RNA (1999), *5*:1135–1137. Cambridge University Press. Printed in the USA.

LETTER TO THE EDITOR

Genomic-scale quantitative analysis of yeast pre-mRNA splicing: Implications for splice-site recognition

PASCAL J. LOPEZ and BERTRAND SÉRAPHIN
European Molecular Biology Laboratory, D-69117, Heidelberg, Germany

Keywords: exon; genome; intron; microarray; *Saccharomyces cerevisiae*; sequence consensus; spliceosome; transcription frequency

The availability of the complete sequence of the genome of *Saccharomyces cerevisiae* (Goffeau et al., 1996) and other organisms has provided a substantial amount of information on their chromosomal organization. We need now to understand the coordinated expression of the many genes present in a given genome and the function of the various encoded proteins. Microarrays (DNA chips) offer the opportunity to analyze nucleic acids at the genomic scale and have been used to perform global gene expression studies (Lander, 1999). However, these techniques also have the potential to give us a new perspective on posttranscriptional processes involved in the regulation of gene expression. We provide here a striking example through the investigation of yeast pre-mRNA splicing. Our analysis demonstrates that pre-mRNA splicing is quantitatively a much more important process in this species than previously thought and shows that splicing-signal conservation is correlated with transcription efficiency.

It has often been suggested that pre-mRNA splicing is a minor process in yeast because only a tiny proportion of yeast genes contains introns (Dujon, 1996). However, the real question is what fraction of all transcripts is spliced? In other words, what proportion of total transcripts is dealt with by the splicing machinery? To answer this question quantitatively, we first created a database of known and predicted yeast introns. For this purpose, we retrieved relevant information from several databases (SGD: *Saccharomyces* Genome Database, YPD: Yeast Protein Database, MIPS: Munich Information center for Protein Sequences, EMBL: European Molecular Biology Laboratory) and recent articles (Long et al., 1997; Burge et al., 1999; Spingola et al., 1999). This information was manually edited to remove entries with annotation errors or where experimental evidence indicated the absence of an intron. We ended up with a list of 253 introns (available at http://www.embl-heidelberg.de/ExternalInfo/seraphin/yidb.html) present in 248 genes. The presence of intron is not uniform among the different class of genes; in particular, ~70% of genes encoding for ribosomal protein contain an intervening sequence (Woolford, 1989; Rymond & Rosbash, 1992; Planta & Mager, 1998; Spingola et al., 1999). For 228 of these genes, a transcription frequency (i.e., number of transcripts produced per unit of time) has been determined (Holstege et al., 1998). We could therefore estimate the fraction of pre-mRNAs among all transcribed cellular RNAs. This revealed that although only 3.8% of the genes contain introns, 27.1% of the transcripts synthesized by the cell are spliced. This large difference occurs because a large fraction of intron-containing RNAs, notably ribosomal proteins pre-mRNAs, are highly expressed and demonstrates that pre-mRNA splicing is much more preponderant in yeast than previously suspected.

Our intron database also allowed us to examine the conservation of pre-mRNA splicing signals present at the 5′ splice site, branchpoint, and 3′ splice site by calculating the frequencies of each nucleotide in these regions. To assess the significance of this analysis, we performed χ^2 tests, taking into account the biased composition of yeast coding sequences (see Fig. 1). This confirmed the importance of previously known splicing signals and demonstrated significant conservation of the flanking sequences (Fig. 1). Interestingly, tabulating the representation of the corresponding nucleotides in the pool of cellular pre-mRNAs, taking expression levels into account, reveals greater representation of the less-conserved positions (Fig. 1). This indicates that the selection pressure on splicing signals is stronger for highly expressed transcripts. Notably, our analysis revealed conservation of As at positions −2 to −4 of

Reprint requests to: Bertrand Séraphin, European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117, Heidelberg, Germany; e-mail: Seraphin@EMBL-Heidelberg.de.

https://doi.org/10.1017/S135583829999091X Published online by Cambridge University Press

Exon 1 | 5' Splice site

	A	A	A	G	G	U	A	U	G	U	U
f	49	53	46	36	100	97	99	88	100	90	47
f_E	52	52	44	41	100	100	99	87	100	91	50

Branch point*

	U	U	U	A	C	U	A	A*	C	A	A/U
f	52	62	95	99	98	100	100	100	100	62	83
f_E	60	74	97	99	99	100	100	100	100	77	86

3' Splice site | Exon 2

	U	U	U	U	U	U	U	U	U	U	N	U	A	U/C	A	G
f	42	42	47	45	48	54	57	72	58	45		46	54	99	100	100
f_E	46	51	63	50	53	62	65	86	79	49		46	62	100	100	100

FIGURE 1. Effect of variable gene expression on the composition of the conserved splicing signal of yeast introns. Conserved residues of the 5′ splice site, branchpoint, and 3′ splice site sequences (commonly used consensus sequences are underlined) that are significantly different from the composition of yeast coding sequences (χ^2 test, 3 degrees of freedom, $p < 0.001$) are presented. The global coding sequence composition used for the statistical tests is G = 0.20, A = 0.33, C = 0.19, and T = 0.28 (http://www.embl-heidelberg.de/ExternalInfo/seraphin/ydata.html); however, similar results were obtained using the composition of either highly or poorly expressed coding sequences (data not shown; Hani & Feldmann, 1998). f corresponds to the frequency of the most significantly represented base at each position. f_E corresponds to the frequency of the most significantly represented base at each position corrected for expression frequency. To calculate f_E we weighted the frequency of each base at a given position by the expression frequency of the corresponding transcripts for the 233 yeast introns analyzed. That is, if i represent a transcript, and TF_i represents its transcription frequency per hour, and this transcript contains base B_i at a (the position analyzed), then the expressed frequency of base X at that position is:

$$f_E(X) = \frac{\sum_i (TF_i \cdot a_i)}{\sum_i TF_i}$$

where $a_i = 1$ if $B_i = X$ and $a_i = 0$ if $B_i \neq X$ for all possible transcripts i. Note the stronger conservation of splicing signals in the transcript population.

exon 1 consistent with previous findings (Long et al., 1997; Spingola et al., 1999). We also found a statistically significant conservation of G at the last position of exon 1, as independently noted by Spingola et al. (1999). Indeed, although the G frequency at this position is only 36%, it is the 5′ splice site nucleotide that is the most affected by taking into account expression levels (+5%). The significance of these values is further reinforced by the relatively low level of G in yeast coding sequences. This conclusion differs strikingly from the one reached by Long et al. (1997). The functional conservation of a G at the end of exon 1 is furthermore consistent with a functional base-pairing interaction between this nucleotide and the U1 snRNA (Kandels-Lewis & Séraphin, 1993). However, the positions that were most affected by taking into account transcription levels were located in the polypyrimidine tract upstream of the 3′ splice site (Fig. 1; Parker & Patterson, 1987). The increase in nucleotide frequency within the polypyrimidine tract when comparing genome versus transcript data can be as high as 21% for U at position −8. This suggests that yeast and metazoan introns may share more similarities in this region than previously assumed.

Our approach reveals that analysis of genome sequences may give a biased view, and that careful quantitative evaluation should therefore be considered for a full understanding of the corresponding cellular processes.

ACKNOWLEDGMENTS

We are grateful to I. Mattaj and members of the laboratory for careful reading of the manuscript and constructive criticism. B.S. is on leave from Centre National de la Recherche Scientifique.

Received April 30, 1999; returned for revision May 21, 1999; revised manuscript received June 15, 1999

REFERENCES

Burge CB, Tuschl T, Sharp PA. 1999. Splicing of precursors to mRNAs by the spliceosomes. In: Gesteland RF, Cech TR, Atkins JF, eds. *The RNA world.* Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. pp 525–560.

Dujon B. 1996. The yeast genome project: What did we learn? *Trends Genet 12*:263–270.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hohcisel JD, Jacq C, Johnston II, Louis EJ, Newco HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 genes. *Science 274*:546–547.

Hani J, Feldmann H. 1998. tRNA genes and retroelements in the yeast genome. *Nucleic Acids Res 26*:689–696.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell 95*:717–728.

Kandels-Lewis S, Séraphin B. 1993. Involvement of U6 snRNA in 5′ splice site selection. *Science 262*:2035–2039.

Lander ES. 1999. Array of hope. *Nat Genet 21*:3–4.

Long M, de Souza S, Gilbert W. 1997. The yeast splice sites revisited: New exon consensus from genomic analysis. *Cell 91*: 739–740.

Parker R, Patterson B. 1987. Architecture of fungal introns: Implications for spliceosome assembly. In: Dudock B, Inouye M, eds. *Molecular biology of RNA, new perspectives*. New York: Academic Press. pp 133–149.

Planta RJ, Mager WH. 1998. The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. *Yeast 14*:471–477.

Rymond B, Rosbash M. 1992. Yeast pre-mRNA splicing. In: Broach JR, Pringle J, Jones EW, eds. *The molecular and cellular biology of the yeast Saccharomyces cerevisiae*, Vol. 2. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. pp 143–192.

Spingola M, Grate L, Haussler D, Ares M Jr. 1999. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA 5*:221–234.

Woolford JL Jr. 1989. Nuclear pre-mRNA splicing in yeast. *Yeast 5*:439–457.

https://doi.org/10.1017/S135583829999091X Published online by Cambridge University Press