). The location of the alternative 5′ splice sites is indicated in Figure 1 (closed arrows). The nucleotide composition at the 5′ splice site reflects binding to U1 snRNA (Zhuang and Weiner, 1986) as well as an interaction with U6 snRNA (Wise, 1993). Overall, alternative exons deviate most strongly at the +4 and +5 positions, which is the case for exons from all grups analyzed. Interestingly, in exons that are expressed exclusively in neurons and muscle cells, there is an additional deviation at the −3 position, where splice sites from both categories use more adenosine than the consensus.

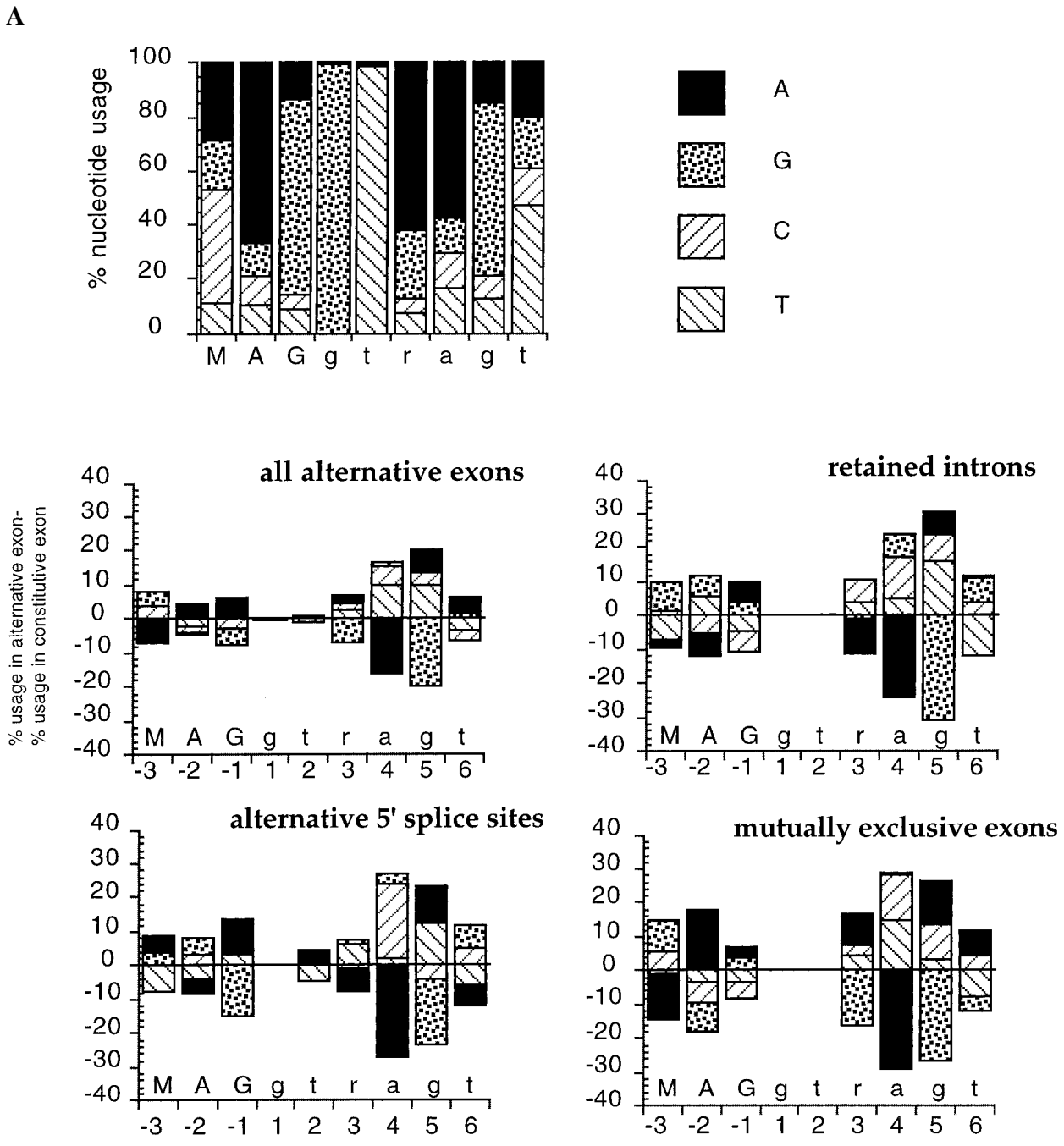


FIG. 6. (*Continued on next page.*) Nucleotide usage at the 5′ splice site. (**A**) Percent nucleotide usage of constitutive exons (top). The 5′ splice site consensus sequence is indicated at the bottom. Below, the deviation in nucleotide usage is plotted. The Y axis is (% usage in alternative exons—% usage in constitutive exons). The splicing type is indicated. M = C or A; R = A or G. (**B**) Deviation of nucleotide usage in subgroups of cassette exons.

As in the −3 position of the 3′ splice site, this would introduce a mismatch to G11 of U1 snRNA (Steitz, 1992). Furthermore, the nucleotide at the +4 position is a U in the yeast consensus sequence and was postulated to bind to A49 of U6 snRNA in a late step of splicing (Sontheimer and Steitz, 1993). In mammalian systems, this A49 of U6 snRNA was shown to be a N6-methyl adenosine (Reddy, 1989). Furthermore, the corresponding bases U5 and U6 of U1 snRNA that would bind to the +4 and +3 positions are modified to pseudouracils (Reddy, 1989). It is interesting that alternative exons deviate from the consensus at positions of the 5′ splice site that are modified in the complementary base position of U1 and U6 snRNA.

B

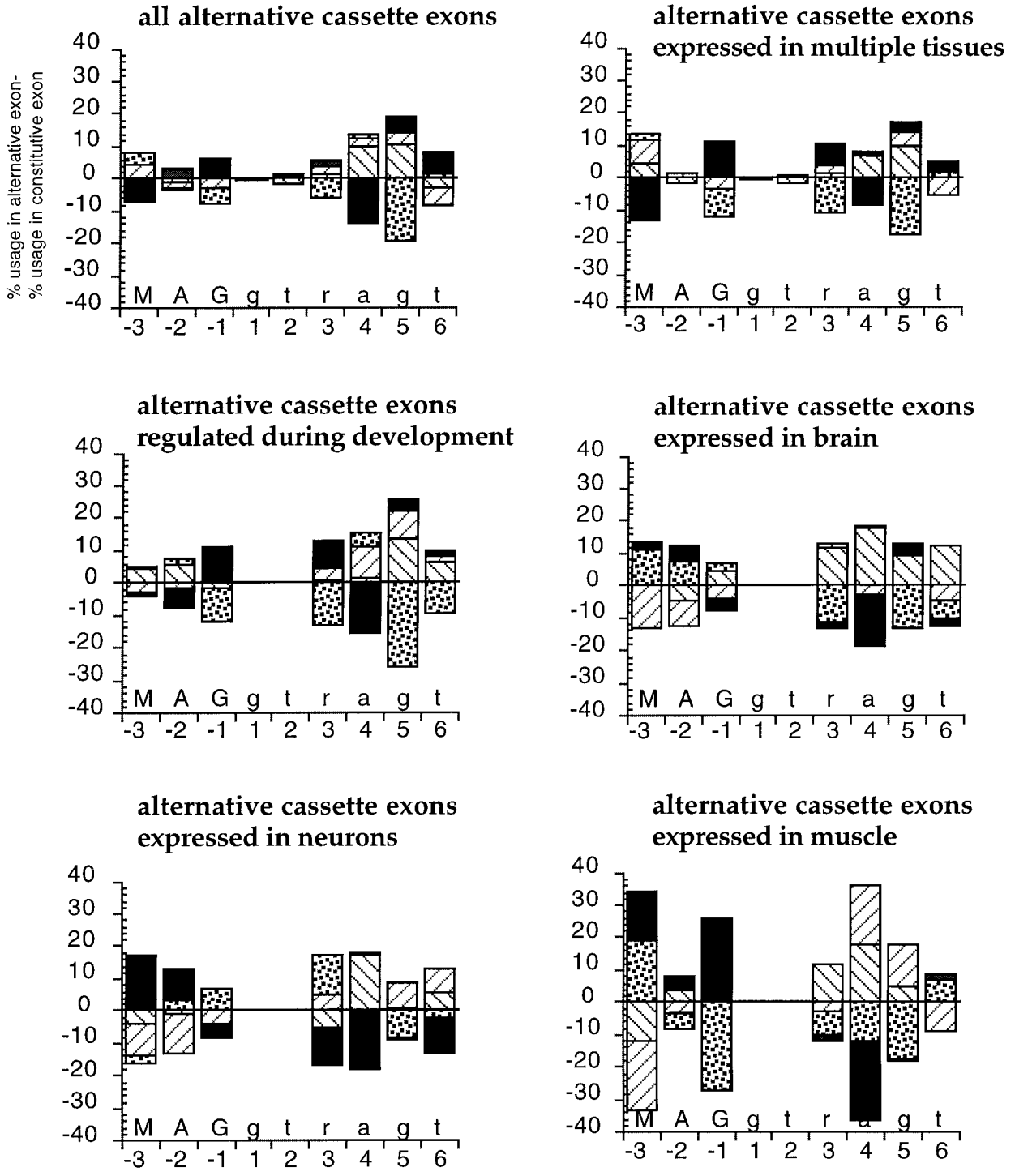


FIG. 6. (*Continued*)

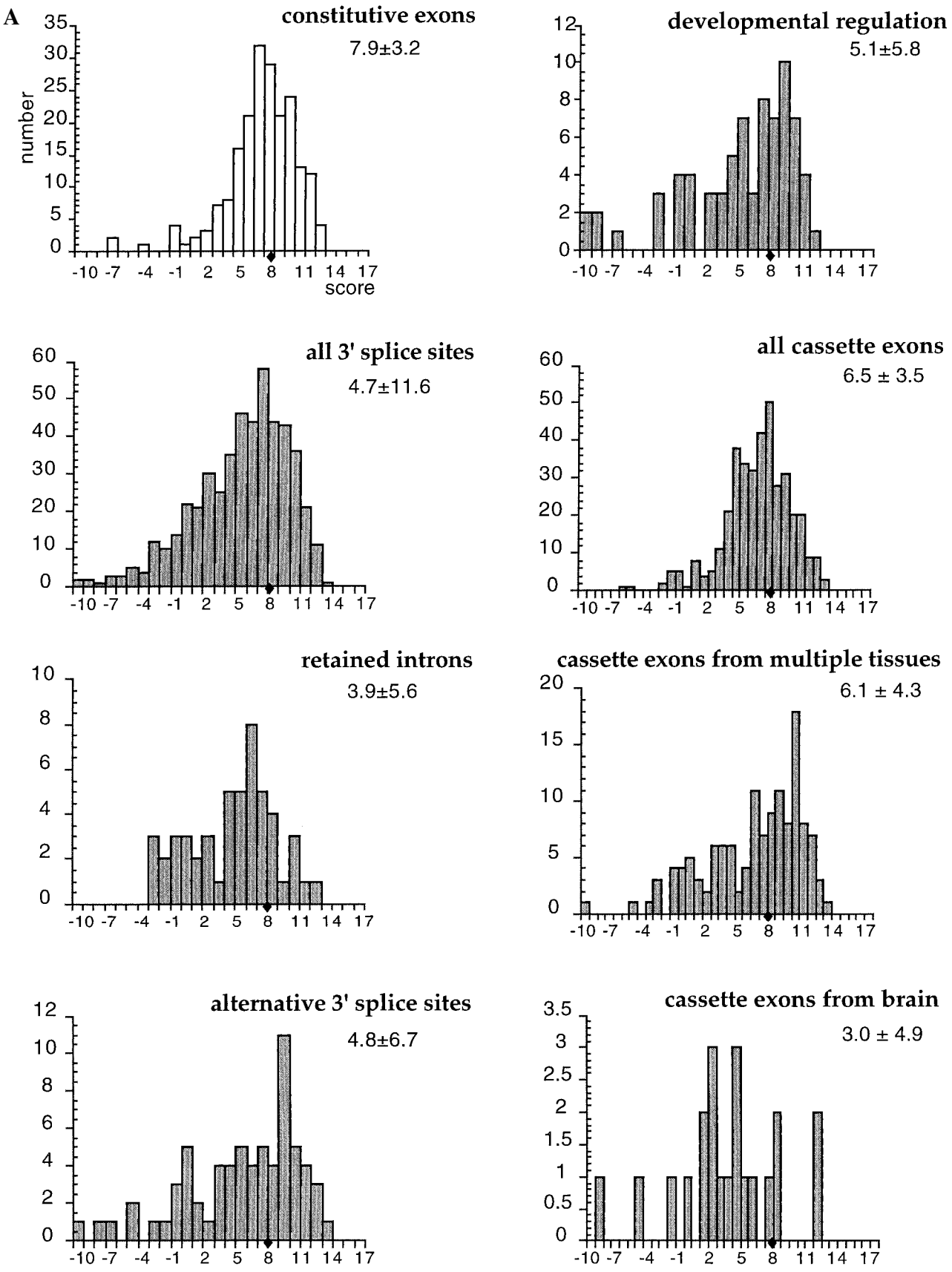


FIG. 7. (*Continued on the next 3 pages.*)

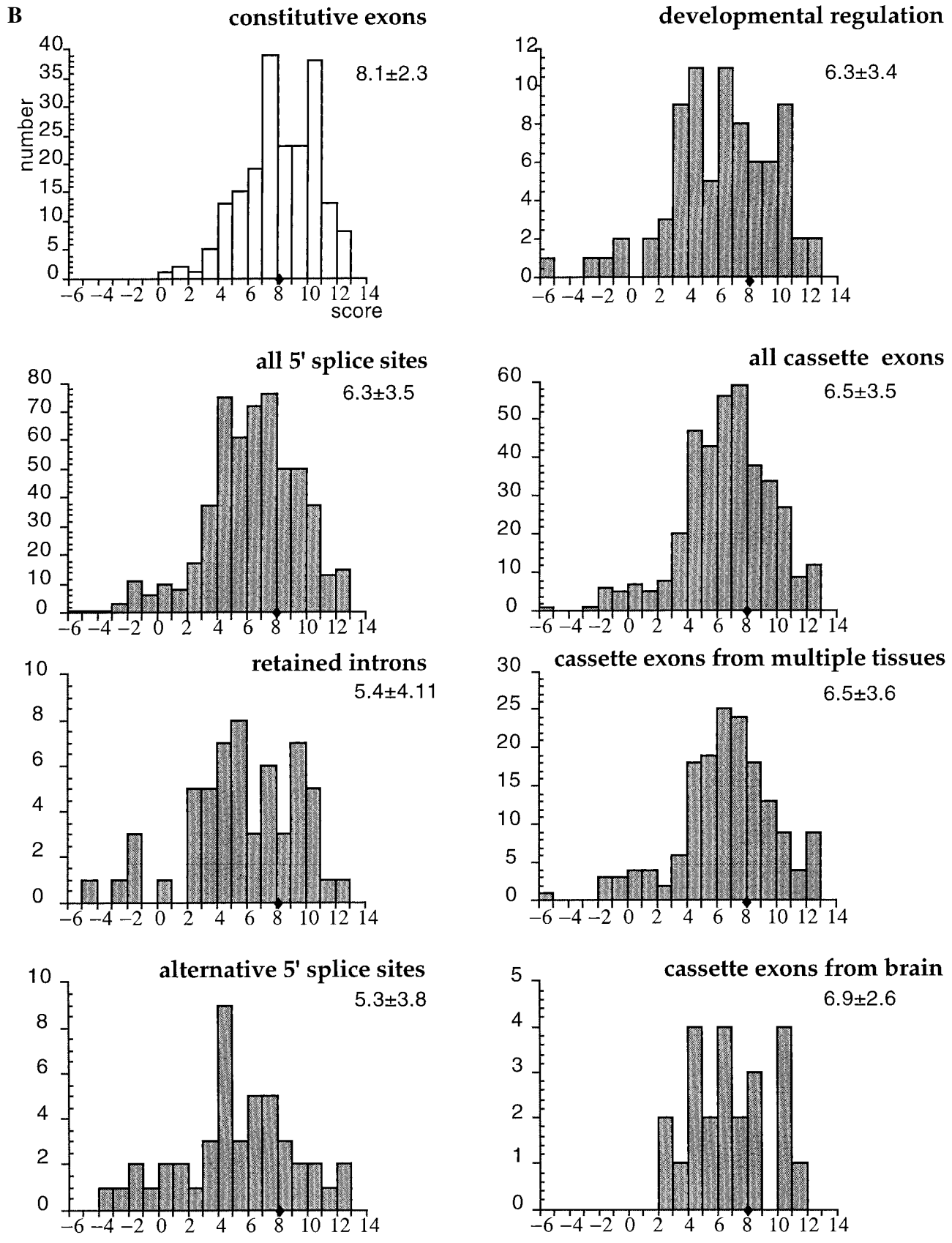


FIG. 7. (*Continued*)

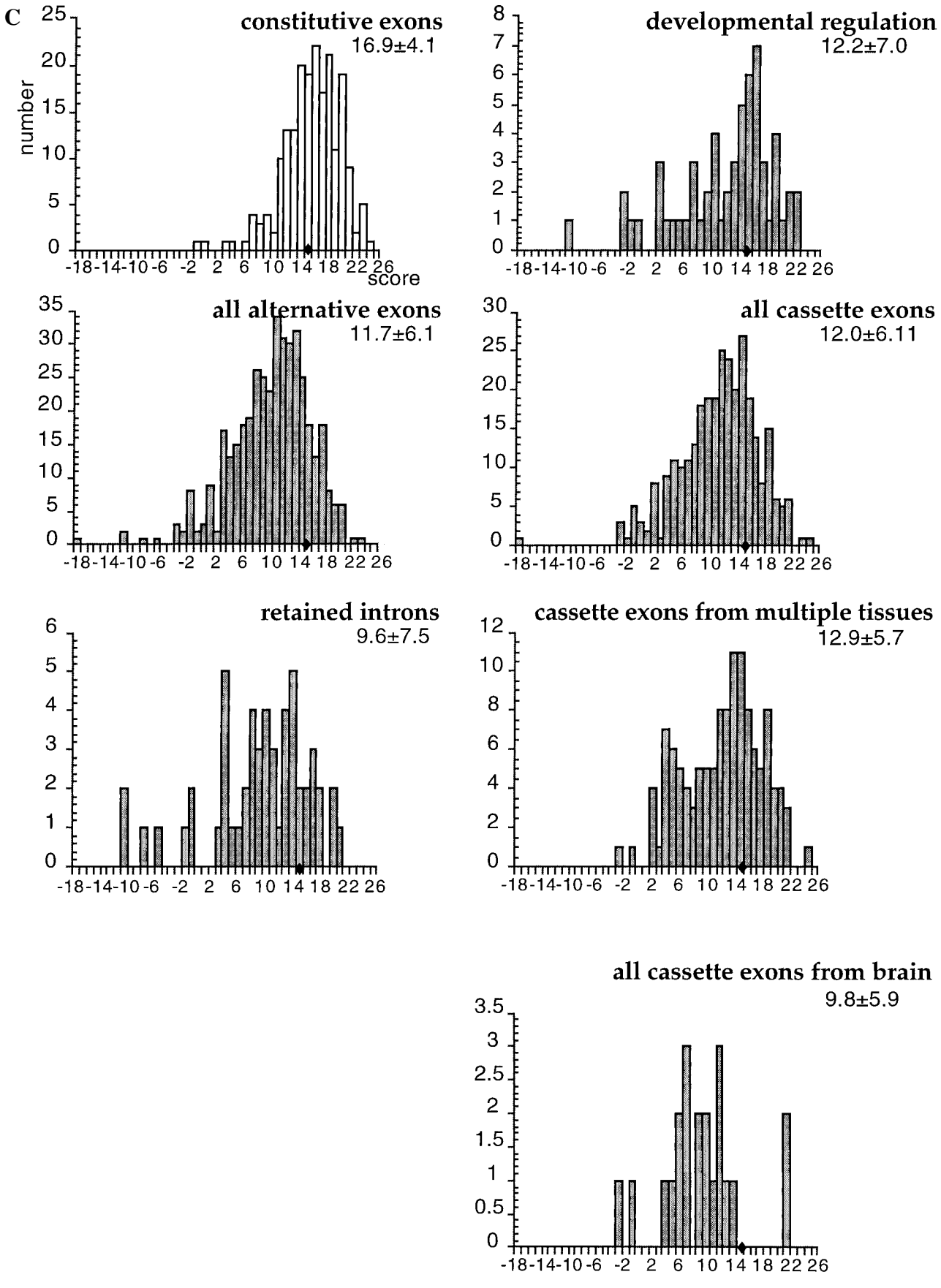


FIG. 7. (*Continued*)

Splice site scores

In order to assess the quality of individual splice sites, we calculated LOG-ODD scores for each site. The LOG-ODD scores express the coincidence of a splice site with the consensus sequence: a higher coincidence generates a higher score.

The LOG-ODD scores were calculated according to Zhang and Marr (1994) using the GeneId dataset and randomly extracted splice sites for comparison. The scoring function is defined as:

$$S_i(X) = \log_2 \frac{P_i(X)}{Q(X)}$$

Where $P_i(X)$ is the frequency of finding X at position i that is equal to $C_i(X)/D_i$, the normalization D_i is the sum of the counts $C_i(X)$ over X (= A, C, G, T), and $Q(X) = {}^1/_4$ was used for all X as the random background frequency. The score for a splice site is the sum of the scores for each individual nucleotide.

We first calculated the scores for individual 5′ and 3′ alternative splice sites of several groups of exons. Because it has been postulated that the splice sites surrounding an exon define its borders in a concerted way (Robberson *et al*., 1990; Berget, 1995), we then calculated the LOG-ODD scores for the splice sites surrounding different classes of exons.

Compared with the control splice sites, the distribution of alternative 5′, 3′, and combined splice sites is broader, indicated by a higher standard deviation, and the mean of the distribution is at a lower score (Fig. 7). A statistical evaluation using the Wilcoxon rank test (Conover, 1980) demonstrated that these differences are significant.

Inspection of individual classes of exons demonstrated that on average, the weakest 3′ splice sites are found in retained introns and alternative 3′ splice sites (Fig. 7A). Similarly, the weakest individual 5′ splice sites are found in retained introns and exons with an alternative 5′ splice site (Fig. 7B). These properties are reflected in the scores of combined splice sites, where a subgroup of retained introns is characterized by splice sites that deviate considerably from the consensus (Fig. 7C). Of the tissue-specific exons analyzed, exons from brain have the weakest 3′ and combined splice sites (Fig. 7). Suboptimal splice site scores are also found in exons expressed only in neurons and muscle, which show a combined score distribution of 13.2 ± 3.8 and 10.4 ± 4.2, respectively, (data not shown). Furthermore, a large subgroup of exons that are known to be regulated during development are characterized by suboptimal splice sites.

This analysis indicates that a substantial proportion of alternative exons are characterized by splice sites with strong deviations from the consensus. For recognition of these exons, specific factors might be needed.

Correlation between pairs of mutually exclusive exons

The regulation of alternative splicing in mutually exclusive exons can be viewed as the result of a competition between two pairs of splice sites. We therefore asked whether there is a correlation between the splice site strengths, expressed as its score, in pairs of mutually exclusive exons. We plotted the 3′, 5′, and combined scores of each pair and found a correlation between $r = 0.5$ and $r = 0.69$. The 5′ splice sites show a slightly higher correlation, $r = 0.64$, than 3′ splice sites, $r = 0.5$. On average, the difference between the splice site scores of two mutually exclusive exons is 1.5. (Fig. 8A–C). Next, we asked whether there is a relation between the length of mutually exclusive exons. We found a strong correlation between the length of individual pairs (Fig. 8D). In most cases, the exons are identical in length.

Mutually exclusive exons have been studied in several model systems (Pret *et al*., 1999; Selvakumar and Helfman, 1999; Southby *et al*., 1999), and the spacing of regulatory elements was found to be important. It remains to be determined whether the strong correlation of the length of mutually exclusive exons is caused by a mechanistic or a functional selective pressure.

FIG. 7. Splice site scores of alternatively spliced exons. The scores are plotted on the X axis. The higher the score, the better the match with the consensus sequence. The number of splice sites that correspond to a certain score is plotted on the Y axis. Constitutive exons (top) are in white; alternative exons are in gray. The ◆ sign indicates the mean of the distribution in human cassette exons. (**A**) Scores of 3′ splice sites. The splicing mode is indicated on the top right of each histogram; the number below indicates the mean and SD of the distribution. A "perfect" 3′ splice site, $(u)_{11}$cagG, would have a score of 15.6. The mean of the distribution is 7.9 for constitutively spliced exons. The deviation in each subclass is statistically significant. Parameters of the Wilcoxon rank test are: developmental regulation: t = 4.99, df = 273, $P = 0$; all 3′ splice sites: t = 4.9265, df = 268, $P = 0$; all cassette exons: t = 4.6256, df = 580, $P = 0$; retained introns: t = 6.616, df = 251, $P = 0$; cassette exons from multiple tissues: t = 4.2241, df = 329, $P = 0$; alternative 3′ splice sites: t = 4.9265, df = 268, $P = 0$; cassette exons from brain: t = 2.2964, df = 219, $P = 0.0226$. (**B**) Scores of 5′ splice sites. A perfect match to U1 snRNA would have a score of 12.2. The mean of the distribution is 8.1 for constitutively spliced exons. The deviation in each subclass is statistically significant. Parameters of the Wilcoxon rank test are: developmental regulation: t = 5.154, df = 277, $P = 0$; all 5′ splice sites: t = 6.9511, df = 746, $P = 0$; all cassette exons: t = 5.9315, df = 579, $P = 0$; retained introns: t = 6.2722, df = 256, $P = 0$; cassette exons from multiple tissues: t = 5.0759, df = 361, $P = 0$; alternative 5′ splice sites: t = 6.4235, df = 243, $P = 0$; cassette exons from brain: t = 0.8635, df = 223, $P = 0.3888$. (**C**) Distribution of combined 5′ and 3′ splice sites flanking a single exon or retained intron. An exon surrounded by a "perfect" 3′ and 5′ splice site would have a score of 27.8. The mean of the distribution is 16.9 for constitutively spliced exons. The deviation in each subclass is statistically significant. Parameters of the Wilcoxon rank test are: developmental regulation: t = 5.166, df = 257, $P = 0$; all alternative exons t = 9.0193, df = 612, $P = 0$; all cassette exons: t = 8.0743, df = 504, $P = 0$; retained introns: t = 8.2561, df = 251, $P = 0$; cassette exons from multiple tissues: t = 5.6352, df = 322, $P = 0$; alternative 5′ splice sites: t = 6.4235, df = 243, $P = 0$; cassette exons from brain: t = 6.2901, df = 219, $P = 0$.

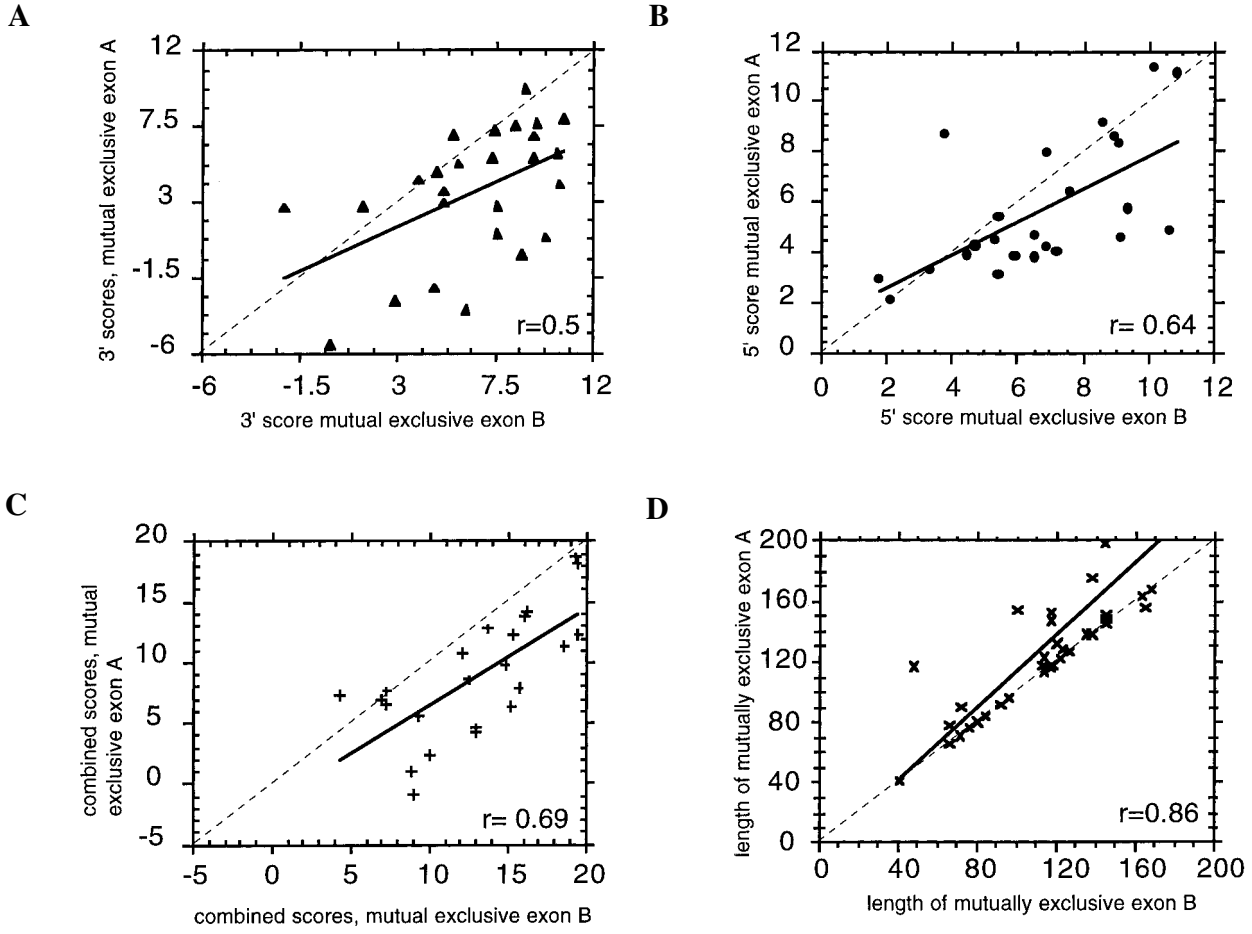


FIG. 8. Analysis of properties of mutual exclusive exons. Splice site scores and length of pairs of mutually exclusive exons (A and B) were determined, and the values of individual pairs were compared. The r number in the bottom right corner of each graph indicates the correlation coefficient. The thick line shows the linear regression from the data. The thin dotted line indicates a correlation of $r = 1$. (**A**) The 3′ scores of splice sites located on two mutually exclusive exons. (**B**) The 5′ scores of splice sites located on two mutually exclusive exons. (**C**) Combined scores of splice sites located on two mutually exclusive exons. (**D**) Length in nucleotides of two mutually exclusive exons.

Sequence elements present on alternative and constitutive exons

Recent experiments have revealed the existence of regulatory elements known as exonic sequence enhancers (ESE) that are present on constitutive and alternatively spliced exons (Cooper and Mattox, 1997; Tacke and Manley, 1999). These sequences function in constitutive and alternative splicing by recruiting RNA-binding proteins; e.g., SR proteins (Manley and Tacke, 1996) and hnRNPs (Weighardt *et al.*, 1996), to the vicinity of splice sites and act at early (Staknis and Reed, 1994; Selvakumar and Helfman, 1999) and late (Chew *et al.*, 1999) steps of splicing. It is currently debated whether tissue- or developmental stage-specific alternative splicing is the result of cell type-specific concentration differences or is attributable to cell type-specific factors. In these models, either a specific combination of factors or a specific factor would bind to a regulated alternative exon.

In order to investigate the possibility of common RNA motifs, we applied the Gibbs algorithm to subclasses of exons (Fig. 9). This algorithm allows the determination of redundant sequence motifs that are least likely to occur by chance in a given nucleotide composition (see Neuwald *et al.*, 1995 for the technical details). First, we analyzed exons with suboptimal splice sites. Those exons were defined as having a combined exon score <10. In this group, we were able to identify the motif GNCNTCCAA as being highly abundant when we applied the Gibbs sampling method with the maximum *a posteriori* probability <0.02 (Neuwald *et al.*, 1995). We then employed this algorithm to look for common motifs in subgroups of alternative exons that were expressed only in neurons, brain, blood, liver, muscle, testis, and thymus (Fig. 9). Only short motifs, from 4 to 9 nt, could be found. Common to all motifs is a high degree of degeneracy, which has also been found when SELEX procedures were employed *in vivo* or *in vitro* (Tacke and Manley, 1999). Furthermore, the occurrence of a given motif was always restricted to certain exons in a given subclass. When all motifs are considered together, the nucleotide composition is A = 22%, G = 28%, C = 33%, and T = 17%, indicating that overall, these motifs are not purine rich. Inspection of the 24

Exons with splice site score < 10

motif	incidence	occurence	#
GNCNTCCAA	(53/100/100/100/100/100/80/60/60)	(10 of 52)	1

Neurons

CCYCCA	(70/94/76/89/94/94/74)	(11 of 48)	2
GRGGAGG	(93/95/93/80/80/93/93)	(21 of 48)	3
(C/G)CNGNCNNCCC	(87/94/n/93/94/n/n/87/94/87)	(7 of 48)	4
TCCCAG(G/C)	(81/88/94/94/87/93/79)	(9 of 48)	5
CCTCCAG	(81/81/78/81/81/79/80)	(5 of 48)	6
TCCCAGN(C/G)	(90/91/91/91/91/90/91/n/94)	(5 of 48)	7

Brain

AAA(A/T)AAA	(93/86/93/95/93/86/79)	(5 of 62)	8
AGCCCCA	(53/93/93/93/83/93/93)	(9 of 62)	9
CNCTCCTC	(93/77/85/93/93/93/60)	(11 of 62)	10
GAGRG(T/A)G	(93/93/87/90/51/96/93)	(12 of 62)	11
CCARYCC	(93/93/85/87/96/93/69)	(11 of 62)	12
CCACYCC	(93/93/93/93/95/93/93)	(10 of 62)	13

Blood

TGGT(G/T)(A/T)C	(83/73/83/83/95/96/93)	(9 of 16)	14
(A/C)AGAGRA	(96/93/93/93/68/95/60)	(8 of 16)	15

Liver

TGA(C/G)TGG	(94/80/80/95/94/93/93)	(7 of 14)	16
TTTTATT	(94/84/94/94/53/78/94)	(6 of 14)	17
CCAG(ACG)Y	(85/93/86/93/98/70/96)	(5 of 14)	18

Muscle

CT(C/G)NNCC(A/T)G	(88/41/97/NN/93/88/95/77)	(7 OF 19)	19
GAGGRAG	(94/87/82/71/91/76/94)	(3 OF 19)	20

TESTIS

R(G/T)GACTG	(94/95/70/93/94/93/82)	(5/11)	21
TGCAGCC	(88/77/94/88/93/88/88)	(4/11)	22

THYMUS

AGRAGYY	(80/93/95/93/67/95/85)	(7 of 12)	23
TCA(A/T)GTT	(93/70/71/96/93/93/93)	(5 of 12)	24

FIG. 9. Nucleotide motifs found in exons from different tissues. The motifs were identified using a Gibbs algorithm on the subset of exons expressed specifically in the tissues indicated. The incidence is the percent probability of finding a nucleotide at a given position. The occurence shows how many exons of this subgroup contain the exon.

motifs showed only four GAR-like sequences (3, 11, 15, and 20) that have been described as exon enhancers (Watakabe *et al.*, 1993; Tanaka *et al.*, 1994). Because AAC and ACC triplets are missing in the motif list, triplet CCA could be indicative of A/C-rich enhancer (ACE) motifs (Coulter *et al.*, 1997) in these data and is present nine times (1, 2, 5, 6, 7, 9, 12, 13, and 18). Together, these data show that in addition to the previously described GAR and ACE sequences, other motifs exist in alternatively spliced exons. Often, these motifs are pyrimidine rich. Interestingly, such pyrimidine-rich motifs have been identified

as activating motifs in *in vitro* SELEX procedures (Tian and Kole, 1995). However, the *trans*-acting factors binding to these motifs remain to be determined.

DISCUSSION

We have created a database of alternatively spliced exons that is based on sequences published in the literature. This data acquisition allowed us to introduce biologic features, such as splicing mode, tissue specificity, and regulatory functions, that are usually not present in GenBank entries.

Currently, the database consists of approximately 2×10^5 nucleotides in about 1000 exons. Because we limited ourselves to experimentally characterized exons, this size is smaller than the actual number of alternatively spliced genes, estimated to approach one of three human genes (Croft *et al.*, 1999; Mironov *et al.*, 1999), and is also smaller than a recent GenBank-generated database (Dralyuk *et al.*, 2000). However, we were able to make statistically relevant conclusions after subgrouping exons on the basis of their biologic properties. The regulation of alternatively spliced genes is most likely achieved by a combination of intrinsically weak interactions that achieve specificity by a presumably cooperative interaction of multiple partners (Krämer, 1996; Varani and Nagai, 1998). Not surprisingly, therefore, our analysis did not reveal strong sequences or motifs that regulate a single type of alternative splicing. However, several features can be delineated. Cassette exons are the most common form of alternatively spliced exons accounting for more than half of the entries analyzed. Compared with constitutive cassette exons, their 3′ splice sites contain more purine nucleotides, and their 5′ splice sites deviate mostly at the +4 and +5 positions. The next common alternative splicing form is alternative 3′ splice sites. These exons are characterized by their short length and a high usage of A and T at the −3 position of the alternative (downstream) 3′ splice site. This deviation is also reflected in the average splice site score, which is about 5 units below the average for constitutive 3′ splice sites. Mutually exclusive exons are slightly less frequent and are marked by a strong correlation of the length and splice site strength within a pair of mutually exclusive exons. Seven per cent of the exons analyzed were generated by alternative 5′ splice site usage. Similar to alternative 3′ exons, these exons are short. On average, their alternative splice site strength is about three score points below that of the control exons and deviates most at the +4 and +5 position. The least frequent mode of splicing are retained introns. On average, they are longer than constitutive exons but significantly shorter than constitutive introns. Their combined splice site scores are the weakest ones observed, about six units below the constitutive controls. This class of exons has the highest purine content in their 3′ splice sites, giving rise to the strongest deviation from the U2AF consensus sequence. Weak splice sites have been pointed out by numerous studies, and our analysis confirms this feature for some, but not all, alternative exons.

We then looked for common features of exons that have similar biologic regulation, such as a restricted expression pattern or developmental regulation. Such exons show the strongest deviation from the norm: they are shorter and have weaker splice sites than other alternative exons. Common to tissue-specific splice sties are deviations at the −3 position of the 3′ splice site and at the +4 and +5 positions of the 5′ splice site. A deviation of alternative splice sites from the consensus sequences has been observed by numerous investigators. Our quantitative analysis shows that this deviation is concentrated on certain nucleotides and is more pronunced when a strict tissue-specific regulation is observed.

We then looked for putative regulatory elements in this group of exons and, using a Gibbs algorithm, identified several motifs that are used more often in tissue-specific exons. Some of these motifs are reminiscent of ACE- and GAR-type splicing enhancers, but about half of the motifs found have not been previously described. Common to all these motifs is a high degree of degeneracy relative to a consensus sequence. Most likely, this degeneracy allows exon enhancer sequences to be present regardless of the codon usage dicated by the amino acid composition of the protein.

It is interesting to note that GAR-rich enhancer sequences are not found more often in alternative exons than in constitutive exons. For example, searching with the motif GARGAR in our database revealed that it is found in 22.1% of alternative exons but also in 22.9% and 17.6% of the 5′ and 3′ constitutive flanking exons. This result could indicate that purine-rich enhancers of the GAR type do not have a special role in alternative splicing or that the purine content over a short stretch of nucleotides is more important than a strict sequence *in vivo*. Furthermore, at least one of the motifs identified by us (Fig. 9, No. 8) was shown to be involved in a disease. A change of AATAA into AAGAA in exon 10 of neurofilament tau can cause aberrant splicing of this exon by disrupting a putative exon enhancer, resulting in frontotemporal dementia (Gao *et al.*, 2000). The functionality of the other motifs identified remains to be determined. Together, our data show that alternative splice site selection is governed by a delicate balance of weakly conserved RNA sequence elements. It remains the challenge to determine their binding specificities for *trans*-acting factors and to combine multiple binding equilibria between RNA and proteins, as well as protein–protein interactions, which would ultimately allow the accurate prediction of alternative splice site usage.

ACKNOWLEDGMENTS

This work was supported by the HSFP (RG562/96 to S.S. and K.N.), the Max-Planck Society, a Grant-in-Aid (Genome Science) for Scientific Research on Priority Areas of MESSC in Japan (K.N.), and in part by a National Institutes of Health (R01HG01696) to M.Q.Z. We thank Adrian R. Krainer for critical discussions.

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.

AYANE, M., PREUSS, U., KÖHLER, G., and NIELSEN, P.J. (1991).

A differentially expressed murine RNA encoding a protein with similarities to two types of nucleic acid binding motifs. Nucleic Acids Res. **19,** 1273–1278.

BELL, M.Y., COWPER, A.E., LEFRANC, M.-P., BELL, J.I., and SCREATON, G.R. (1998). Influence of intron length on alternative splicing of CD44. Mol. Cell. Biol. **18,** 5930–5941.

BERGET, S.M. (1995). Exon recognition in vertebrate splicing. J. Biol. Chem. **270,** 2411–2414.

BLACK, D.L. (1991). Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non neuronal cells? Genes Dev. **5,** 389–402.

BLACK, D.L. (1995). Finding splice sites within a wilderness of RNA. RNA **1,** 763–771.

BOGGS, R.T., GREGOR, P., IDRISS, S., BELOTE, J.M., and MCKEOWN, M. (1987). Regulation of sexual differentiation in *D. melanogaster* via alternative splicing of RNA from the transformer gene. Cell **50,** 739–747.

BREITBART, R.E., ANDREADIS, A., and NADAL-GINARD, B. (1987). Alternative splicing: A ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annu. Rev. Biochem. **56,** 467–495.

CÁCERES, J., STAMM, S., HELFMAN, D.M., and KRAINER, A.R. (1994). Regulation of alternative splicing in vivo by overexpression of antogonistic splicing factors. Science **265,** 1706–1709.

CHANDLER, D., MCGUFFIN, M.E., PISKUR, J., YAO, J., BAKER, B., and MATTOX, W. (1997). Evolutionary conservation of regulatory strategies for the sex determination factor transformer-2. Mol. Cell. Biol. **17,** 2908–2919.

CHEW, S.L., LUI, H.-X., MAYEDA, A., and KRAINER, A.R. (1999). Evidence for the function of an exonic splicing enhancer after the first catalytic step of pre-mRNA splicing. Proc. Natl. Acad. Sci. **96,** 10655–10660.

COOPER, T.A., and MATTOX, W. (1997). The regulation of splice-site selection, and its role in human disease. Am. J. Hum. Genet. **61,** 259–266.

CONOVER, W.J. (1980). *Practical Nonparametric Statistics*, ed 2. (Wiley, New York).

COULTER, L.R., LANDREE, M.A., and COOPER, T.A. (1997). Identification of a new class of exonic splicing enhancers by in vivo selection [published erratum appears in Mol. Cell. Biol. 1997;17(6): 3468]. Mol. Cell. Biol. 17:2143–2150.

CRAMER, P., CACERES, J.F., CAZALLA, D., KADENER, S., MURO, A.F., BARALLE, F.E., and KORNBLIHTT, A.R. (1999). Coupling of transcription with alternative splicing: RNA polII promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. Mol. Cell **4,** 251–258.

CROFT, L., SCHANDORFF, S., CLARK, F., BURRAGE, K., ARCTANDER, P., and MATTICK, J.S. (1999). ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nature Genet. **24,** 340–341.

DRALYUK, I., BRUDNO, M., GELFAND, M.S., ZORN, M., and DUBCHAK, I. (2000). ASDB: Database of alternatively spliced genes. Nucleic Acids Res. **28,** 296–297.

DU, L., and WARREN, L. (1996). A functional interaction between the carboxy-terminal domain of RNA polymerase II and pre-mRNA splicing. J. Cell Biol. **136,** 5–18.

ELLIOT, D.J. (2000). Splicing and the single cell. Histol. Histopathol. **15,** 239–249.

EPSTEIN, J.A., GLASER, T., CAI, J., JEPEAL, L., WALTON, D.S., and MAAS, R.L. (1994). Two independent and interactive DNA-binding subdomains of the Pax6 paired domain are regulated by alternative splicing. Genes Dev. **8,** 2022–2034.

GAO, Q.S., MEMMOTT, J., LAFYATIS, R., STAMM, S., SCREATON, G., and ANDREADIS, A. (2000). Complex regulation of tau exon 10, whose missplicing causes frontotemporal dementia. J. Neurochem. **74,** 490–500.

GRABOWSKI, P. (1998). Splicing regulation in neurons: Tinkering with cell-specific control. Cell **92,** 709–712.

GREEN, M.J. (1991). Biochemical mechanism of constitutive and regulated pre-mRNA splicing. Annu. Rev. Cell Biol. **7,** 559–599.

HANAMURA, A., CACERES, J.F., MAYEDA, A., FRANZA, B.R., and KRAINER, A.R. (1998). Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. RNA **4,** 430–444.

HANKE, J., BRETT, D., ZASTROW, I., AYDIN, A., DELBRUCK, S., LEHMANN, G., LUFT, F., REICH, J., and BORK, P. (1999). Alternative splicing of human genes: More the rule than the exception? Trends Genet. **15,** 389–390.

HAWKINS, J.D. (1988). A survey of intron and exon lengths. Nucleic Acids Res. **16,** 9893–9908.

HERBERT, A., and RICH, A. (1999). RNA processing and the evolution of eukaryotes. Nature Genet. **21,** 265–269.

JENSEN, K.B., DREGE, B.K., STEFANI, G., ZHONG, R., BUCHANOVICH, R.J., OKANO, H.J., YANG, Y.Y., and DARNELL, R.B. (2000). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron **25,** 359–371.

KRÄMER, A. (1996). The structure and function of proteins involved in mammalian pre-mRNA splicing. Annu. Rev. Biochem. **65,** 367–409.

KRAWCZAK, M., REISS, J., and COOPER, D.N. (1992). The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: Causes and consequences. Hum. Genet. **90,** 41–54.

LORSON, C.L., HAHNEN, E., ANDROPHY, E.J., and WIRTH, B. (1999). A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Proc. Natl. Acad. Sci. USA **96,** 6307–6311.

LIU, H.-X., ZHANG, M., and KRAINER, A.R. (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. **12,** 1998–1012.

LIU, H.-X., CHEW, S.L., CARTEGNI, L., ZHANG, M.Q., and KRAINER, A.R. (2000). Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol. Cell. Biol. **20,** 1063–1071.

MANLEY, J.L., and TACKE, R. (1996). SR proteins and splicing control. Genes Dev. **10,** 1569–1579.

MATTOX, W., PALMER, M.J., and BAKER, B.S. (1990). Alternative splicing of the sex determination gene transformer-2 is sex-specific in the germ line but not in the soma. Genes Dev. **4,** 789–805.

MAYEDA, A., and KRAINER, A.R. (1992). Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. Cell **68,** 365–375.

MCCRACKEN, S., FONG, N., YANKULOV, K., BALLANTYNE, S., PAN, G., GREENBLATT, J., PATTERSON, S.D., WICKENS, M., and BENTLEY, D.L. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature **385,** 357–361.

MIRONOV, A.A., FICKETT, J.W., and GELFAND, M.S. (1999). Frequent alternative splicing of human genes. Genome Res. **9,** 1288–1293.

MOORE, M.J. (2000). Intron recognition comes of AGe. Nature Struct Biol **7,** 14–16.

NAKAI, K., and SAKAMOTO, H. (1994). Construction of a novel database containing aberrant splicing mutations of mammalian genes. Gene **14,** 171–177.

NAYLER, O., CAP, C., and STAMM, S. (1998a). Human transformer-2-beta gene: Complete nucleotide sequence, chromosomal localisation and generation of a tissue specific isoform. Genomics **53,** 191–202.

NAYLER, O., STRÄTLING, W., BOURQUIN, J.-P., STAGLJAR, I., LINDEMANN, L., JASPER, H., HARTMANN, A.M., FACKELMEYER, F.O., ULLRICH, A., and STAMM, S. (1998b). SAF-B

couples transcription and pre-mRNA splicing to SAR/MAR elements. Nucleic Acids Res. **26,** 3542–3549.

NEUBAUER, G., KING, A., RAPPSILBER, J., CALVIO, C., WATSON, M., AJUH, P., SLEEMAN, J., LAMOND, A., and MANN, M. (1998). Mass spectometry and EST-database searching allows characterization of the multiprotein spliceosome complex. Nature Genet. **20,** 46–50.

NEWALD, A.F., LIU, J.S., and LAWRENCE, C.E. (1995). Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. Protein Sci. **4,** 1618–1632.

PEREZ, I., LIN, C.-H., MCAFEE, J.G., and PATTON, J.G. (1997). Mutation of PTB binding sites causes misregulation of alternative 3′ splice site selection in vivo. RNA **3,** 764–778.

PHILIPS, A.V., and COOPER, T.A. (2000). RNA processing and human disease. Cell Mol. Life Sci. **57,** 235–249.

PRET, A.M., BALVAY, L., and FISZMAN, M.Y. (1999). Regulated splicing of an alternative exon of beta-tropomyosin pre-mRNAs in myogenic cells depends on the strength of pyrimidine-rich intronic enhancer elements. DNA Cell Biol. **18,** 671–683.

REDDY, R. (1989). Compilation of small nuclear RNA sequences. Methods Enzymol. **180,** 521–532.

ROBBERSON, B.L., COTE, G.J., and BERGET, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. **10,** 84–94.

SCHUL, W., DE JONG, L., and VAN DRIEL, R. (1998). Nuclear neighbours: The spatial and functional organization of genes and nuclear domains. J. Cell Biochem. **70,** 159–171.

SCREATON, G.R., CACERES, J.F., MAYEDA, A., BELL, M.V., PLEBANSKI, M., JACKSON, D.G., BELL, J.I., and KRAINER, A.R. (1995). Identification and characterization of three members of the human SR family of pre-mRNA splicing factors. EMBO J. **14,** 4336–4349.

SELVAKUMAR, M., and HELFMAN, D.M. (1999). Exonic splicing enhancers contribute to the use of both 3′ and 5′ splice site usage of rat beta-tropomyosin. RNA **5,** 378–394.

SMITH, K.P., MOEN, P.T., WYDNER, K.L., COLEMAN, J.R., and LAWRENCE, J.B. (1999). Processing of endogenous pre-mRNAs in association with SC-35 domains is gene specific. J. Cell Biol. **144,** 617–629.

SONTHEIMER, E.J., and STEITZ, J.A. (1993). The U5 and U6 small nuclear RNAs as active site components of the spliceosome. Science **262,** 1989–1996.

SOUTHBY, J., GOODING, C., and SMITH, C.W.J. (1999). Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of alpha-actinin mutually exclusive exons. Mol. Cell. Biol. **19,** 2699–2711.

STAKNIS, D., and REED, R. (1994). SR proteins promote the first specific recognition of pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. Mol. Cell. Biol. **14,** 7670–7682.

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.

STEITZ, J.A. (1992). Splicing takes a holliday. Science **257,** 888–889.

STOSS, O., SCHWAIGER, F.W., COOPER, T.A., and STAMM, S. (1999a). Alternative splicing determines the intracellular localization of the novel muscle-specific protein Nop30 and its interaction with the splicing factor SRp30c. J. Biol. Chem. 10951–10962.

STOSS, O., STOILOV, P., HARTMANN, A.M., NAYLER, O., and STAMM, S. (1999b). The *in vivo* minigene approach to analyze tissue-specific splicing. Brain Res. Prot. **4,** 383–394.

TACKE, R., and MANLEY, J.L. (1999). Determinants of SR protein specificity. Curr. Opin. Cell Biol. **11,** 358–362.

TANAKA, K., WATAKABE, A., and SHIMURA, Y. (1994). Polypurine sequences within a downstream exon function as a splicing enhancer. Mol. Cell. Biol. **14,** 1347–1354.

TARN, W.-Y., and STEITZ, J.A. (1997). Pre-mRNA splicing: The discovery of a new spliceosome doubles the challenge. TIBS **22,** 132–137.

THANARAJ, T.A. (2000). Positional characterisation of false positives from computational prediction of human splice sites. Nucleic Acids Res. **28,** 744–754.

TIAN, H., and KOLE, R. (1995). Selection of novel exon recognition elements from a pool of random sequences. Mol. Cell. Biol. **15,** 6291–6298.

USUKA, J., and BRENDEL, V. (2000). Gene structure prediction by spliced alignment of genomic DNA with protein sequences: Increased accuracy by differential splice site scoring. J. Mol. Biol. **297,** 1075–1085.

VARANI, G., and NAGAI, K. (1998). RNA recognition by RNP proteins during RNA processing. Annu. Rev. Biophys. Biomol. Struct. **27,** 407–445.

WANG, J., and MANLEY, J.L. (1995). Overexpression of the SR proteins ASF/SF2 and SC35 influences alternative splicing in vivo in diverse ways. RNA **1,** 335–346.

WATAKABE, A., TANAKA, K., and SHIMURA, Y. (1993). The role of exon sequences in splice site selection. Genes Dev. **7,** 407–418.

WEIGHARDT, F., BIAMONTI, G., and RIVA, S. (1996). The role of heterogeneous nuclear ribonucleoproteins (hnRNP) in RNA metabolism. BioEssays **18,** 747–756.

WISE, J.A. (1993). Guides to the heart of the spliceosome. Science **262,** 1978–1979.

WU, S., ROMBO, C.M., NILSON, T.W., and GREEN, M.R. (1999). Functional recognition of the 3′ splice site AG by the splicing factor U2AF. Nature **402,** 832–835.

ZAHLER, A.M., NEUGEBAUER, K.M., LANE, W.S., and ROTH, M.B. (1993). Distinct functions of SR proteins in alternative pre-mRNA splicing. Science **260,** 219–222.

ZHANG, M.Q. (1998). Statistical features of human exons and their flanking regions. Hum. Mol. Genet. **7,** 919–932.

ZHANG, M.Q., and MARR, T.G. (1994). Fission yeast gene structure and recognition. Nucleic Acids Res. **11,** 1750–1759.

ZHUANG, Y., and WEINER, A.M. (1986). A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. Cell 827–835.

Address reprint requests to:
Dr. Stefan Stamm
Institute of Biochemistry
University of Erlangen-Nuremberg
Fahrstrasse 17
91054 Erlangen, Germany

E-mail: stefan@stamms-lab.net

Received for publication May 15, 2000; received in revised form July 25, 2000; accepted August 14, 2000.