Supporting Information

Dataset Construction

Datasets of nonredundant, cDNA-verified transcripts were created from GenBank release 118 (Sept., 2000) using the GENOA genome annotation script (unpublished data). Briefly, GENOA extracts all available genomic and cDNA sequence data for a given organism from GenBank, and masks repetitive elements with RepeatMasker (A. Smit and P. Green, personal communication). BLASTN (1) is then used to identify cDNA/genomic pairs with significant blocks of identity. Such pairs are then aligned by using the spliced alignment algorithm MRNAVSGEN (unpublished work), which produces GenBank format annotation of the inferred exon and intron locations. The MRNAVSGEN algorithm is similar in concept to the SIM4 program (2), but is designed specifically for aligning cDNAs rather than expressed sequence tags. The GENOA script then resolves genomic sequences aligning to multiple cDNAs into separate regions containing single genes and checks overlapping alignments for evidence of alternative splicing. Finally, a nonredundant set was created by using BLASTP comparison of the encoded peptides. Because our goal here was to study constitutive splicing, transcripts found to be alternatively spliced were not used. For human, the Refseq collection of full-length cDNAs was used. For all other organisms except yeast, all GenBank cDNAs were used. In yeast, a comprehensive survey of intron-containing genes in yeast has recently been carried out (3), so we used this manually curated set for yeast rather than using the output of the GENOA script. This set of transcripts was downloaded from the websites (http://www.cse.ucsc.edu/research/compbio/yeast_introns) and (http://genome-www.stanford.edu/Saccharomyces). Intronless transcripts produced by GENOA in human were not considered because many of these are likely to represent pseudogenes.

Splice Site Statistics

We observed significant statistical dependencies between nucleotides at adjacent positions in both the 5′ and 3′ splice signals of all organisms considered in this study, except yeast for which too few introns were available to assess such dependence. Specifically, using consensus indicator chi-square analysis (as in ref. 4), significant dependencies ($p < 0.001$) were observed

between all pairs of adjacent positions from (-20,-19) to (-5,-4) in the 3′ss of all four multicellular organisms. The nature of these dependencies was in many cases a bias toward pairs of adjacent pyrimidine nucleotides, likely reflecting a preferred 3′ splice signal composition involving a tract of consecutive pyrimidine nucleotides at somewhat variable location relative to the 3′ splice junction. In 5′ splice signals, significant dependencies were observed between positions (-2, -1) and (+5, +6) relative to the 5′ splice junction in all four organisms, and at certain other pairs of positions that differed between organisms. The nature of the (-2, -1) effect is a slightly increased frequency of G at position –1 when the –2 base is A (both consensus nucleotides at the respective positions). The (+5, +6) effect is an increased frequency of T (consensus) at +6 when the +5 base is G (also consensus). Both of these effects have been observed previously in human 5′ splice signals and can be interpreted in terms of the thermodynamics of RNA duplex formation between U1 small nuclear RNA and the 5′ss and/or U6 small nuclear RNA and the 5′ss in the case of the (+5,+6) interaction (5). As expected, identification of introns was more accurate by using I1M splice site models than WMMs in all organisms studied (data not shown).

Monte Carlo Simulations

The purpose of the Monte Carlo simulations illustrated in Fig. 3 was to determine how much information, in a relative entropy sense, the splice signal motifs would need to contain to accurately determine the locations of short introns in transcripts from each organism. One way to accomplish this objective would be to steadily add biased nucleotides to the existing splice signal motifs (e.g., adding additional pyrimidine nucleotides to the pyrimidine tracts of introns) and then to measure the accuracy of short intron identification after each additional base. In practice, we used a slightly different procedure which puts the problem into a somewhat more general context and accomplishes the same goal.

For each organism, a set of artificial transcript sequences with the same lengths as the original transcripts was generated at random from a uniform nucleotide distribution (<0.25, 0.25, 0.25, 0.25> for the frequencies of A, C, G, and T, respectively). Next, a pair of artificial splice site motif WMMs, each of length 17 bases, was generated. The length 17 is similar to the lengths of real splice signal motifs and is long enough to generate motifs of high relative entropy.

Each artificial splice site motif was created by randomly selecting 17 frequency vectors from the set of 99 frequency vectors, <0.01, 0.33, 0.33, 0.33>, <0.02, 0.3266, 0.3266, 0.3266 >, <0.03, 0.3233, 0.3233, 0.3233 >, … <0.99, 0.0033, 0.0033, 0.0033>. Fusing these 17 frequency vectors generates a WMM. Such models were constructed for both the 5′ss and 3′ss and then these models were used to generate random splice signal sequences that were inserted at the exact locations of the original splice junctions of short introns in each transcript. Next, a simple splice site pair model (PAIRSCAN) was used to predict the short intron locations by using the artificial splice signal WMM models. The accuracy achieved by PAIRSCAN was then recorded, together with the sum of the relative entropies of the two splice signal motifs used, relative to a uniform background nucleotide distribution (<0.25, 0.25, 0.25, 0.25>). In this way, the exact length and exon-intron structure of each transcript was preserved, but the information content of the splice signal motifs was varied over a large range. Although the RelEnt of a motif is clearly very strongly related to its ability to specify intron locations (as seen in Fig. 3), the degree of degeneracy of the WMM (as measured by the RelEnt per position) also affects the results slightly. For example, slightly different results are obtained when using 12-bit splice signal motifs with 2 bits per position than with longer but more degenerate 12-bit motifs containing 1 bit per position because of the greater variability in splice site scores in the latter case.

Gibbs Sampling

Branch site motifs were identified by searching for 7-nt motifs in branch region sequences 15 to 45 nucleotides upstream of the 3′ss by using the Gibbs sampling software (6) downloaded from the ftp site at the National Center for Biotechnology Information (ftp://ftp.ncbi.nlm.nih.gov). We assumed that all possible branch sites would contain a canonical branch adenosine at position 6 of the 7-mer motif. Thus, sequence positions that did not overlap any 7-mer with adenosine at position 6 were masked (replaced by Ns) to increase the signal-to-noise ratio. For example, the sequence TTTTTTTTACTAACTTTTTTTTATTTT would be replaced with the sequence NNNTTTTTACTAACNNNTTTTTATNNN. The Gibbs motif sampler was run repeatedly on the masked sequences, varying the expected count parameter from 10% of the total number of sequences to 100%. A motif that satisfied the two criteria of being consistently generated by Gibbs sampling and showing significant complementarity to U2 small nuclear RNA was

observed in all organisms except *C. elegans*, where the motif generated most commonly by the Gibbs sampler showed little complementarity to U2 small nuclear RNA (Fig. 2*B*). For each organism, a representative motif generated by the Gibbs sampler was chosen and used to generate a WMM. In *S. cerevisiae*, the branch signal is known to be longer and is commonly located further upstream of the 3′ss. Therefore, the procedure was modified to search for an 11-mer motif in branch region sequences 15 to 200 nt upstream of the 3′ss.

Intron Composition Models

Let $\vec{N}^5 =< N_1^5, N_2^5, ..., N_{1024}^5 >$ represent the counts of the 1,024 possible pentanucleotides in all the introns from a given organism. For example, listing the pentamers in alphabetical order, N_1^5 is the count of the pentamer AAAAA, N_2^5 the count of AAAAC, etc. The count vectors for tetramers, triplets, doublets and nucleotides ($\vec{N}^4, \vec{N}^3, \vec{N}^2, \vec{N}^1$) are defined analogously. The pentamer frequency vector is then defined as $\vec{f}^5 =< f_1^5, f_2^5, ..., f_{1024}^5 >$ where $f_i^5 = N_i^5 / L'$, and L' is the total number of pentamers in all of the introns (sum of intron lengths less four times the number of introns), and similarly for $\vec{f}^4$, $\vec{f}^3$, etc. Let $\vec{g}^5, \vec{g}^4, \vec{g}^3, \ldots$ be the corresponding oligomer frequencies measured in transcripts as a whole (i.e. including exons). Now let $\vec{n}^5 =< n_1^5, n_2^5, ..., n_{1024}^5 >$ represent the pentamer count vector of a particular intron I, so n_2^5 is the number of occurrences of the pentamer AAAAC in that particular intron, for example. Similarly, let $\vec{n}^4$ represent the vector of tetramer counts in the intron, $\vec{n}^3$ the vector of triplet counts, etc. In calculating the pentamer count vector, pentamers that overlap the 5′ or 3′ splice signal or branch signal motifs of the intron are excluded. In calculating the tetramer count vector, tetramers that overlap or fall within one base of the 5′ss, 3′ss or branch signal are excluded. For the triplet count vector, triplets that overlap or fall within two bases of the 5′ss, 3′ss or branch signal are excluded, and so forth (this convention simplifies the notation below). The intron composition score of an intron I is then defined as: $s(I) = \sum_i n_i^5 \log_2(f_i^5 / g_i^5) - \sum_j n_j^4 \log_2(f_j^4 / g_j^4)$, where the first sum is taken over all 1,024 possible pentamers and the second sum is over all 256 tetramers. Subtracting the second sum effectively corrects for the over counting of tetramers overlapped by adjacent pentamers. When intron oligomer counts are defined as described above, this score is equivalent to the log-odds ratio of a homogeneous fourth-order Markov model of

intron composition over a homogeneous fourth-order Markov model of transcript composition. The notation used above can be more easily generalized to models that score only a subset of the 1,024 pentamers (as in Fig. 5) than can standard Markov chain notation. For example, suppose that we only want to assign scores to two pentamers, AATTG and TTGCC. The intron score for this model is then defined as:

$s'(I) = n^5_{AATTG} \log_2 (f^5_{AATTG} / g^5_{AATTG}) + n^5_{TTGCC} \log_2 (f^5_{TTGCC} / g^5_{TTGCC}) - n^3_{TTG'} \log_2 (f^3_{TTG} / g^3_{TTG})$ where $n^3_{TTG'}$ is the number of occurrences of the triplet TTG that occur at overlaps between the scored pentamers (i.e., in heptamers AATTGCC), which is a natural generalization of the Markov formula to models involving subsets of oligomers only. The intron score for other subsets is defined analogously, always subtracting terms corresponding to the oligomers (of size 1, 2, 3 or 4), which are generated by overlaps of scored pentamers.

Using INTRONSCAN for Gene Finding

INTRONSCAN was run on a sample of *Drosophila* genomic sequences using the 5′ss, 3′ss, branch and intron length models and a score cutoff of 16.25 bits. Given the frequency with which INTRONSCAN predicts short introns in *Drosophila* genomic sequences, a cluster of 4 or more predicted short introns within 1 kb is expected to occur only rarely in the genome (approximately once per 166 kb), if the predicted introns were randomly distributed. Such intron clusters are often associated with genes, as illustrated in Fig. 6.

References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997) *Nucl. Acids Res.* **25,** 3389-402.

2. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W (1998) *Genome Res.* **8,** 967-74.

3. Spingola, M., Grate, L., Haussler, D., & Ares, M., Jr. (1999) *RNA* **5,** 221-34.

4. Burge, C. (1998) in *Computational Methods in Molecular Biology*, eds. Salzberg, S. L., Searls, D. B. & Kasif, S. (Elsevier, Amsterdam), pp. 129-164.

5. Burge, C., & Karlin, S. (1997) *J. Mol. Biol.* **268,** 78-94.

6. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. (1993) *Science* **262,** 208-14.