Testing the ‘‘Proto-splice Sites’’ Model of Intron Origin: Evidence from Analysis of Intron Phase Correlations

Manyuan Long and Carl Rosenberg†*

*Department of Ecology and Evolution, University of Chicago; and †Falling Rain Genomics, Inc., Palo Alto, California

A few nucleotide sites of nuclear exons that flank introns are often conserved. A hypothesis has suggested that these sites, called ‘‘proto-splice sites,’’ are remnants of recognition signals for the insertion of introns in the early evolution of eukaryotic genes. This notion of proto-splice sites has been an important basis for the insertional theory of introns. This hypothesis predicts that the distribution of proto-splice sites would determine the distribution of intron phases, because the positions of introns are just a subset of the proto-splice sites. We previously tested this prediction by examining the proportions of the phases of proto-splice sites, revealing nothing in these proportion distributions similar to observed proportions of intron phases. Here, we provide a second independent test of the proto-splice site hypothesis, with regard to its prediction that the proto-splice sites would mimic intron phase correlations, using a CDS database we created from GenBank. We tested four hypothetical proto-splice sites G|G, AG|G, AG|GT, and C/AAG|R. Interestingly, while G|G and AG|GT site phase distributions are not consistent with actual introns, we observed that AG|G and C/AAG|R sites have a symmetric phase excess. However, the patterns of the excess are quite different from the actual intron phase distribution. In addition, particular amino acid repeats in proteins were found to partially contribute to the excess of symmetry at these two types of sites. The phase associations of all four sites are significantly different from those of intron phases. Furthermore, a general model of intron insertion into proto-splice sites was simulated by Monte Carlo simulation to investigate the probability that the random insertion of introns into AG|G and C/AAG|R sites could generate the observed intron phase distribution. The simulation showed that (1) no observed correlation of intron phases was statistically consistent with the phase distribution of proto-splice sites in the simulated virtual genes; (2) most conservatively, no simulation in 10,000 Monte Carlo experiments gave a pattern with an excess of symmetric (1, 1) exons larger than those of (0, 0) and (2, 2), a major statistical feature of intron phase distribution that is consistent with the directly observed cases of exon shuffling. Thus, these results reject the null hypothesis that introns are randomly inserted into preexisting proto-splice sites, as suggested by the insertional theory of introns.

Introduction

A major feature of eukaryotic genes is their exon-intron structure. Mature mRNAs form after nuclear introns are spliced from a pre-mRNA transcript by complex machinery, the spliceosome, made of proteins and small nuclear ribonucleoproteins (snRNPs). (For simplicity, we will ignore the existence of the very small portion of introns that are spliced by other machinery.) During the splicing process, the components of a spliceosome need to establish particular interactions with parts of the intron and its flanking exons to ensure accurate and efficient splicing (for a review, see Moore, Query, and Sharp 1993; Burge, Tuschl, and Sharp 1999). This assertion was tested and confirmed in many experiments which used either spontaneous or suppression mutations in yeasts (Newman and Norman 1991, 1992) and mammals (Treisman et al. 1982). To establish these interactions in the splicing process of modern mRNA, conserved nucleotide sequence patterns have evolved within introns and exons as splicing signals. The most highly conserved sequences in an intron are the donor and acceptor sites and the branching site; exons retain a limited degree of conservation (Long et al. 1997, 1998).

Despite these observations of the role of the conserved sequences in exons in the splicing reaction, an alternative view of flanking exon sequence conservation is that these are relics of recognition signals for the insertion of introns, which began soon after the rise of eukaryotes. Thus, these sites, dubbed ‘‘proto-splice sites’’ (Dibb and Newman 1989), provide a physical entity for a possible mechanism of intron origin and have become a central feature of the intron-late hypothesis (for a recent review, see Logsdon 1998; Logsdon, Stoltzfus, and Doolittle 1998). However, this proposal has rarely been put to the test (Long et al. 1998), although it was often invoked in the introns-late argument (e.g., Lee, Stapleton, and Huang 1991). To help define proto-splice sites, one can, on a case-by-case basis, compare the exon sequences flanking a site occupied by an intron in one gene but lacking an intron in its homology if there are an adequate number of such pairs of homologous genes (Dibb and Newman 1989; Logsdon 1998). However, a powerful test of the proto-splice model has to be built on a statistical description of general states of introns, because the definition of proto-splice sites was based on hypothetical ancient eukaryotic genomes and thus would determine all introns of subsequent origin. The distribution of intron phases in eukaryotic genomes provides the first opportunity to develop such a test.

Intron phases were defined as relative positions of an intron within or between codons (an intron is of phase 0, 1, or 2 if it is located between two intact codons or within a codon after the first or second nucleotide, respectively). Because introns are thought to be func-

Key words: intron phase, proto-splice sites, exon shuffling, evolution of intron-exon structures, intron insertion.

Address for correspondence and reprints: Manyuan Long, Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637. E-mail: mlong@midway.uchicago.edu.

Mol. Biol. Evol. 17(12):1789–1796. 2000

tionless in general and thus are usually viewed as neutral evolutionary units, an obvious prediction was that the distribution of the three intron phases should be random, like the distribution of point mutations in other functionless genetic elements (e.g., pseudogenes). However, when a large number of introns in GenBank DNA sequence database were examined, making use of the great progress of genome projects in recent years, a series of unexpected distributions of intron phases were discovered.

First, the proportions of the three intron phases were significantly not equal (Fedorov et al. 1992; Long, Rosenberg, and Gilbert 1995; Tomita, Shimizu, and Brutlag 1996). Phase 0 is the most abundant (~50%), followed by phase 1 (~30%), with phase 2 being the least abundant (~20%). Although new introns have continually been added to the databases, new analyses always reveal a similar distribution (see Long and Deutsch 1999; Sakharkar et al. 2000). Second, more interestingly, multiple introns within a gene showed a significant correlation with respect to the association of their phases. Exons flanked by introns of the same phase significantly outnumbered those predicted based on random association of intron phases, a condition termed ‘‘symmetric exon excess’’ (Long, Rosenberg, and Gilbert 1995; Tomita, Shimizu, and Brutlag 1996). In this correlation, the symmetric exons flanked by phase 1 introns ((1, 1) exons) always showed higher excess than the other two symmetric exons, (0, 0) and (2, 2), in accordance with the observation that most of the identified cases of exon shuffling involved the same (1, 1) exons (Patthy 1995). Finally, the excess phase 0 introns and excess symmetric exons were also observed in ancient conserved regions (ACRs; Green et al. 1993), suggesting that the same mechanism creating the distinctive distribution of intron phases also worked in such regions of ancient genes.

These observations show that distribution of intron locations within the coding sequences is nonrandom and thus reject a simple form of the insertional hypothesis of intron origin. However, the observed phenomena might be interpreted as a result of intron insertions into nonrandomly distributed proto-splice sites. For this hypothesis, the issue is whether or not the distribution of the phases of proto-splice sites is similar to that of intron phases or, more strictly, whether or not the observed intron phase distribution is a randomly sampled subset of the total proto-splice site distribution. Long et al. (1998) tested this hypothesis by investigating whether or not the observed proportions of intron phases were consistent with the phase proportions of hypothetical proto-splice sites as predicted by dicodon distributions in humans and other organisms. No consistency was found between the distribution of the three intron phases and the phase distribution of proto-splice sites, thus negating any explanatory power of present-day proto-splice sites with regard to the nonrandomness in phase proportions.

In this paper, we extend our analysis of proto-splice sites from the first observation to the second phenomenon of intron phase correlation. We take a hypothesis test approach to investigate the validity of the proto-splice site model. In this text, the random insertion of introns into nonrandomly distributed proto-splice sites is the null hypothesis that serves as a basis to calculate the probability of observed intron phase correlation. We first ask if there are phase correlations among adjacent proto-splice sites. Then, we test whether or not the proto-splice sites can explain the correlation of intron phases as observed. Finally, we simulate a process of intron insertion into proto-splice sites and ask how often such an insertional process can generate observed distribution of intron phases. We will show that the proto-splice site model cannot yield any distribution resembling observed intron phase correlations.

Materials and Methods

General Approach

Based on the sequence database GenBank, we first constructed an exon database that contained coding sequence (CDS). Then, we calculated the phases and positions of introns based on the information in annotation feature tables for each gene. This would give two distributions: the frequencies for proportions of three intron phases $f(i)$ (i = 0, 1, and 2) and the frequencies for the nine states (possible (5′, 3′) pairs) of intron phase associations within genes. We then calculated phases of various hypothetical proto-splice sites by scanning the CDS for each gene and created a gene structure delineated by the phases and positions of proto-splice sites. Because the phases of proto-splice sites were viewed as the phases of potential but unrealized insertion sites of introns, we called these calculated phases ‘‘pseudo-intron phases’’ (the corresponding coding region between two adjacent proto-splice sites [pseudo-introns] was called a ‘‘pseudo-exon’’). Like the analysis of real introns, this would also generate two distributions: the proportions of pseudo-intron phases and the association of pseudo-intron phases. We tested the statistical difference of intron phase associations from pseudo-intron phase associations as expectations using G-tests (Sokal and Rohlf 1995). Moreover, under the hypothesis that the introns we see today are the result of insertion into proto-splice sites, the distribution of today's introns should be a random sample drawn from the total distribution of proto-splice sites. We carried out Monte Carlo simulation to test this hypothesis by calculating the probability of observed intron phase distribution over all randomly drawn intron phases as defined by proto-splice sites.

Exon Database with CDS Sequence

Using a computer program similar to one we wrote to construct an exon database (Long, Rosenberg, and Gilbert 1995), we collected the entries of all intron-containing genes in GenBank release 114 to form a raw database including the DNA sequence for each locus. We then wrote a program to filter out all questionable entries, such as pseudogenes and entries with inconsistent feature tables. We then developed a final exon database in which we calculated the positions and phases

```
$ID 221
LOCUS       AF006573     5770 bp    DNA             INV       31-OCT-1997
DEFINITION  Drosophila virilis maltase 1 (Mav1) and maltase 2 (Mav2) genes,
cds no: 1
ORGANISM  Drosophila virilis
Proto AG|GT(No: 6, phase:020010 position:120,143,375,1449,1558,1710)

1-atgagtttgaagtggagcttagttttgggcctaagttggctcctttttgtggcttcgtcagaattaaaaaagcataag
ccaaacgagttggacgacaatattaattggtggcgacacgaggtcttttatcagatttatcctaggtcctttaaggac
agcgatggcgatggcattggtgatcttaagggcatcacctccaagctgcagtactttgtggacactggcatcacggccat
ctggctaagtcccatttacaagtcacccatggttgactttgggtacgatatatccgactacagggacatccagccggagt
atggcaccctggaggactttgatgcgctgatcgccaaggccaatcaactgggcatcaaggtcattttggactttgtgcc
c(396).....(1441)cattataaggtctatcagtcgctgatcaagctgagacaatcgcgagtcttgcgggatggctcatt
tacagcccaggcccttaatcgtaatgtttttgctattaagcgcgaattgagaggtcagcccaccctgctaactgtcatt
aatgtgagcaatcgcacccagcaagtcgatgtcagtaactttatcgatttgcccaatcgtcttacattgttggttgtggg
cgtttgctcccagcatagggtgagcgagcgccttaagcccgccgaggtcaaactgtcgccccatgagggtctagttata
cagctcaaagctcgctag-1761
```

FIG. 1.—An example of coding sequence (CDS), with marked AGGT sites along the sequence of *Drosophila virilis* maltase 1. The six AGGT sites are distributed in 5′ and 3′ portions of the CDS; the middle portion (397–1440) of the sequence is not shown. The phases of these six hypothetical proto-splice sites were calculated as 020010 from the 5′ to the 3′ end. The positions of these sites are calculated and shown in the feature line (the sixth line, "Proto AG|GT(No: . . .)").

of introns and the positions and phases of proto-splice sites by scanning the entire CDS for each proto-splice site. An example for the phases and positions of pseudo-introns is given in figure 1. We also calculated other statistical parameters, such as the lengths of exons and proteins and the protein sequences. In this database, we also constructed CDSs based on the information of feature tables. The major computing challenges were that very often the sequences of some exons as defined in one feature table are in different entries. We wrote a program to collect all of those entries that contained the sequence of a single exon into a subdatabase. Then, when we generated CDSs, the computing process automatically visited the subdatabase to fetch the exon sequence to form a complete CDS sequence.

Purging Redundancy from the Database

Because a large proportion of sequences in the database were redundant, as shown by previous work (Long, Rosenberg, and Gilbert 1995; Long and Deutsch 1999; Rubin et al. 2000), we purged the redundant sequences to avoid possible bias from that source. We used the program GBPURGE (Falling Rain Genomics, Inc., Palo Alto, Calif.) to group the homologous proteins and then kept only one sequence in each protein family (Long, Rosenberg, and Gilbert 1995). The purging process was carried out in an alpha workstation (DIGITAL). The threshold we set to judge homology was 20% identity as calculated by FASTA3 (Pearson 2000): when two protein sequences were compared, if the identity was ≥20%, we deleted the sequence that had fewer introns and kept the other one.

Proto-splice Sites and Pseudo-intron Phases

Following our previous study (Long et al. 1998), we chose four hypothetical sites, G|G, AG|G, AG|GT, and C/AAG|R, where "|" indicates a possible insertion site and "/" indicates two alternative states of one nucleotide site. These sites were chosen either because they were suggested by the proponents of proto-splice sites or because they had higher frequencies for real introns in the databases.

Distributions of Intron Phases and Pseudo-intron Phases

Both intron phases and pseudo-intron phases will generate two separate distributions: the proportions of the three phases and the association of phases within genes, $f(i, j)$, where $(i, j) = (0, 0), (1, 1), (2, 2), (0, 1), (0, 2), (1, 2), (1, 0), (2, 0)$, or $(2, 1)$. As shown previously (Long, Rosenberg, and Gilbert 1995; Tomita, Shimizu, and Brutlag 1996; Fedorov et al. 1998), the distribution of intron phases shows significant bias toward phase 0 introns and excess symmetric exons. The frequencies of the two sets of phase associations of introns and pseudo-introns were compared using a likelihood ratio test (*G*-test; Sokal and Rohlf 1995),

$$G = 2\sum_{i=0}^{2}\sum_{j=0}^{2} O(i, j)\ln\frac{O(i, j)}{P(i, j)},$$

where $O(i, j)$ and $P(i, j)$ are the frequencies of the associations of introns and pseudo-introns, respectively.

Monte Carlo Simulation

A Monte Carlo simulation was performed as a direct statistical test of the model in which the real introns in each gene are a result of random insertion into proto-splice sites. We generated all possible pseudo-introns in each gene in the purged database, then randomly targeted these sites once for each real introns in the gene. Figure 2 gives an example of the molecular model of intron insertion into proto-splice sites in the simulation process. Having treated all the genes from the purged database, we created a comparable array of virtual genes, for which we then calculated the distribution of the simulated pseudo-intron phases and associations.

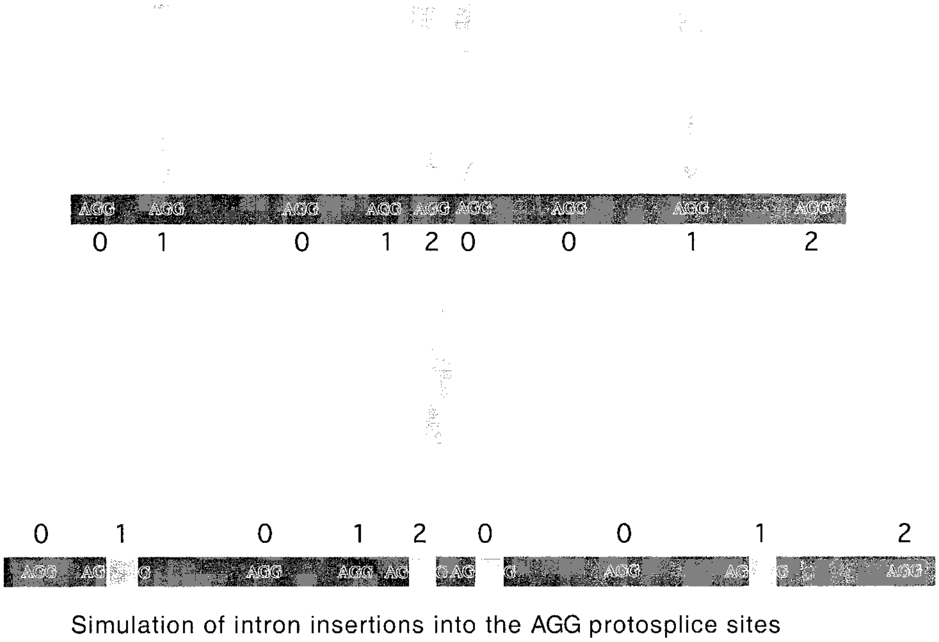
AGG
0
1
2
Simulation of intron insertions into the AGG protosplice sites

FIG. 2.—The molecular model for insertion of introns into AGG sites in the simulation. The two black boxes indicate a CDS, with the phases of AGG shown in nine positions. This hypothetical CDS is assumed to contain four introns initially. In the simulation, the four introns, as shown in the gray boxes, are randomly inserted into the four AGG sites between the two G nucleotides in the sites and form a virtual gene with four introns of phases 1201 from the 5′ to the 3′ end.

One hundred thousand such arrays and distributions were generated. In each array, we measured the difference between the observed intron phase associations and the distribution predicted by the insertion model using

$$\mathbf{X}^2 = \sum_{i=0}^{2}\sum_{j=0}^{2}\frac{(O(i, j) - S(i, j))^2}{S(i, j)},$$

where $O(i, j)$ and $S(i, j)$ are observed and simulated frequency of association (i, j) between two adjacent introns or pseudo-introns. The frequency of pseudo-intron association $S(i, j)$, normalized to the total number of observed intron associations, was calculated in two ways: (1) by direct counts from simulated pseudo–exon-intron structures; (2) as the product of the proportions of two pseudo-introns, i and j.

We then generated a frequency distribution of the obtained values of X^2 to describe the dissimilarity between all simulated results and the observed intron associations. Meanwhile, we also calculated the probability of the statistical patterns of the observed relative excess of symmetric pseudo-exons. The pattern was described by

$$\mathbf{R}(0, 0) > 0.05 \cap \mathbf{R}(1, 1) > \mathbf{R}(0, 0) + 0.05$$
$$\cap\ \mathbf{R}(1, 1) > \mathbf{R}(2, 2) + 0.05 \cap \mathbf{R}(2, 2)$$
$$> 0.05 \cap \mathbf{F}(i, j) < \mathbf{E}(i, j)$$

for $(i \neq j)$, where (1) $\mathbf{R}(i, i) = (\mathbf{F}(i, i) - \mathbf{E}(i, i))/\mathbf{E}(i, i)$, $\mathbf{R}(i, i)$ is the measurement of the excess of the (i, i) type of symmetric pseudo-exons; (2) $\mathbf{E}(i, j) = \mathbf{F}(i) \times \mathbf{F}(j) \times \mathbf{N}$, $\mathbf{E}(i, j)$ is the expected frequency of the (i, j) pseudo-exon and $\mathbf{F}(\mathrm{i})$ and $\mathbf{N}$ are the observed proportion of pseudo-intron i and total internal exon number, respectively; and (3) the logic sign "∩" means that the given conditions are met simultaneously.

This is a very conservative test of symmetric phase description. Actual observed excess of symmetric exons is just a portion of the excess measured here, and the observed excess of (0, 0) exons is higher than 0.05. The differences in the excesses between (0, 0) and (1, 1) and between (0, 0) and (2, 2) are also larger than 0.05, which we used in the computing. Simulation was carried out in the UNIX environment of the alpha workstation, where the function drand48( ) was used as a random number generator.

Results

We collected 53,542 intron-containing sequences from a recent version of GenBank (release 114). After purging redundant sequences using GBPURGE, we created a final exon database that contained 12,805 independent or quasi-independent sequences (the identity among any two protein sequences was lower than the threshold, 20%). We calculated the phases and positions of introns and pseudo-introns defined by four types of proto-splice sites using the CDS DNA sequence of each of the 12,805 genes.

The proportions of three intron phases and frequencies of nine associations of introns showed significant nonrandom distribution (table 1). Similar to previous version of the database, there are unequal intron proportions (48% phase 0, 28% phase 1, and 24% phase 2; $G = 7{,}787$, $P \ll 10^{-100}$). The arrangements of intron phases within genes showed strong correlation: all symmetric exons, (0, 0), (1, 1), and (2, 2), showed significant excess over a random prediction ($G = 867$, $P = 7.4 \times$

Table 1
Observed and Expected Symmetric and Asymmetric Exons

	Symmetric Exons			Asymmetric Exons						
	(0, 0)	(1, 1)	(2, 2)	(0, 1)	(0, 2)	(1, 2)	(1, 0)	(2, 0)	(2, 1)	Exon No.
Observed	18,492	7,205	5,021	8,584	8,301	4,848	8,958	8,406	4,742	74,557
Expected	16,881	5,848	4,443	9,936	8,660	5,097	9,936	8,660	5,097	
Excess (%)	10	23	13							

Note.—G = 867, $P = 7.4 \times 10^{-182}$. The excess is measured as (observed − expected)/expected. The expected number of intron phase association (i, j) = $f(i) \cdot f(j) \cdot N$, where $f(i)$ and N are the frequency of intron phase i and the total number of internal exons, respectively.

10^{-182}), with (1, 1) showing the highest relative excess (23%). These observations are consistent with previous reports (Long, Rosenberg, and Gilbert 1995; Tomita, Shimizu, and Brutlag 1996; Fedorov et al. 1998). Our previous investigation (Long et al. 1998), which focused on the relationship between intron phase proportions and the proto-splice sites, rejected the model that the proto-splice sites could predict the distribution of intron phase proportions. This report focuses on a similar analysis of associations of proto-splice sites as predictors of the intron phase correlation, i.e., the correlation analysis of pseudo-intron phases.

The analysis of pseudo-intron phases revealed an interesting phenomenon: the pseudo-introns are not randomly scattered within genes, and for some symmetric exons there are excesses over random predictions (table 2). For all proto-splice sites, G values that measure the difference between observed distribution and expected distribution are highly significant (in the smallest one, $G = 7{,}534$ with $P \ll 10^{-100}$). Furthermore, in AG|G and C/AAG|R sites, all three symmetric pseudoexons showed excess over expected occurrences. This phenomenon, previously unknown, suggests that some adjacent proto-splice sites seem to prefer particular phase associations, especially symmetric pseudo-exons for AG|G and C/AAG|R sites. However, can these distributions be interpreted as the cause of the intron phase correlation?

The patterns of distributions of symmetric and asymmetric exons in both G|G and AG|GT were different from the observed intron phase association: the former showed excess in some symmetric exons and some asymmetric pseudo-exons, while the latter had excesses in all of the symmetric exons and none of the asymmetric exons. Table 2 shows that G|G proto-splice sites do not show significant correlation of symmetric phases. (0, 0) exons had 2% excess, (1, 1) exons had 6% excess, and (2, 2) exons had −12% deficiency. In its asymmetric phase pairs, (0, 1), (1, 2) and (2 ,0) showed higher excess over expectation. Thus, the distribution of G|G site-defined pseudo-intron phases did not share any similarity with the intron phase correlations. Similarly, the pseudo-intron phase distribution defined by AG|GT did not show excesses in two symmetric pseudo-exons, with (0, 0) showing −8% deficiency compared with its expectation.

AG|G and C/AAG|R seem better candidates because all three of their symmetric pseudo-exons have

Table 2
Observed and Expected Symmetric and Asymmetric Pseudo-exons

	Symmetric Pseudo-exons			Asymmetric Pseudo-exons						
	(0, 0)	(1, 1)	(2, 2)	(0, 1)	(0, 2)	(1, 2)	(1, 0)	(2, 0)	(2, 1)	Exon No.
G|G proto-spice sites										
Observed	187,423	147,810	42,647	181,091	69,149	113,403	120,527	126,901	54,706	1,043,657
Expected	183,504	139,515	48,270	160,005	94,116	82,064	160,005	94,116	82,064	
Excess (%)	2	6	−12							
Compared with observed intron associations: G = 9,321, very significant										
AG|G proto-splice sites										
Observed	74,605	35,572	12,488	35,126	21,826	15,574	36,278	21,874	14,239	267,582
Expected	66,190	27,357	8,951	42,553	24,341	15,648	42,553	24,341	15,648	
Excess (%)	13	30	40							
Compared with observed intron associations: G = 3,680, very significant										
AG|GT proto-splice sites										
Observed	15,077	10,270	3,569	6,403	10,375	4,995	10,420	10,498	4,191	75,798
Expected	16,437	10,766	1,879	13,302	5,558	4,498	13,302	5,558	4,498	
Excess (%)	−8	−5	90							
Compared with observed intron associations: G = 4,100, very significant										
C/AAG|R proto-splice sites										
Observed	87,104	98,943	20,061	69,870	28,923	34,174	69,257	29,847	33,076	471,155
Expected	73,474	86,377	14,749	79,665	32,919	35,693	79,665	32,919	35,693	
Excess (%)	19	14	36							
Compared with observed intron associations: G = 13,924, very significant										

Note.—$P \ll 10^{-100}$ for all G values. Excess measured as (observed − expected)/expected. The expected number is $E(i, j)$ (see *Materials and Methods*).

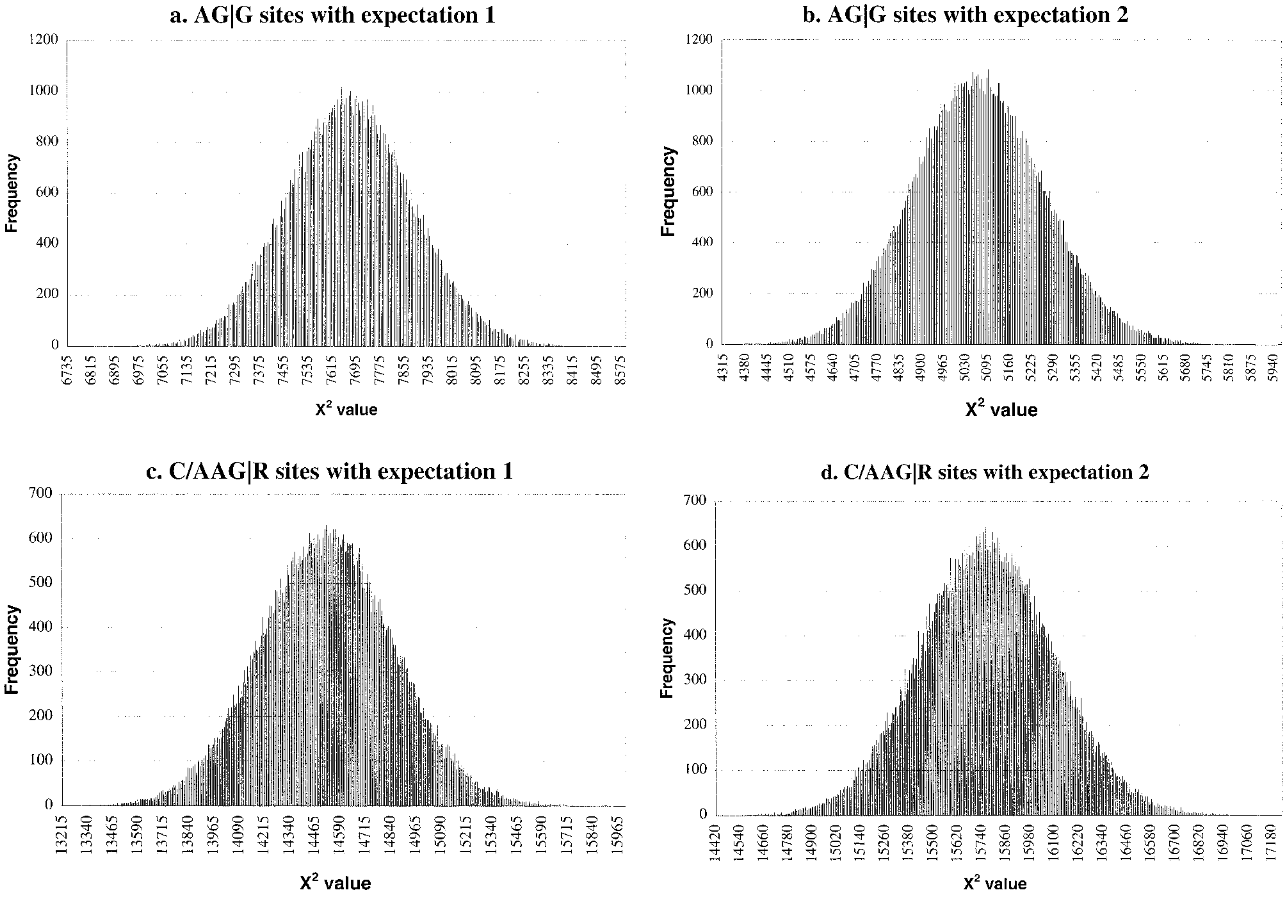

FIG. 3.—Monte Carlo simulations generate the distributions of dissimilarity of the phase associations between the real introns and those expected from insertions into proto-splices sites. *a* and *b*, AGG sites with two expectations. *c* and *d*, AGG sites with two expectations. Expectation 1 includes the frequencies of nine associations of the phases; expectation 2 includes the frequencies calculated from the product F(*i*) × F(*j*) × *l*, where F(*i*) is the phase proportion of phase *i* in *l* virtual genes, where *l* is set to the number of introns in the purged database.

excesses and all their six asymmetric pseudo-exons have deficiencies compared with expectations. However, both sites show conflicting patterns of symmetric pseudo-exons: (2, 2) pseudo-exons show the highest excesses among three symmetric types, 40% for AG|G sites and 36% for C/AAG|R. This differs from intron phase associations which show an excess pattern, $(0, 0) < (1, 1) > (2, 2)$, i.e., (1, 1) pseudo-exons having highest excess, consistent with observed cases of exon shuffling (Patthy 1995).

The difference as shown in pattern analysis was further supported by simple statistical comparison between the phase distributions of introns and pseudo-introns. All proto-splice sites showed significant difference (the smallest *G*, from AG|G sites, was 3,680, with $P \ll 10^{-100}$). Thus, the phase distributions as generated by proto-splice sites were not the same as the distribution of intron phases. Besides the statistical comparison between observed intron phase association and the overall distribution of pseudo-intron phase associations, another biologically more sensible statistical test was developed by directly simulating the process of intron insertion into proto-splice sites under the hypothesis that introns are results of insertion into preexisting proto-splice sites.

On average, the numbers of most hypothetical proto-splice sites (G|G, AG|G, and C/AAG|R) in each CDS were 4–15 times as numerous as introns (3.7/kb CDS; Deutsch and Long 1999). This allows randomization of intron positions among proto-splice sites in each gene in computer simulation processes. Each randomization of the available introns in each gene generated one set of outcomes following the simulated insertions into a portion of proto-splice sites. Then, we investigated the associations of the pseudo-intron phases in each resulting set of outcomes and compared them with the observed intron associations.

In the program written for the simulation process, each simulation experiment began with the first gene in the database. The number of introns in this gene, *n*, was counted before the random number generator was called to randomly assign introns into *n* of the *m* previously defined proto-splice sites ($m > n$). Then, a string of pseudo-intron phases and their positions was calculated from the assignment. After this process had reached the last gene of the database, we had one new ‘‘database’’ with a set of randomly inserted positions. The associations of pseudo-intron phases for this set of randomized genes was then calculated and compared with the observed intron phase associations in the ‘‘real’’ database.

We repeated this simulation experiment 100,000 times and generated a frequency distribution of statistical comparisons.

Figure 3*a–d* plots the frequency distributions of X^2 measures for the comparisons of all 100,000 simulated data sets. These distributions actually contain two sets of comparisons. First, for a particular X^2 value that contains a particular expectation from one sampling of proto-splice sites, the standard χ^2 distribution can be used to test whether the observed intron phase correlation is significantly different from that expectation. Second, by examining the frequency distributions of X^2 values, we ask how often, in general, the observed intron phase correlation can be found in those 100,000 simulated data sets.

The lowest X^2 value is 4,315 (fig. 3*b*), corresponding to a probability level of $P \ll 10^{-100}$ in a χ^2 distribution with 8 df. This indicates that of all those random insertions of hypothetical introns into proto-splice sites, none yielded an expected phase distribution that was similar to the observed intron phase correlations. Thus, both statistics, frequency of pseudo-intron association and the product of pseudo-intron proportions, reject the null hypothesis that the observed intron phase correlations can be a subset of inserted introns into proto-splice sites.

It should be noted that, given the huge sample size in this investigation, this test might have used too strict a criterion of fit to the hypothesis. A more conservative test was conducted by directly examining particular patterns of intron phase correlation. However, when using a very conservative test to calculate the probability of the observed pattern of excess symmetric exons (defined as $\mathbf{R}(0, 0) > 0.05 \cap \mathbf{R}(1, 1) > \mathbf{R}(0, 0) + 0.05 \cap \mathbf{R}(1, 1) > \mathbf{R}(2, 2) + 0.05 \cap \mathbf{R}(2, 2) > 0.05 \cap \mathbf{F}(i, j) < \mathbf{E}(i, j)$; $i \neq j$; see *Materials and Methods*), none of 100,000 simulations generated an excess of symmetric pseudo-intron associations that fell into the broad range of excess patterns including the observed intron phase associations ($P < 10^{-5}$).

Discussion

Our statistical questions are, first, whether or not the proportions of pseudo-intron phases show a biased distribution toward phase 0 introns; second, whether the association of pseudo-intron phases are correlated; and third, if the association of pseudo-introns are correlated, can such a correlation yield the correlations of the intron phases? While the first question has been addressed in Long et al. (1998), two observations in this investigation are remarkable. First, we found that some proto-splice sites showed a degree of autocorrelation with respect to their phases. Second, the simulated insertion model generated significantly different distributions of pseudo-intron phases from the observed intron phases.

We observed that the arrangement of proto-splice sites is nonrandom with respect to their phases. At the first glimpse, the phase correlations of some proto-splice sites seem to mimic the correlation of intron phases, suggesting the possibility that the excess symmetric exons might be a consequence of intron insertions into the correlated proto-splice sites. Upon closer inspection, the sites that show phase correlations are those that have very low conservation, like any random sites in the genomes, and thus cannot be taken as vestigial candidates for insertions. The site that has highest conservation among those tested sites, G|G, does not show any significant correlations.

The nonrandom distribution of proto-splice sites provides a unique opportunity to model a hypothetical process of intron insertion. The Monte Carlo simulation for such a model showed a significantly low probability that proto-splice sites underlie the distribution of intron phases. This is consistent with the statistical test of the difference between the distributions of intron phases and proto-splice sites. These two statistical tests rejected the model of proto-splice sites. Finally, we tried several candidates for proto-splice sites. When more sites, albeit randomly selected ones, are taken as proto-splice sites, one always can find, by chance, some sites that show some similarity to any statistical pattern. For example, we tried AT|T and TT|C, which also show some similarity.

The three introns in Xdh found by Tarrio et al. (1998) were viewed as evidence for the proto-splice site model of intron gains (Logsdon et al. 1998). Two clouds weaken this argument. First, the definition of the proto-splice sites is ad hoc. Altogether, four sites were defined, CAG|G, AAG|G, GAA|A, and TCN|G (N refers to any of the four nucleotides, A, T, G, or C). The last two sites have nothing to do with known conserved sequences surrounding splice sites, although the first two sites have some similarity to the consensus sequence of splice sites. Dibb and Newman (1989) proposed that proto-splice sites existed in intron-lacking ancestral genes. This proposal may offer a specific prediction for distribution of the insertion sites: the flanking exon sequence motif should be in both intron-containing and intron-absent sequences. This concept, however, also predicts some conservation in the proto-splice sites as a recognition signal for intron insertion. This criterion is violated in this case. Second, the argument that these introns are recently acquired is based on a standard but questionable approach of phylogenetic distribution. In this line of logic, once an intron appears in a small number of branches in a tree, by what is thought of as a parsimony approach, this intron is viewed as a recent gain. However, considering the biological reality that intron loss may occur much more frequently than intron gain, this approach is not justified. For example, introns in the whole genome of *Saccharomyces cerevisiae* are almost all lost by gene conversion (Fink 1987). Thus, it is not reasonable to dismiss the alternative hypothesis that many lines of independent intron loss are more likely than a single intron insertion that may bring deleterious effects to the target genes.

Finally, one cause can be inferred for the distributions of psuedo-intron phases: the repetition of amino acid residuals with particular dicodon and codon usage, which we found to contribute to the correlation of some pseudo-intron phases. For AG|G proto-splice sites, for example, a string of glutamic acids, Glu.Glu.Glu.Glu, encoded by gag.gag.gag.gag, will create two (0, 0) symmetric pseudo-exons. This scenario was supported by the biased dicodon usage of gag.gag, which is highest among all dicodons in mammals, plants, and invertebrates (1,000-fold higher than the lowest dicodon usages) (Long et al. 1998) and thus will make a contribution to the (0, 0) symmetric exons. In fact, when we deleted all adjacent amino acid repeats, we found that the excess of symmetric exons significantly dropped. Similarly, particular nonadjacent amino acid repeats and codon usages can also contribute to particular symmetric and asymmetric pseudo-exon distribution.

This investigation, along with a previous companion study, showed that the actual distribution of proto-splice sites in eukaryotic genes differs significantly from the distribution of intron phases. This rejects the proto-splice sites model as a null hypothesis to account for the unique distribution of intron phases. Alternatively, the best explanation for such distribution is that a large amount of exon shuffling, as predicted by the exon theory of genes (Gilbert 1987), created excess symmetric exons and overrepresented phase 0 introns (Long, Rosenberg, and Gilbert 1995; Long et al. 1998), because observed patterns of symmetric and asymmetric exons are consistent with observed cases of exon shuffling (e.g., Patthy 1991, 1995, 1999).

Acknowledgments

We thank the members of the Long laboratory and Walter Gilbert (Harvard University) for valuable discussions. Thanks to Ana Llopart and Ether Betran for their help in the preparation of the manuscript. This work was supported by the National Science Foundation (MCB-9977990) and the Packard Foundation (a Packard Fellowship in Science and Engineering).

LITERATURE CITED

BURGE, C. B., T. TUSCHL, and P. A. SHARP. 1999. Splicing of precursors to mRNAs by the spliceosomes. Pp. 525–560 *in* R. F. GESTELAND, T. R. CECH, and J. F. ATKINS, eds. The RNA world. 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

DEUTSCH, M., and M. LONG. 1999. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. **15**:3219–3228.

DIBB, N. J., and A. J. NEWMAN. 1989. Evidence that introns arose at proto-splice site. EMBO J. **8**:2015–2022.

FEDOROV, A., L. FEDOROVA, V. STARSHENKO, V. FILATOV, and E. GRIGOR'EV. 1998. Influence of exon duplication on intron and exon phase distribution. J. Mol. Evol. **46**:263–271.

FINK, G. R. 1987. Pseudogenes in yeast? Cell **49**:5–6.

GILBERT, W. 1987. The exon theory of genes. Cold Spring Harb. Symp. Quant. Biol. **52**:901–905.

GREEN, P., D. LIPMAN, L. HILLIER, R. WATERSTON, D. STATES, and J. M. CLAVERIE. 1993. Ancient conserved regions in new gene-sequences and the protein databases. Science **259**: 1711–1716.

LEE, V. D., M. STAPLETON, and B. HUANG. 1991. Genomic structure of Chlamydomonas caltractin: evidence for intron insertion suggests a probable genealogy for the EF-hand superfamily of proteins. J. Mol. Evol. **221**:175–191.

LOGSDON, J. M. 1998. The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. **8**:637–648.

LOGSDON, J. M., A. STOLTZFUS, and W. F. DOOLITTLE. 1998. Molecular evolution: recent cases of spliceosomal intron gain? Curr. Biol. **8**:R560–R563.

LONG, M., S. J. DESOUZA, and W. GILBERT. 1997. The yeast splice site revisited: new exon consensus from genomic analysis, Cell **91**:739–740.

LONG, M., S. J. DESOUZA, C. ROSENBERG, and W. GILBERT. 1998. Relationship between ''proto-splice sites'' and intron phases: evidence from dicodon analysis. Proc. Natl. Acad. Sci. USA **94**:219–313.

LONG, M., and M. DEUTSCH. 1999. Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol. Biol. Evol. **16**:1528–1534.

LONG, M., C. ROSENBERG, and W. GILBERT. 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. Proc. Natl. Acad. Sci. USA **92**:12495–12499.

MOORE, M. J., C. C. QUERY, and P. A. SHARP. 1993. Splicing precursors to messenger RNAs by the spliceosome. Pp. 303–357 *in* R. GESTELAND and J. ATKINS, eds. The RNA world. 1st edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

NEWMAN, A. J., and C. NORMAN. 1991. Mutations in yeast U5 snRNA alter the specificity of 5′ splice-site cleavage. Cell **65**:115–123.

———. 1992. U5 snRNA interacts with exon sequences at 5′ and 3′ splice sites. Cell **68**:743.

PATTHY, L. 1991. Modular exchange principles in proteins. Curr. Opin. Struct. Biol. **4**:351–361.

———. 1995. Protein evolution by exon shuffling. Springer-Verlag, New York.

———. 1999. Genome evolution and the evolution of exon-shuffling. Gene **238**:103–114.

PEARSON, W. R. 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. **132**:185–219.

RUBIN, G. M., M. D. YANDELL, J. R. WORTMAN et al. (55 co-authors). 2000. Comparative genomics of the eukaryotes. Science **287**:2204–2215.

SAKHARKAR, M., M. LONG, T. W. TAN, and S. J. DE SOUZA. 2000. ExInt: an exon/intron database. Nucleic Acids Res. **28**:191–192.

SOKAL, R. R., and F. J. ROHLF. 1995. Biometry. 3rd edition. Freeman, New York.

TARRIO, R., F. RODRIGUEZ-TRELLES, and F. J. AYALA. 1998. New Drosophila introns originate by duplication. Proc. Natl. Acad. Sci. USA **95**:1658–1662.

TOMITA, M., N. SHIMIZU, and D. L. BRUTLAG. 1996. Introns and reading frames: correlation between splicing sites and their codon positions. Mol. Biol. Evol. **13**:1219–1223.

TREISMAN, R., N. J. PROODFOOT, M. SHANDER, and T. MANIATIS. 1982. A single-base change at a splice site in a beta 0-thalassemic gene causes abnormal RNA splicing. Cell **29**: 903–911.

ANTONY DEAN, reviewing editor

Accepted August 7, 2000