Letter to the Editor

The Yeast Splice Site Revisited: New Exon Consensus from Genomic Analysis

Saccharomyces cerevisiae is an excellent model system to study the splicing of pre-mRNA introns (Rymond and Rosbash, 1992). However, different compilations of yeast introns have yielded conflicting exon consensus patterns (Csank et al., 1990; Rymond and Rosbash, 1992). The entire sequence of the yeast genome (Goffeau et al., 1996) now permits one to examine all the information for this organism. In doing so, we find a hitherto unnoticed consensus pattern near the boundaries of the exons and suggest that this pattern is the result of the interaction of the exon sequence with a small nuclear RNA (snRNA).

Among the several snRNAs involved in the splicing process, U5 is thought to interact with 5′ exon sequences to help define 5′ splice sites (reviewed by Steitz, 1992; Horowitz and Krainer, 1994). Wyatt et al. (1992) demonstrated that the mammalian U5 snRNA can be cross-linked to the region upstream of the 5′ splice sites and that Watson–Crick pairing to the 5′ exon is not required. However, in S. cerevisiae, Newman and Norman (1992) found such a pairing between U5 and the −2 and −3 exon positions. Recently, O'Keefe et al. (1996) showed in vitro that mutations involving the loop I region of U5 do not abolish the first catalytic step of intron splicing reaction but affect the second catalytic step. Furthermore, their in vivo experiments showed that all such mutations are lethal. Sontheimer and Steitz (1993) demonstrated that U5 continues to hold the 5′ exon throughout the splicing reaction. Hence, O'Keefe et al. argue that the role of the loop I of U5 is to tether the free 5′ exon after the first splicing step in the correct orientation for a nucleophilic attack at the 3′ splice site in the second catalytic step. Nonetheless, it is not clear how an interaction between loop I of U5 and the end of the 5′ exon could occur, because the sequence of loop I is highly conserved while exon sequences are variable.

By scanning the Saccharomyces Genome Database, we constructed an intron/exon subdatabase that contains all the introns for S. cerevisiae, using programs from Long et al. (1995). By eliminating the ORFs that contain questionable introns (for instance, hypothetical introns one nucleotide long), we obtained 214 introns in 210 genes (Table 1). (Forty-four of these were identified solely by computer prediction; however, their presence or absence does not alter the conclusions of the analysis.)

Table 1 shows the nucleotide pattern for the 10 exon sites that flank each intron. A quantitative measure of conservation is the information content at the i′th site, Rs(i), which ranges from zero to two for maximum information:

$$Rs(i) = 2 + \Sigma f_b \log 2(f_b) - e(n),$$

where f_b and $e(n)$ are the frequency of nucleotides and a correction term for sample size (n) (Schneider et al., 1986).

There is no conservation in the exon (positions 1 to 10) to the 3′ side of the splice site. However, significant conservation appears in the 5′ exon at positions −2, −3, and −4, evident both in the information content and in the frequencies. To identify a consensus, we take as a criterion that a single base should represent at least 40% of the total. Thus, we define an S. cerevisiae exon consensus for the splice site as $A_{47}A_{53}A_{44}N|N$ (the subscripts show the actual percentages of the bases and the bar "|" shows the intron position). This consensus is clearly different from those derived before, using smaller samples: $T_{44}G_{50}|N$ in Csank et al. (1990) from 18 introns, and $(G/A)_{74}|N$ in Rymond and Rosbash (1992) from 54 introns.

A biological role for this consensus would be in a pairing of the 5′ exon sequence to an snRNA. Newman and Norman (1992) demonstrated that nucleotides 5 and 6 in the highly conserved loop I of U5 snRNA pair with bases −2 and −3 in the 5′ exon sequence of the mRNA, based on the suppression of point mutations. The sequence of loop I is (from positions 1 to 9) -GCCUUUUAC. Thus, the −2 and −3 positions in the exon, which have a consensus AA, could form Watson–Crick base-pairings with nucleotides 5 and 6. The −4 exon consensus A is also complementary to nucleotide 7 in the loop, although an actual pairing would have to be proved by experiment.

This biological model of partial pairing between U5 and the exon sequences could be satisfied through random matching of a few bases at each site or could involve specific sequences. Thus, we asked if there were any pattern, such as an excess of AAA's, which might

Table 1. Conservation Analysis of Exon Regions Flanking S. cerevisiae Introns[a]

i	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	1	2	3	4	5	6	7	8	9	10
Rs (i):	.01	.02	.07	.05	.03	.13	.16	.27	.15	.07	.06	.01	.03	.08	.07	.02	.01	.07	.01	.00
Percentages:																				
a	28	30	33	37	34	38	47	53	44	28	34	22	28	30	33	33	33	29	29	31
c	18	18	15	23	20	11	16	13	15	15	13	26	22	15	22	22	21	16	21	24
g	24	23	19	18	21	19	16	13	14	37	27	21	17	19	13	20	23	18	21	22
t	30	29	33	22	25	32	21	21	27	20	26	31	33	36	32	25	23	37	29	23

[a] The accession codes of all introns used in this study are available at http://www.cell.com/supplemental/91/6/739.

Table 2. The Random Distribution of Consensus Nucleotides at Positions −4, −3, −2

Phase	AAA	AA*	A*A	*AA	A**	*A*	**A	***	χ^2	P
All	22/23.5	32/29.9	18/20.8	33/26.4	28/26.5	26/33.7	21/23.5	34/29.9	4.9	0.74
0	8/10.6	17/14.6	6/7.1	15/10.6	11/9.7	10/14.6	6/7.1	11/9.7	5.0	0.67
1	6/6.4	10/9.2	5/7.8	10/7.2	14/11.3	7/10.4	9/8.8	13/12.7	4.0	0.78
2	8/6.4	5/6.0	7/5.5	8/9.3	3/5.1	9/8.6	6/7.9	10/7.3	3.5	0.83

* : nucleotides G, or C, or T. The data are expressed as observation/expectation (Obs/Exp). Expected numbers (Exp) are calculated based on the assumption of random distribution. For example, in the first row, the expected number of AA* = $0.47 \times 0.53 \times (1 − 0.44) \times 214 = 29.9$ since the frequencies of A at positions −4, −3, and −2 are 0.47, 0.53, and 0.44, respectively, while in the second row (phase 0), since the consensus is $A_{50}A_{60}A_{42}$, the expected number of AA* = $0.50 \times 0.60 \times (1 − 0.42) \times 84 = 14.6$. The degree of freedom for the χ^2 test is 7.

show the presence of such specific enhancers. The first row of Table 2 shows the distribution of the consensus nucleotides among the 214 introns. Eighty-four percent of introns have at least one A at the consensus positions (most of the rest have at least one G; only 3% of introns contain neither A's nor G's). The eight trinucleotide patterns follow the random expectation (p = 74%). Therefore, the consensus pattern is not a consequence of the overuse of specific sequences, such as AAA, AAN, or NAA.

Furthermore, an alternative model would be that the consensus sequence is the result of some systematic bias, such as one caused by codon usage. Such a codon bias might appear if intron phase were involved in the determination of consensus sequences. We resorted the 214 introns into three groups according to intron phase and recalculated the consensus at −2, −3, −4 for each group. The phase 0 group (84 introns) has the consensus $A_{50}A_{60}A_{42}$; phase 1 (74 introns), $A_{47}A_{45}A_{41}$; and phase 2 (56 introns), $A_{41}A_{54}A_{52}$. While there is considerable variation (9–15 percentage points) between groups, this variation does not seem to reflect a conflict between the local amino acid composition and the requirement for matching with U5 snRNA, a possibility suggested by Fichant et al. (1992). For example, for phase zero, the position −3, the first nucleotide of a codon, has a higher percentage of A's than the wobble position at −4.

Table 2 also lists the distribution of the consensus nucleotides within the three groups. All show random distributions, suggesting again that codon or amino acid usage is not determining the high percentages of A's at positions −2, −3, and −4.

U1 snRNA also collaborates with U5 snRNA to bring the splice site together in the assembled spliceosome (Steitz, 1992). Nucleotides 9, 10, 11 of the yeast snRNA U1, CUU, may also pair with the −1, −2, and −3 nucleotides at the 5′ splice site (nucleotides 3–8 of U1 pair with the highly conserved sequence within the intron). Our consensus sequence partially matches these nucleotides.

In short, these exon consensus sequences support the idea that there is some interaction between the snRNAs and exon nucleotides. This work shows the advantage of doing computational analysis based on a full genome set rather than examining the small data sets that were previously available.

Manyuan Long, Sandro J. de Souza,
and Walter Gilbert
The Biological Laboratories, Harvard University
16 Divinity Avenue, Cambridge, Massachusetts 02138

References

Csank, C., Taylor, F.M., and Martindale, D.W. (1990). Nucleic Acid Res. 18, 5133–5141.

Fichant, G.A. (1992). Human Mol. Genet. 1, 259–267.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., et al. (1996). Science 274, 546–567.

Horowitz, D.S., and Krainer, A.R. (1994). Trends Genet. 10, 100–106.

Long, M., Rosenberg, C., and Gilbert, W. (1995). Proc. Natl. Acad. Sci. USA 92, 12495–12499.

Newman, A.J., and Norman, C. (1992). Cell 68, 743–754.

O'Keefe, R.T., Norman, C., and Newman, A.J. (1996). Cell 86, 679–689.

Rymond, B.C., and Rosbash, M. (1992). The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression, J.R. Broach, J.R. Pringle, and E.W. Jones, (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press), 143–192.

Schneider, T.D., Stormo, G.D., Gold, L., and Ehrenfeucht, A. (1986). J. Mol. Biol. 188, 415–431.

Sontheimer, E.J., and Steitz, J.A. (1993). Science 262, 1989–1996.

Steitz, J.A. (1992). Science 257, 888–889.

Wyatt, J.R., Sontheimer, E.J., and Steitz, J.A. (1992). Genes Dev. 6, 2542–2553.