Association of Intron Phases with Conservation at Splice Site Sequences and Evolution of Spliceosomal Introns

Manyuan Long and Michael Deutsch

Department of Ecology and Evolution, University of Chicago

How exon-intron structures of eukaryotic genes evolved under various evolutionary forces remains unknown. The phases of spliceosomal introns (the placement of introns with respect to reading frame) provide an opportunity to approach this question. When a large number of nuclear introns in protein-coding genes were analyzed, it was found that most introns were of phase 0, which keeps codons intact. We found that the phase distribution of spliceosomal introns is strongly correlated with the sequence conservation of splice signals in exons; the relatively underrepresented phase 2 introns are associated with the lowest conservation, the relatively overrepresented phase 0 introns display the highest conservation, and phase 1 introns are intermediate. Given the detrimental effect of mutations in exon sequences near splice sites as found in molecular experiments, the underrepresentation of phase 2 introns may be the result of deleterious-mutation-driven intron loss, suggesting a possible genetic mechanism for the evolution of exon-intron structures.

Introduction

Most eukaryotic nuclear genes share the common structural feature that nuclear introns disrupt protein-coding regions. Studies of such gene structures have revealed that intron-exon structures were subjected to various evolutionary changes (Long, Desouza, and Gilbert 1995; Gilbert, Desouza, and Long 1997) which include intron loss, intron gain, and intron drift as major forms. An ample literature documents clear cases of intron loss in many gene families (e.g., Fink 1987; Charlesworth, Liu, and Zhang 1998), but concrete evidence for the other two forms of intron evolution is still being sought, and a few cases have been reported (Giroux et al. 1994; Stoltzfus et al. 1997; Logsdon, Stoltzfus, and Doolittle 1998; O’Neill et al. 1998). The evolutionary forces driving such evolutionary changes in intron-exon structures remain unexplained, despite extensive studies (Palmer and Logsdon 1991; Long, Desouza, and Gilbert 1995; Desouza, Long, and Gilbert 1996; DeSouza et al. 1996; Gilbert, Desouza, and Long 1997; Long and Desouza 1998). One challenge is that the distribution across species of introns in a given gene often provides little information. Recently arising genomic analysis, or database analysis (Long, Desouza, and Gilbert 1995; Long, Rosenberg, and Gilbert 1995), may provide an efficient approach to characterization of the variation of eukaryotic gene structures.

Intron phases, the positions of introns between or within codons (Sharp 1981), are a conservative character of eukaryotic gene structures, because any phase change requires either compensatory double mutations or more complex molecular mechanisms. There are three intron phases: phase 0 introns are located between codons, phase 1 introns are located between the first and second nucleotides of a codon, and phase 2 introns are located between the second and third nucleotides. General functions or phenotypic effects of spliceosomal introns have not been reported except in exceptional cases where introns harbor regulatory elements or code for snRNAs (Tycowski, Shu, and Steitz 1996; Long and DeSouza 1998). Remarkably, when a large number of introns in independent genes in the databases were analyzed, it was found that a dominant proportion of introns are of phase 0, showing that introns tend to keep codons intact, and the proportions of both phase 1 and phase 2 introns are significantly lower than random expectation (Fedorov et al. 1992; Long, Rosenberg, and Gilbert 1995; Tomita, Shimizu, and Brutlag 1996).

For overrepresented phase 0 introns, two alternative explanations were offered based on alternative hypothetical origins of introns. The direct explanation is that because primitive genes might code for short peptides, such as independent functional or structural units, the modern genes evolved greater complexity by assembling primitive genes in an intron-exon structure, and all introns were of phase 0. According to this, the exon theory of genes (Gilbert 1987), phase 1 and 2 introns were features derived by intron movement. Alternatively, overrepresented phase 0 introns could also be explained by insertion to preferred sites of genes, “proto-splice sites” (Dibb and Newman 1989), although so far no such sites that fit the distribution of intron phases have been found (Long et al. 1998). Although both approaches seek historical reasons for phase 0 overrepresentation, several interesting problems remain. Consider the hypothesis that phase 1 and 2 introns result from movement of phase 0 introns as a consequence of recombination of primitive genes. In this scenario, a one-nucleotide shift toward the 5’ codon would create a phase 2 intron, while a one-nucleotide shift toward the 3’ codon would yield a phase 1 intron. If shifts in either direction are equally likely, then the proportions of phase 1 and 2 introns should be similar. This prediction differs from the observed unequally probable phase distribution.

Since the introns have to be spliced before mature mRNAs are made from pre-mRNAs, the mechanism involved in the splicing process may be related to the distribution of intron phases. One important factor is...
how signals for recognizing splice sites are distributed and changed, a mutation parameter that may lead to the loss of introns. In the early 1980s, some conserved sequences were found adjacent to splice sites in both exons and introns (Mount 1982). Stephens and Schneider (1992) further characterized the distribution of conserved nucleotides in the splice sites of 1,800 human introns using logo analysis, a quantitative graphical display for information contents in the intron and exon nucleotide sites near the splice boundaries. A similar analysis was extended to other model species later (Long et al. 1998). These statistical analyses have located important nucleotide sites and elucidated their relative importance within and among the sites at splice boundaries and branching points.

A general picture of the molecular interactions between the nucleotides at splice sites and the components of spliceosomes, the machinery to recognize and cut introns accurately, has been described (Horowitz and Krainer 1994; Reed 1996). Using cross-linking in a mammalian system, the human HeLa cell line, Wyatt, Sontheimer, and Steitz (1992) and Sontheimer and Steitz (1993) identified interactions between the two nucleotides immediately upstream of intron 5′ donor sites and U5 snRNA. By suppressing point mutations in Saccharomyces cerevisiae, Newman and Norman (1991, 1992) demonstrated that Watson-Crick pairing between U5 snRNA and exon nucleotides near both 5′ and 3′ splice sites plays an essential role in splicing reactions, which was further confirmed to be general in a genomic analysis of splice sites in yeast (Long, DeSouza, and Gilbert 1997). Many point mutations immediately upstream of 5′ donor sites have been identified and shown to abort splicing (Rymond and Rosbash 1992; Aebi et al. 1987). All of these observations and experiments show interactions between spliceosomes and nucleotide sequences in the exon regions near splice sites; perturbation of these interactions would have deleterious effects on an organism.

Are the exon regions close to splice sites (donor and acceptor) related to the distribution of intron phases? We analyzed the phase distribution of spliceosomal introns and the degree of conservation of the exon sequences near splice sites in a large exon database that contains all recently published introns and exons in eukaryotic genes. (For the sake of convenience, we will simply use the word “intron” to refer to the spliceosomal intron). We report here a significant correlation between the distribution of intron phases and conservation of the splicing signals in the exon nucleotides near splice sites and explore the evolutionary implications of such a correlation for the evolution of eukaryotic genes.

Materials and Methods

Splice Site Sequence Databases and Exon Databases

Two types of databases were constructed from GenBank 106: (1) exon databases and (2) splice site sequence databases. The procedure for constructing exon databases was similar to that of Long, Rosenberg, and Gilbert (1995). Briefly, using an earlier computer program (Long, Rosenberg, and Gilbert 1995), we selected all the genes that contain known intron-exon structures as defined in the feature tables to form a raw database. Pseudogenes, genes with questionable protein lengths as defined by feature tables, and putative Caenorhabditis elegans genes identified by computer programs were discarded to avoid possible artifacts. Then we wrote a program to calculate various parameters such as the positions and phases of introns. In total, there are 16,989 genes in the raw database, each with defined intron positions and phases. To avoid possible bias brought about by uneven representation of different gene families, we used the GBPURGE program to purge all redundant sequences at a criterion of 20% amino acid identity: If two protein sequences shared an identity defined by fasta3 program (Pearson 1996) as greater than 20%, we kept the sequence with more exons and deleted the other sequence. We deleted all splice sequences for introns lacking consensus GT…AG, because such sequences may not be correct splice sequences (the small number of AT…AC introns [Hall and Padgett 1994; Tarn and Steitz 1997] that were also excluded in the computation does not change the result of statistical analyses).

Conservation Measurement of Splice Sites

To measure the conservation of the exon sequences that flank the introns, we aligned all splice sequences in the splice sequence databases and calculated the percentages of four nucleotides, A, T, G, and C, in each of 10 sites upstream or downstream of the splice boundaries of donor and acceptor sites. We then used two approaches to measure conservation. The first one, a now standard approach for analyzing sequence conservation around splice sites, is an information measurement method proposed by Schneider et al (1986) and Schneider (1991) which has the advantage over conventional goodness-of-fit tests, such as the chi-square test, that it is additive across sites and free from the influence of sample size. The second approach is simply to calculate the percentage of the most conservative consensus sites.

Information Analysis

The information for each base is calculated according to the equation

\[ R_e(l) = 2 - H(l) - e(n), \]

where \( H(l) = -\sum f_i(l) \log_2 f_i(l), \) \( f_i(l) \) being the frequency of nucleotide \( i \) (A, T, G, or C) at base \( l \), and \( e(n) = 3/ (2n \log_2 2) \) is a term for correcting sample size \( n \) (the total number of splice sequences). In the case of the splice sequence database, \( e(n) \) is negligible, since \( n \) is large. \( R_e(l) \) reaches the maximum of 2, when only one nucleotide is present, indicating the highest conservation.

Logo analysis (Schneider and Stephens 1990), a graphic description of sequence conservation, was used. Based on the additive property of \( R_e(l) \), we calculated the total information by summing over all five con-
served exon sites (positions −3, −2, and −1 at the 5′ splice site and positions 1 and 2 at the 3′ splice site):

\[ R = \sum_{-1 \text{ to } -3} R_i(l_{\text{donor}}) + \sum_{1 \text{ to } 2} R_i(l_{\text{acceptor}}), \]

with the variance

\[ V = 5 \left( \sum_{\text{all } nb} P_{nb} H_{nb}^2 - (E H_{nb})^2 \right). \]

where \( P_{nb} \) is the probability \((n!/(n_a!n_c!n_g!n_t!))\) for the particular combination \( n = n_a + n_c + n_g + n_t \) (see Schneider et al. 1986). The maximum test for the significance of \( R \) values of three phases followed the method developed by Schneider et al. (1986) and Schneider (http://www-lecb.ucsf.edu/~toms/delila/ttest.html), which used a \( t \)-test \((t = (R_1 - R_2)/(S_1^2/N_1 + S_2^2/N_2)^{1/2})\) for comparison of two \( R \) values with small sample sizes, where \( S^2 \) and \( N \) are variance and sample size, respectively, and subscripts 1 and 2 indicate samples 1 and 2, respectively. In this investigation, the sample sizes for three intron phases are very large (larger than 3,000), and the standard normal distribution is used in the test.

**Results**

The final exon database included 2,895 genes that were independent or quasi-independent and contained 17,044 introns. We calculated the proportions of the three intron phases in this database. The results were similar to those reported previously (Long, Rosenberg, and Gilbert 1995; Tomita, Shimuzu, and Brutlag 1996): 51% phase 0 introns, 27% phase 1 introns, and 22% phase 2 introns.

For the splice sequence database, we wrote a program to collect sequences of 10 nucleotides on both the intron and exon sides of each splice site (both donor and acceptor sites) from all genes for which defined intron-exon structures with intron sequences were available. Then, we used the splice sequences in the genes of the purged exon databases to form a raw splice sequence database for further analysis. After we deleted redundant sequences, we generated a final splice sequence database that contained 16,295 independent splice sequences. Intron phase proportions from this database are approximately equal to those calculated from the exon database (the differences are smaller than 1%).

The logo in figure 1 describes the distribution of conservation of the nucleotides near donor and acceptor splice sites for eukaryotic intron-containing genes in general. It shows that the highest conservation is within introns and that there is also a limited conservation in the adjacent exon sites, a similar situation to that described for human introns (Stephens and Schneider 1992) and those of other model organisms (Long et al. 1998). The logos for the total splice sequence database (fig. 1) and the three subdatabases (unpublished data) show a steady decline of information from position −1 to position −3 in the exon regions near donor sites (5′ splice sites) (after position −4, the information becomes indistinguishable from random noise). In the exon regions near acceptor sites (3′ splice sites), there is limited conservation in positions 1 and 2.

We conducted similar conservation analyses on the subdatabases of splice sequences for the three intron phases separately. Since the signals within intron regions are similarly high among three intron phases, we focus on the exon regions near the splice sites.

Table 1 shows the results of these calculations. Surprisingly, there is a significant correlation between the proportions and degrees of conservation among the three
Table 1
Intron Phase Proportion and Conservation

<table>
<thead>
<tr>
<th>PHASE</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion (%)</td>
<td>51</td>
<td>27</td>
<td>22</td>
</tr>
<tr>
<td>Conservation</td>
<td>Information</td>
<td>2.03 (0.58 x 10^-3)</td>
<td>1.38 (1.18 x 10^-3)</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>G percentage</td>
<td>41 (5.41 x 10^-3)</td>
</tr>
</tbody>
</table>

NOTE.—The individual R(i) values in five exon positions (−3, −2, −1, 1, and 2) were as follows: for phase 0, 0.16, 0.55, 0.89, 0.30, and 0.13; for phase 1, 0.10, 0.18, 0.89, 0.14, and 0.07; and for phase 2, 0.07, 0.37, 0.47, 0.12 and 0.05. The standard errors for information contents are listed in parentheses, and the standard deviations for G|G percentages in parentheses were calculated as square root of p(1 − p)N; p and N are the proportion of G|G type sequences and the sample size, respectively. The differences in the G|G percentages between the three intron phases are significant at P > 10^-6. The sample sizes for calculating information contents of phase 0, 1, and 2 were 8,280, 4,395, and 3,609, respectively. The database contained 17,044 introns.

Table 2
Intron Phase and Conservation at Splice Sites in Six Model Organisms

<table>
<thead>
<tr>
<th>PHASE</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (4,620 introns)</td>
<td>Percentage</td>
<td>46</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Information (bits)</td>
<td>2.42 (0.002)</td>
<td>1.79 (0.003)</td>
</tr>
<tr>
<td>Mouse (1,565 introns)</td>
<td>Percentage</td>
<td>45</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Information (bits)</td>
<td>2.31 (0.007)</td>
<td>1.74 (0.009)</td>
</tr>
<tr>
<td>Rat (513 introns)</td>
<td>Percentage</td>
<td>47</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Information (bits)</td>
<td>2.54 (0.020)</td>
<td>1.95 (0.027)</td>
</tr>
<tr>
<td>Chicken (329 introns)</td>
<td>Percentage</td>
<td>50</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Information (bits)</td>
<td>2.45 (0.030)</td>
<td>2.23 (0.044)</td>
</tr>
<tr>
<td>Arabidopsis thaliana (8,214 introns)</td>
<td>Percentage</td>
<td>57</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Information (bits)</td>
<td>2.18 (0.001)</td>
<td>1.69 (0.003)</td>
</tr>
<tr>
<td>Schizosaccharomyces pombe (1,387 introns)</td>
<td>Percentage</td>
<td>44</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Information (bits)</td>
<td>1.08 (0.007)</td>
<td>1.03 (0.012)</td>
</tr>
<tr>
<td>Drosophila (1,173 introns)</td>
<td>Percentage</td>
<td>45</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Information (bits)</td>
<td>1.43 (0.009)</td>
<td>0.900 (0.014)</td>
</tr>
</tbody>
</table>

NOTE.—“Percentage” refers to the proportions of three intron phases in various organisms. The standard errors for information contents are listed in parentheses.

Discussion

We observed that the proportion of an intron phase is correlated with the extent of conservation in the adjacent exon sequences. The sequences flanking phase 2 introns are the most variable. The two observations that lead to the same conclusion suggest that the observed correlations are not statistical artifacts but may implicate biological connections between the two parameters we
observed, the proportions of intron phases and the changes in splice sequences in exon regions.

The first interpretation is that the three classes of introns differ with regard to the recognition signaling role of the exon sequences immediately adjacent to the splice site. The exon sequences near the phase 2 introns are the least important to recognition by spliceosomes, so less splicing functional constraint leads to more variable sequences to meet the coding requirement in various genes, a situation similar to the neutral evolution of homologous genes (Kimura 1983). Functionally, the exon sequences flanking phase 0 introns, which have developed the most conservative use of nucleotides, are most important to splicing. This scenario is unlikely, because there is no indication that a spliceosome is “aware” of the reading frame (E. J. Sontheimer, personal communication).

The second interpretation is based on the empirical fact that mutations in the exon sequences flanking introns are often detrimental to splicing. The direct change in exon nucleotides at 5′ splice sites of phase 0 introns was found to yield reduced or abolished splicing in both yeast and rabbit genes (Felber, Orkin, and Hamer 1982; Seraphin and Rashbash 1990). Spontaneous mutations involving the GT dinucleotide at 5′ splice sites were found to inactivate normal splicing but activate cryptic splicing sites G|GT in human thalassemic genes (Treisman et al. 1982). The phases of the mutated introns in these observations included phases 0 and 2, suggesting that the detrimental effects of the mutations may not depend on the phases of introns. Thus, there could be stronger negative selection on the more variable exon boundaries, since they would more frequently generate deleterious changes. Therefore, the low frequency of phase 2 introns could be the consequence of deleterious-mutation-driven intron loss, since the exon regions flanking these introns are the most variable and create the highest deleterious mutation rates. This hypothesis, which differs from the first interpretation, offers an explanation for the correlations between intron phase distributions in independent eukaryotic genes and exon conservation.

While the consequence of the correlation between conservation of exon sequences flanking introns and intron phase proportions is clear, the molecular mechanism responsible for the variation of the conservation among the three intron phases remains unknown. A most attractive model, proposed by Fichant (1992), holds that the nucleotide composition in exon regions might depend on the balance between the conflicting demands of coding constraint and spliceosomal functional constraint. It is plausible that the variation in conservation reflects such a balance. However, one simple prediction from this theory, that nucleotides in wobble positions should meet the requirement for constraint better than first and second codon positions, is not supported by the information distribution at the five exon sites. For example, the amounts of information at position −1 for phases 0 and 1 are equal, although the −1 site in phase 0 introns is a wobble position, while the −1 site in phase 1 is the first codon position (table 1). Evidently, the balance between spliceosome and amino acid choice is more complex than this simple prediction. One possibility is that the exon sequences across phase 1 and 2 introns may be subject to more protein-coding constraint than the exon sequences near phase 0 introns, whose −1 nucleotide at the 5′ side is in wobble position. Thus, different genes would choose different amino acids in the boundary of phase 1 and 2 introns. This is not an unreasonable assumption, given the discovery that there are excess phase 0 introns, instead of phase 1 and 2 introns, in the boundary regions between protein modules, where the coding constraints may be lower than those in the regions inside the modules (DeSouza et al. 1998).

The uneven distribution of intron phases is likely a mixed consequence of historical events and natural selection. The ancient origin of introns may yield over-representation of phase 0 introns; subsequent intron drift could produce equal portions of phase 1 and 2 introns. Then, as a result of negative selection, phase 2 introns have been lost more quickly than phase 0 or 1 introns and thus become the least represented.

Compared with the previous search for historical reasons, the correlation found in this report sheds additional light on the evolution of the eukaryotic gene: the impact of evolutionary forces. Thus, the evolution of eukaryotic gene structures may not be entirely a process of neutral evolution. A more complete picture requires both historical and mechanistic approaches. We propose that deleterious-mutation-driven intron loss may have played a significant role in the evolution of eukaryotic gene structures, because variation in exon regions may, along with other factors that also affect splicing reactions (Sontheimer and Steitz 1993; Reed 1996), significantly affect recognition of introns by spliceosomes. The study reported in this paper may represent a new approach to dealing with the mechanisms responsible for the evolution of gene structures.

Finally, the roles that various evolutionary forces play in molecular evolution have mainly been concerned with interpreting variation in DNAs or proteins (Kimura 1983; Gillespie 1991; Li 1997). In particular, natural selection and genetic drift have become increasingly recognized as factors that shape the patterns of sequence variation (Kreitman and Akashi 1995). This study suggests that a new evolutionary force, deleterious-mutation-driven intron loss, may also control the evolution of gene structures and provides a useful conceptual extension of molecular evolution to the understanding of the origin and evolution of new gene structures.

Acknowledgments

We thank E. J. Sontheimer, C.-I. Wu, and J. Spoford for critical reading of the manuscript; W. Gilbert, M. Kreitman, P. A. Sharp, T. Nagylaki, C. Bergman, and the members of our laboratory for stimulating discussions; T. Schneider and R. Block for help in compiling the programs for the information analysis downloaded from Schneider’s web sites (http://www-lecb.ncifcrf.gov/~toms); the David and Lucile Packard Foundation.
for a Packard Fellowship in Science and Engineering; and the National Science Foundation for support.

LITERATURE CITED


ANTONY DEAN, reviewing editor

Accepted August 3, 1999