Intron phase correlations and the evolution of the intron/exon structure of genes
(exon shuffling/ancient conserved regions)

MANYUAN LONG*, CARL ROSENBERG†, AND WALTER GILBERT*

*Department of Molecular and Cellular Biology, The Biological Laboratories, Harvard University, Cambridge, MA 02138; and †Whitehead Institute/ Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02139

Contributed by Walter Gilbert, September 13, 1995

ABSTRACT Two issues in the evolution of the intron/exon structure of genes are the role of exon shuffling and the origin of introns. Using a large data base of eukaryotic intron-containing genes, we have found that there are correlations between intron phases leading to an excess of symmetric exons and symmetric exon sets. We interpret these excesses as manifestations of exon shuffling and make a conservative estimate that at least 19% of the exons in the data base were involved in exon shuffling, suggesting an important role for exon shuffling in evolution. Furthermore, these excesses of symmetric exons appear also in those regions of eukaryotic genes that are homologous to prokaryotic genes: the ancient conserved regions. This last fact cannot be explained in terms of the insertional theory of introns but rather supports the concept that some of the introns were ancient, the exon theory of genes.

How did the intron/exon structure of genes evolve? Broadly speaking, there are two alternative theories for the origin of introns. The insertional theory of introns, also called the "introns-late" theory, invokes recent events of intron insertion into eukaryotic genomes (1–3). On the other hand, the "introns-early" hypothesis (4–7) takes the position that introns were used in the progenote to assemble genes and that the evolutionary pattern since was one of loss down many lines, some retention, and, possibly, some gain. On this model one expects some correlation of exons with functional elements of the protein gene products.

The introns-late hypothesis assumes that introns were added to preexisting genes at some time during evolution. One form would suggest that introns were added early in eukaryotic evolution and were used to shuffle parts of genes to account for the Cambrian explosion of metazoan species. Such models would expect exon shuffling to appear only in evolutionarily late gene products. A still more extreme model is that introns were added only very late in evolution and that no exon shuffling has occurred. In all introns-late models one expects no correlation between exons and elements of protein function for ancient genes, and observations of correlations between three-dimensional structure and exon structures (8–10), of correlations between intron positions in plants and animal genes (11–15), or of correlations between intron positions in genes that diverged in the progenote (16–19) are all viewed as statistical coincidences (20–22).

This study began with a search for different testable predictions derived from these two competing theories. Intron phase [the position of the intron within a codon, phase 0, 1, or 2 bying before the first base, after the first base, or after the second base, respectively (23)], is an important evolutionary character for whom the distribution the two theories generate different predictions. Since the insertion of introns into pre-

viciously uninterrupted genes, as the introns-late view advocates, does not affect the structure of the gene product, introns in each phase should have similar chances to survive. Thus, the simple form of the introns-late theory predicts a random phase distribution for introns. On the contrary, the introns-early theory predicts a nonrandom phase distribution for introns as a consequence of exon shuffling events, since exon shuffling works better if introns are in the same phase. Furthermore, exon shuffling should produce correlations in intron phases, since symmetric exons shuffle more easily, while insertional models predict that intron phases are uncorrelated. The prediction of a random distribution of intron phases and random correlations allows us to detect exon shuffling and to test the insertional theory of intron origins. In this report, we show that this population analysis indicates both an important role for exon shuffling throughout evolution and also support for the hypothesis of the ancient origin of introns.

METHODS

Exon Data Bases. We used GenBank (24) entries released in 1994 [release 84, which also includes European Molecular Biology Laboratory (EMBL) and DNA Database of Japan (DDBJ) entries] to construct a raw data base that contains information about the intron/exon structures, definitions of genes, locus names, and, finally, protein sequences. Then we calculated the intron positions and phases and the numbers and sizes of the exons for these 9276 sequences (45,095 exons) to create a processed data base. We discarded entries with obvious problems, such as wrong feature tables. In testing our extraction program, we identified all entries in GenBank that had the word exon in the definition line and sampled randomly about 10% of these entries to check how many include the "cds...join" and "cds...complement join" command that we used to identify exon-containing genes. We found that about 3% of exons lacked this reference, and we excluded them. A further 5% lacked peptide sequence and were discarded. Therefore, we recovered about 97% of all annotated genes in the data base. To discard duplicates and partial sequences, we compared all the protein sequences with BLASTP (25) and purged all related entries but the longest, using a similarity score of 99% (99% match over the length of the shorter sequence), yielding 6611 genes and 33,150 exons. We then further distilled the data base, using FASTA (26) (similarity scores are calculated from the FASTA alignment as the match percentage times the length of the overlapping region divided by the length of the shorter sequence) and purging down to a criterion of a 20% match to the shorter sequence, keeping the sequence with more exons each time (the purging yielded 4681 genes at the 80% level, 3704 at 60%, 2942 at 40%, and 1925 at 20%). The plant genes are from FASTA purging of a complete data base of plant intron-containing genes to the 20% simi-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.
RESULTS

Construction of Exon Data Base. To survey the intron phase distribution, we first built a data base of eukaryotic intron-containing genes extracted from GenBank release 84. To discard duplicates and partial sequences in this data base, we purged all related sequences by using BLASTP and FASTA. The final data base, purged to 20%, contains 1925 independent or quasi-independent protein sequences containing 13,042 exons and 11,117 introns (we use the prefix quasi here because one cannot infer with certainty that the sequences are truly independent; these low criteria for similarity still miss a few related genes). We found 1600 genes that have more than one intron, and there are 9192 internal exons. Fig. 1 shows a histogram of internal exon lengths: a distribution peaked at 35 residues.

Distribution of Intron Phases. The three classes of intron phases are far from evenly distributed. Table 1 shows that there is an excess of phase zero introns, as was found in a previous survey of a small sample (28). While this result from a large sample is consistent with the prediction of the exon theory of genes and contradicts the simple form of the insertional theory, there are alternative hypotheses for the insertional theory which cannot be rejected. To disfavor the appearance of phase one or phase two introns, one might assume that the insertion was into long nucleotide sequences which might be correlated with the local amino acid sequence of the gene product, such as has been suggested for an AG/GT sequence at the exon boundaries (28). However, we can distinguish between the two theories by studying intron phase correlations.

Intron Phase Correlations. The association of two adjacent introns in eukaryotic genes can be in any of nine different intron phase combinations, which can be classified into two groups—symmetric exons (0, 0), (1, 1), and (2, 2) and asymmetric exons (0, 1), (0, 2), (1, 2), (1, 0), (2, 0), and (2, 1) (the first number is the phase of the 5' intron and the second number is the phase of the 3' intron). Symmetric exons will spread more easily by exon shuffling (29) because the addition of a symmetric exon into an intron of the same phase does not disturb the reading frame. On the contrary, the intron insertional theory predicts a random correlation of intron phases across exons.

The simplest null hypothesis again is the random expectation that intron phases should have equal frequencies, 1/3, and that each pair of introns, each exon combination, should be equiprobable at 1/9. Compared to this, the intron phase correlation is very far from random, and there is a 125% excess of symmetric (0, 0) exons over the null expectation. This dramatic excess of symmetric exons could be taken as support for exon shuffling, with the corollary that the excess of phase zero introns would then be a result of the shuffling of symmetric (0, 0) exons. However, let us consider the more
Table 2. Observed and expected symmetric intron phase associations

<table>
<thead>
<tr>
<th>No. of exons</th>
<th>(0, 0)</th>
<th>(1, 1)</th>
<th>(2, 2)</th>
<th>No. of exons</th>
<th>(\chi^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2325/2075 (12%)</td>
<td>1085/832 (30%)</td>
<td>519/461 (13%)</td>
<td>9192</td>
<td>200.4</td>
</tr>
<tr>
<td>2</td>
<td>1935/1829 (6%)</td>
<td>930/662 (40%)</td>
<td>421/373 (13%)</td>
<td>7556</td>
<td>177.2</td>
</tr>
<tr>
<td>3</td>
<td>1593/1492 (7%)</td>
<td>715/518 (38%)</td>
<td>330/305 (8%)</td>
<td>6202</td>
<td>132.8</td>
</tr>
<tr>
<td>4</td>
<td>1357/1230 (10%)</td>
<td>602/424 (41%)</td>
<td>311/252 (23%)</td>
<td>5116</td>
<td>150.5</td>
</tr>
<tr>
<td>5</td>
<td>1060/1015 (4%)</td>
<td>461/350 (32%)</td>
<td>209/208 (0%)</td>
<td>4224</td>
<td>53.6</td>
</tr>
<tr>
<td>6</td>
<td>899/842 (7%)</td>
<td>390/290 (34%)</td>
<td>183/172 (6%)</td>
<td>3503</td>
<td>60.4</td>
</tr>
<tr>
<td>7</td>
<td>736/606 (6%)</td>
<td>318/240 (33%)</td>
<td>133/142 (6%)</td>
<td>2893</td>
<td>42.6</td>
</tr>
</tbody>
</table>

The observed and expected intron associations are expressed as the observed number/the expectation and excess of observed is in parentheses. For one exon, the expectation is \(PPN\), where \(P\) is the proportion of introns of phase \(i\) and \(N\) is the total number of exons. For sets of exons, the expectation is the Bayesian probability based on the observed proportions of internal exons (15). The \(\chi^2\) values are calculated on the total distribution (nine numbers). The asymmetric exon sets are all less than expectation.

conservative hypothesis, that the phase zero bias might have some cause other than exon shuffling, and compare the observed frequencies of exons with the expected frequencies based on a random association of introns [the expected frequency of exon \((i, j)\) being then the product of the frequencies of introns of phase \(i\) and phase \(j\)].

The \(\chi^2\) test shows that the deviation of the 9192 exons from a random association of phase is very significant (\(\chi^2 = 200\), a \(P\) value less than \(10^{-46}\)). Thus intron phases are not combined in a random fashion. Furthermore, symmetric exons are significantly more frequent than expected and asymmetric exons are fewer. Table 2 shows that there is a 12% excess of \((0, 0)\) exons over expectation and a 30% excess of \((1, 1)\) exons. This is strong evidence for the exon shuffling hypothesis, because, even if there were some hitherto undiscovered rule about phase zero introns and correlations of phase zero introns, one would not expect a significant excess of \((1, 1)\) exons; symmetry per se seems to be favored. One source of nonrandom bias is human error in the data base. We estimate that the feature table is wrong in a few percent of the entries. The most common error is to assign introns to phase zero rather than using \(-\to GT\to AG\)= rules. Such a bias would not lead to an excess of \((1, 1)\) correlations.

One further property of exon shuffling is that groups of exons are likely to shuffle together. Table 2 also surveys the distributions of the outside intron phases for groups of two to seven exons. The observed frequencies of each group are compared with Bayesian conditional probabilities based on the frequencies of the exons making up the group. Since the numbers of exons in the sets of exons falls as the sets get longer (next-to-last column in Table 2), the statistical significance of the deviation from expectation (last column in Table 2) falls as the length of the set of exons increases. However, the striking fact is that the fractional excess over expectation is maintained for each type of symmetrical set; \((0, 0)\) sets are about 7% over expectation, while \((1, 1)\) sets are all about 30% over expectation. This behavior of sets of exons is a further argument for exon shuffling.

Could this excess of symmetric exons and exon sets be due to alternative splicing as distinct from exon shuffling? Alternative splicing is dependent solely on the biological properties of individual genes (30, 31) rather than the distances between the introns. Alternative splicing frequently adds alternative sets of exons to the beginning or ends of genes. Such events do not restrict the phases or the phase combinations of exons. Alternative splicing events that add an additional internal exon to a previously functioning structure do require that the added exon be symmetric, but, of course, also require that the exon/intron structure be related to the structure and function of the gene product.

**Role of Exon Shuffling in Evolution.** How much of the data base would have to be involved in exon shuffling events to produce these observed deviations from randomness? We take the difference between the observed and the expected number of exons in symmetric sets as an estimate of the number of exons which had to be involved in exon shuffling. Since there may be an overlap in the use of exons between the exon sets of different length, we calculate only the excess symmetric exons over expectation, and the specific excesses of pairs and triples that do not contain any symmetric exons within each set. The differences between observations and expectations \((d = O - E)\) are singles, \(d = 561\); doubles, \(d = 400\); and triples, \(d = 124\). Then the number of exons involved in shuffling is \(1 \times 561 + 2 \times 400 + 3 \times 124 = 1733\), which is 19% of the purged data base. This is a lower bound for the true fraction involved in shuffling, since we considered the excesses only for singles, certain pairs, and certain triples of exons to avoid the possibility of counting twice, and other factors like intron drift would weaken the evidence for symmetric exon patterns. (If all of the deviation from the 1/3 expectation of intron phases were to be due to exon shuffling, then at least 28% of the data base had to be involved.) Thus, quantitatively, exon shuffling is very important in the evolution of the intron/exon structure of genes.

This analysis has shown that the intron correlations in eukaryotic genes are not random, and it suggests a dominant role for exon shuffling. However, several authors (32–34) have argued that exon shuffling is a recent phenomenon, limited to vertebrate genes. Further analysis indicates that this may not be the case.

We analyzed the intron phase correlations for both plant and animal genes separately. Table 3 shows that the nonrandom

Table 3. Intron correlations in data base subsets

<table>
<thead>
<tr>
<th>Subset</th>
<th>(0, 0)</th>
<th>(1, 1)</th>
<th>(2, 2)</th>
<th>No. of exons</th>
<th>(\chi^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant</td>
<td>579/515 (12%)</td>
<td>118/88 (34%)</td>
<td>76/71 (7%)</td>
<td>1642</td>
<td>33.5</td>
</tr>
<tr>
<td>All animal</td>
<td>1869/1678 (11%)</td>
<td>992/774 (28%)</td>
<td>464/409 (13%)</td>
<td>7923</td>
<td>160.0</td>
</tr>
<tr>
<td>Animal without Caenorhabditis elegans</td>
<td>1189/1011 (11%)</td>
<td>778/604 (29%)</td>
<td>235/208 (13%)</td>
<td>5010</td>
<td>163.1</td>
</tr>
</tbody>
</table>

The proportions of phase zero, one, and two introns for the internal exons of plant genes are 56%, 23%, and 21% (total 2790 introns); the proportions for animal genes are 46%, 31%, and 23% (total 10,999 introns). The expectations for the symmetric exons are calculated on the basis of these proportions.
correlations of introns and the excess of symmetric exons holds not only for animal genes but also for plant genes. Thus, exon shuffling is not limited to vertebrate or animal genomes. [This statistical argument is also supported by a recent report of exon shuffling in plant genes (35).] The animal genes in the data base include a large number of genes from the C. elegans project whose intron positions were determined by a computer program, GENEFINDER (36). One might worry that these "hypothetical" genes might perturb the calculation. Table 3 also shows that all the C. elegans genes can be dropped from the data base without affecting the phase correlation excesses.

**Intron Phase Correlations in ACRs.** More importantly, we have investigated intron phase correlations in a subset of 296 genes from the final data base that are homologous to prokaryotic genes (27), using only those introns that lie in the region of match between the eukaryotic proteins and the prokaryotic ones. These ACRs represent gene structures that were in existence in the last common ancestor of the prokaryotes and the eukaryotes. Since these genes have no introns in the prokaryotes, the introns-late theory holds that all the introns in this portion of the data base are insertional. No shuffling could have occurred after the prokaryotic--eukaryotic divergence, since these gene regions are collinear. However, just as for the overall data base, the distribution of intron phases and phase combinations in these ACRs are not random. Among the 1496 introns, phase zero introns constitute 55%; phase one introns, 24%; and phase two introns, 21%. Furthermore, there are statistically significant excesses of symmetric exon sets. Table 4 shows the distribution of individual symmetric exon types for sets of one to four exons. The distributions for individual patterns of pairs, triples, and quadruples of exons reach statistical significance. However, in all cases, a two-way test shows clearly that symmetrical exons are in excess at the 1% probability level. Thus the intron/exon structure of these ACRs suggests that they were constructed by exon shuffling, which would have to have occurred in the progenote.

**DISCUSSION**

How robust is this analysis? One might argue that those intron correlations are due to some pattern of repeated genes in the data base: that the purging had not produced a set of independent entries. To test this possibility we repeated the calculation for data bases created at different purging levels. Fig. 2 shows that the intron phase frequencies and the excess of symmetric exons (only pairs are shown) are stable across the purging levels. Thus, the deviations from randomness are not a property of redundancy in the data base. The few quasi-independent sequences in the data base which may come from same-gene families do not bias the statistical analysis. Even if we purge the data base to the 10% level, yielding only 689 genes and 7115 exons (657 genes with 5737 internal exons), the phase correlation excesses still stand.

To give an insertional theory a best shot, we have assumed that the nonrandom intron phase distribution might be a result of introns inserted into specific sequences of bases which could be correlated with local amino acid sequence. One example of this is the suggestion that an AG/GT sequence in the exons at the junctions is a remnant of a preinsertion target. However, Stephens and Schneider (37) find no significant AG/GT sequence within the exons in a study of 1800 human introns, and only a slight excess of AG/G. [An alternative interpretation of this AG/G bias would be that such sequences were selected for, after the fact, by evolutionary pressure from the splicing mechanism, such as better matching to the small RNAs (38).] On such insertional models, one might argue that the base composition of the DNA would bias the third positions of codons and thus that variation in the GC bias would drive such "entry" sequences into different phase relations in different genes. To test if the intron phase correlations are the result of some genes being rich in phase zero introns while other genes are rich in phase one introns, a feature that would create correlations between the phases, we examined the distribution of introns within genes in the data base. Histograms of the distribution of introns within genes show a pattern broadly peaked around the average positions, except for spikes at 100% for the three pure phases. We have dropped those sets of exons (68 genes) and repeated the calculation of symmetric exon excesses to find again a 9% excess of (0, 0) exons and a 24% excess of (1, 1) exons.

**Fig. 2.** Stability of the deviations from expectation at various purging levels. (A) Deviations of the intron phases from the 1/3 expectation. (B) Difference between the frequencies of symmetric pairs of exons and the expectation based on the observed exon frequencies.
This study of intron phase correlations has revealed a signal which we interpret as a mark of exon shuffling. This statistical approach has the advantage that such a signal is only weakened, not abolished, by the drift or loss of introns over evolutionary time or by the addition of a few novel introns. The argument is that the intron correlations in the eukaryotic versions of "ancient" genes, the excess of symmetric sets of exons, shows that these exon sets were once used in exon shuffling which had to occur in the progenote. Hence these exons were elements that were used to compose the original genes. These intron phase correlations are an independent support for the exon theory of genes. The argument from phase correlations thus adds to the previous arguments based on the correlation of intron positions with three-dimensional structure for triosephosphate isomerase (8–10), the coincidence of plant and animal introns (11–15), or the identification of ancient introns on the basis of the patterns of descent (16–19) to support the idea that some introns are extremely ancient. Unlike the previous arguments, the intron phase correlation calculation detects a clear, statistically significant signal for the existence of early introns against a background of loss and gain.

We thank Richard C. Lewontin, Nathan Goodman, Charles H. Langley, Stephen M. Mount, Daniel Weinreich, Andrew Berry, Leslie Gottlieb, and all the members of the Gilbert Lab, especially Keith Robison and Zhiping Liu, for useful discussions and helpful computing advice. We thank Philip Green for the computer program for the Smith–Waterman algorithm and Eric Lander for allowing us access to his computing facility. We thank the National Institutes of Health for their support, Grant GM37997 to W.G. and Grant HG0098 to Eric Lander.