Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 5094–5099, April 1998
Evolution

Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins

(introns-early/introns-late/modules)

SANDRO J. DE SOUZA, MANYUAN LONG, ROBERT J. KLEIN, SCOTT ROY, SHIN LIN, AND WALTER GILBERT*

Department of Molecular and Cellular Biology, The Biological Laboratories, Harvard University, Cambridge, MA 02138

Contributed by Walter Gilbert, February 25, 1998

ABSTRACT **We present evidence that a well defined subset of intron positions shows a non-random distribution in ancient genes. We analyze a database of ancient conserved regions drawn from GenBank 101 to retest two predictions of the theory that the first genes were constructed by exon shuffling. These predictions are that there should be an excess of symmetric exons (and sets of exons) flanked by introns of the same phase (positions within the codon) and that intron positions in ancient proteins should correlate with the boundaries of compact protein modules. Both these predictions are supported by the data, with considerable statistical force (*P* values < 0.0001). Intron positions correlate to modules of diameters around 21, 27, and 33 Å, and this correlation is due to phase zero introns. We suggest that 30–40% of present day intron positions in ancient genes correspond to phase zero introns originally present in the progenote, while almost all of the remaining intron positions correspond to introns added, or moved, appearing equally in all three intron phases. This proposal provides a resolution for many of the arguments of the introns-early/introns-late debate.**

The rapid expansion of knowledge of DNA sequences, rising a factor of 10 every 5 years, has now reached the point where one can survey with great statistical power the intron spectrum of genes. This has enabled us to create critical tests of speculations about the role of introns and their history by studying ancient conserved genes, whose protein products are conserved between prokaryotes and eukaryotes. Such genes have no introns in their prokaryotic forms but introns in their eukaryotic homologs. Introns-late models must predict that all introns in these genes were inserted into previously continuous genes that correspond to ancestral forms that were similar to the current prokaryotic genes (1, 2). Thus, such models predict that these introns should not respect intron phase (the position within a codon), should not show phase correlations, and should not be related to the three-dimensional structure of the protein products of these genes. Alternatively, introns-early models, which hypothesize that introns were used in the progenote to assemble the first genes (3–6), look upon some or all of these introns as residues of that process and expect these introns to have been associated with the process of exon shuffling and, hence, to show restrictions on intron phase, to show phase correlations, and to be related to the three-dimensional structure of the proteins.

Over the last several years, we have published two statistical arguments that suggest that introns share properties of the type predicted by an introns-early theory. One of these arguments is that introns in genes for ancient conserved proteins are correlated in phase (the position within the codon) so that exons, or sets of exons, tend to begin and to end in the same phase, to be multiples of three bases. This argument was shown to hold at about the $P = 0.01$ level (7). A second argument showed that intron positions were correlated with an aspect of the three-dimensional structure of ancient proteins, specifically that intron positions were associated with compact modules of diameters 21, 27, and 33 Å, with *P* values less than 0.01 (8). Both of these regularities are predictions of any theory that holds that some or all of the introns were used in the progenote to assemble the genes for these proteins by exon shuffling; neither of these regularities is predicted by theories which hold that the introns were inserted into DNA by processes that are unrelated to the ultimate structure of the gene product.

However, in the last year two papers have appeared that continue the argument that introns are late. One by Cho and Doolittle (9) tries to study a possible coincidence of intron positions in gene pairs that represent duplications that occurred in the progenote, ancient paralogous genes to ask whether the pattern of intron positions in those genes is more suggestive of intron addition or intron loss. A second paper studying the intron distribution in a large gene family argues that the pattern observed is more one of addition or movement than loss (10).

The continuing increase of DNA sequences in the public databases, increasing by a factor of two every 18 months, has led us to reinvestigate this problem using much more data. In this paper, we shall show that the statistical regularities mentioned above can now be analyzed in greater detail with much higher statistical confidence. We reaffirm the basic regularities that we saw before, but now, since there is more data, we can go further in the analysis of the correlation of introns with three-dimensional structural elements. This further analysis shows that the strong correlation is carried by introns that lie between codons (in phase zero), while the introns that lie within the codons (phase one and phase two) do not show strong correlations with three-dimensional structure. This analysis suggests an explicit description of intron positions in terms of both ancient introns and later additions in a way that resolves the conflict between the two viewpoints. We conclude that about 35% of the introns present in ancient genes are ancient, lie primarily in phase zero between codons, and are related to compact elements of protein structure, modules, ranging in diameter between 21 and 33 Å. About 65% of the introns have been added to pre-existing genes, equal fractions in each of the other phases uncorrelated to structure. This division explains why certain analyses see a large fraction of introns as being added to previously existing genes, while the theory that the original genes were constructed through introns remains the simplest and strongest way of predicting the observed regularities.

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.

Abbreviation: ACR, ancient conserved region.

*To whom reprint requests should be addressed at: Department of Molecular and Cellular Biology, The Biological Laboratories, 16 Divinity Ave., Harvard University, Cambridge, MA 02138. e-mail: gilbert@chromo.harvard.edu.

PROCEDURES

Intron Database. We used GenBank release 101 to construct a database containing all entries with an intron/exon organization. We then purged it down to a criterion of a 20% match to the shorter sequence, keeping the sequence with more introns each time, using a program, GBPURGE, written for a DEC Alpha based on a FASTA comparison. The ancient conserved region (ACR) database was constructed, and the expected frequencies of intron phase combinations were calculated as described before (7).

Dataset for Structural Analysis. Forty-four ancient proteins (list below) with 988 intron positions were used in this study. Intron positions were defined by searching the full intron database with the Protein Data Bank reference sequence using FASTA. One difference between the approach used here and that in de Souza *et al.* (8) and Gilbert *et al.* (11) is that a specific program creates two files, one containing the structure coordinates and one containing the sequence of the reference protein in FASTA format (used to search the intron database), using the original Protein Data Bank file as a template. This additional step made sure that the coordinate and FASTA sequence files exactly correspond to each other. The source code of this program will be available on our web site (http://golgi.harvard.edu/gilbert.html).

List of 44 Ancient Proteins. The names inside parentheses correspond to the Protein Data Bank accession codes; the 12 additional proteins are starred: aspartate aminotransferase (1ama); acid amylase (2aaa); acyl-CoA dehydrogenase (3mdd); adenosine deaminase* (1add); alcohol dehydrogenase (1adb); adenylate kinase* (3adk); aldehyde dehydrogenase* (1ad3); aldolase (1ald); alkaline phosphatase (1aja); aldose reductase (1dla); amylase (1ppi); aspartate transcarbamoylase* (1raa); aspartyl-trna synthetase* (1asy); catalase (8cat); citrate synthase (1cts); cu++ superoxide dismutase (1sdy); cytochrome *c* (1ccr); dihydrofolate reductase (1dhf); dihydrolipoamide dehydrogenase* (3lad); elongation factor tu (1eft); enolase (1ebg); glucose 6-phosphate dehydrogenase (1dpg); glyceraldehyde 3-phosphate dehydrogenase (3 gpd); glutamate dehydrogenase* (1hrd); glycogen phosphorylase (1 gpa); glutathione reductase* (1 gra); glutathione *S*-transferase (1gss); hemoglobin (2dhb); high pi amylase (1amy); heat shock protein 70 (1atr); lactate dehydrogenase (2ldx); lysozyme (1laa); malate dehydrogenase (4mdh); mn++ superoxide dismutase (1msd); nucleoside diphosphate kinase* (1ndl); ornithine transcarbamoylase* (1ort); porphobilinogen deaminase* (1pda); phosphofructokinase (3pfk); phosphoglycerate kinase (3pgk); phosphoglycerate mutase (3pgm); pyruvate kinase (from author); thioredoxin* (2trx); triosephosphate isomerase (1tim); xylanase (1clx).

RESULTS

Intron Phase Correlations Revisited. We extracted a subdatabase of genes with introns from GenBank 101 to obtain a database with 25,666 entries. We purged this to eliminate sequences that matched by more than 20% of the shorter sequence, using a FASTA matching program, and saved the versions that have more introns to produce a reduced database of 5,772 members. A large fraction of the genomic genes now come from the *Caenorhabditis elegans* sequencing project. The intron/exon structure of these genes is somewhat problematic because it is predicted by computer. These genes account for 42% of our original database. We constructed purged databases with or without the *C. elegans* material: 5,772 genes with and 1,997 genes without. We identified ACRs as those regions of eukaryotic sequence homologous and colinear to prokaryotic genes using as a criteria a BLAST score greater than 75. We thus obtained a database of introns that lie within the ACRs which we could analyze for intron-phase correlations. All of these introns must have been added to the pre-existing gene on an introns-late model, since there can be no exon shuffling in the history of these particular regions of the eukaryotic genes. However, introns-early models predict that some or all of these introns could be the result of exon shuffling that created the ancestral form of these genes in the progenote.

As we previously observed (7), there is a bias in the intron phase distribution for ACR introns: 54% lie in phase zero, 25% lie in phase one, and 21% lie in phase two. Above this bias, there are correlations in phase between introns on either side of exons producing an excess of symmetric exons. Table 1 shows these excesses, as well as listing the behavior of symmetric pairs, triples, quadruples, and quintuples of exons. For each of these cases, the expected values are calculated based on the observed frequencies for the components. The expectation for symmetric exons is based on the frequency of introns in each phase; the expectation for the symmetric pairs of exons is calculated using the observed frequency for the component exons, beginning and ending in all phases; and similarly for the other sets. Table 1 shows that the excesses of symmetric exons and exon sets above expectation are extremely significant. Table 2 shows the same calculation for the ACR dataset derived from the database that includes *C. elegans*; this larger dataset shows an even greater statistical significance.

The approximately 10% excess of symmetric zero–zero exons and 15% excess of one–one exons (and exon sets) is not the expectation of any insertional model but is consistent with an exon shuffling history. These excesses are above the biased expectations based on the observed intron phase frequencies, in which the greatest number of introns are in phase zero, and the largest excess over that biased expectation is for the one–one symmetric exons. We stress that the actual deviations from randomness are very large. If the original expectation for intron phases had been random, as it would be on the simplest addition model, one-third in each phase, then there are almost three times more zero–zero exons than expected, a 200% excess.

Correlations between Intron Position and Module Boundaries. We have reanalyzed the correlation between intron positions and the three-dimensional structure of the protein products of ancient conserved genes, using a larger set of genes and a more extensive set of intron positions. We expanded the group of 32 proteins (8) to a set of 44 ancient conserved proteins and 988 intron positions from GenBank 101. Where the three-dimensional structure was available, if we had not already used those genes, we added new genes that had been

Table 1. Intron correlations within ancient conserved regions (*C. elegans* sequences excluded)

Length	(0,0)	(1,1)	(2,2)	Number	χ^2	*P*
1	1046/934 (12%)	237/210 (13%)	165/140 (18%)	3241	41.5	1×10^{-6}
2	879/779 (12%)	172/154 (19%)	115/113 (11%)	2599	32.4	1×10^{-4}
3	739/656 (13%)	143/116 (23%)	84/90 (−6%)	2105	39.8	5×10^{-6}
4	616/556 (11%)	111/88 (26%)	64/69 (−7%)	1702	26.4	8×10^{-4}
5	486/456 (7%)	82/71 (15%)	57/53 (8%)	1371	8.4	0.4

The data are given for each exon type as observed number/expected number, with the percent excess of observed over expectation in parentheses. The column labeled "Number" lists the total number of exons or of exon sets of the given length. There are 910 ACR regions in this database, and the overall phase bias is 54, 25, and 21% for phases zero, one, and two.

Table 2. Intron correlations within ancient conserved regions using the entire purged database

Internal Exons	(0,0)	(1,1)	(2,2)	No. of exons	χ^2	P
1	1506/1388 (8%)	392/321 (22%)	300/272 (10%)	5133	51.7	3×10^{-8}
2	1181/1059 (12%)	279/234 (19%)	226/203 (11%)	3857	46.7	5×10^{-8}
3	931/828 (13%)	202/171 (18%)	141/149 (−5%)	2921	40.2	1×10^{-6}
4	731/640 (14%)	153/129 (19%)	108/111 (−3%)	2231	29.9	2×10^{-4}
5	556/507 (10%)	107/100 (7%)	85/78 (9%)	1711	16.7	3×10^{-2}

The data are given for each exon type as observed number/expected number, with the percent excess of observed over expectation in parentheses. The column labeled "Number" lists the total number of exons or of exon sets of the given length. There are 1,916 ACR regions in this database, and the overall phase bias is 52, 25, and 23% for phases zero, one, and two.

used in two recent papers that argued for a pattern of later-moved introns.

We analyzed the three-dimensional structures with a program, INTERMODULE (8). Briefly, we define a module as a segment of the polypeptide chain such that all the distances between the C-alpha carbons are bounded by some maximum diameter. The program dissects each three-dimensional structure into a minimally overlapping set of modules of the specified diameter. The overlaps between the modules, which we call boundary regions, provide a series of regions in which we expect to find an excess of intron positions. These "boundary regions" are such that if an intron were to be placed into each of the boundary regions, the gene product would be dissected into a set of modules all less than the specified diameter. We accumulate a list of all intron positions in genes homologous to the sequence of the known structure, counting each different intron position in the nucleic acid sequence once. We then calculate whether there is an excess of intron positions in the boundary regions over the random expectation, which is that the introns were added to the DNA in a manner that did not respect protein structure. We test the significance of the excess with a simple χ^2 calculation. Fig. 1 shows the output of this calculation for this set of 44 protein and 988 intron positions and displays the χ^2 values for each possible module diameter from 15 Å to 40 Å. Fig. 1 shows that there is a statistically significant excess of intron positions in boundary regions for a range of module sizes, with striking peaks around diameters of 21, 28, and 33 Å [as we observed in de Souza *et al.* (8)]. Minor peaks appear near 25 and 37 Å. The major peaks reach χ^2 values of 11, 13, and 9 and probability values around $P = 0.001$. We interpret this curve as showing that introns tend to mark the boundaries of modules of different sizes in this set of proteins. The linear amino acid sequences that correspond to these module diameters are about 15 amino acids long for the smallest modules and range up to an average length of 30 amino acids for the 33-Å modules. Fig. 1 shows that intron positions are correlated with short elements of polypeptide structure in these 44 ancient conserved proteins. The phenomenon is robust; if one examines the excess of intron positions in these regions, one sees a moderately smooth curve that tracks with the χ^2 result.

Now that we have so much more data, we can examine the separate components of the statistical signal. Slightly more than half the intron positions correspond to phase-zero introns, and so we break the data into a phase-zero portion and a phase-one plus phase-two portion. Fig. 2 shows the χ^2

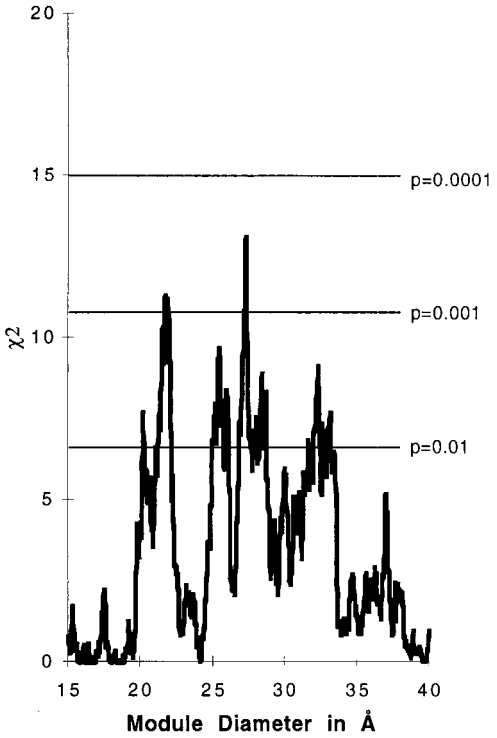

FIG. 1. χ^2 values for the excess of intron positions inside the boundary regions as a function of module diameter. The major peaks are around 21, 27, and 33 Å. The 988 intron positions were drawn from release 101 of GenBank.

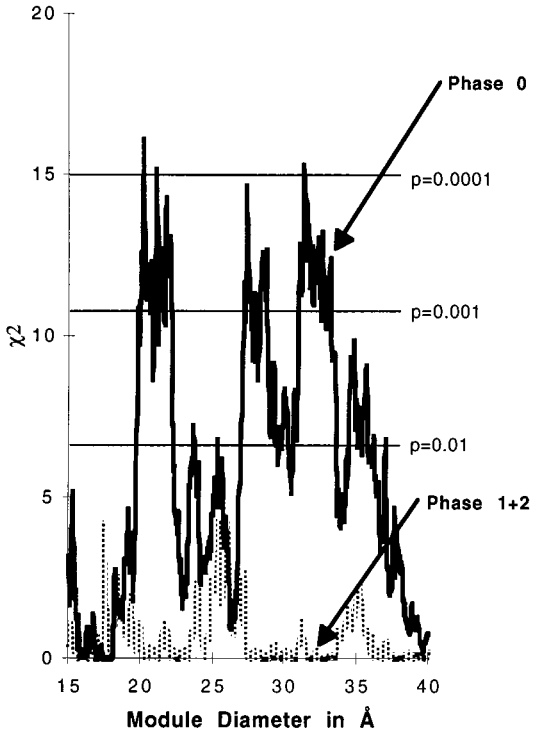

FIG. 2. The χ^2 values for the excess of intron positions inside the boundary regions as a function of module diameter for phase zero and for phases one and two separately. The INTERMODULE calculation was done for phase zero intron positions only (554 positions) (black) and for a set of both phase one and phase two intron positions (434 positions) (gray).

distribution for the excess of phase-zero intron positions in boundary regions as compared with the phase-one and phase-two intron positions. The statistical effect is carried entirely by the phase-zero introns. Phase one and phase two introns do not show enough preference for the boundary regions to be statistically notable. Furthermore, the statistical signal is stronger for the phase-zero data alone, which means that we have taken a random background out of the calculation. The statistical signal now reaches a χ^2 value of 16.5, a P value smaller than 0.0001 for the 21-Å diameter modules. In general, the P values are 10-fold better for the phase zero introns.

Of the 988 introns, 56% are phase zero, 23% are phase one, and 21% are phase two. We suggest that, since the simplest model for intron addition is that equal numbers of introns are added in all three phases, one should interpret the phase two introns as a measure of the background rate of addition and estimate that an equal number of introns were added in phase zero and phase one (added or randomly moved). The excess intron positions over this background, the 35 percentage points of the intron positions in phase zero, are candidates for being ancient introns. About two percentage points of the intron positions that lie in phase one are candidates for being ancient. To put this another way, we suggest that about 65% of all introns are new, added equally in all three phases, and that about 35% of all introns are candidates for being old, are correlated with three-dimensional structure of ancient proteins, and show the excess phase correlations. Of these, almost all are phase zero; 60% of all phase zero intron positions represent old positions, and about 10% of the phase one positions represent old positions.

This analysis can be taken further by asking whether one can see any variation in the different kingdoms in terms of this distribution of phase zero introns. For the 44 proteins, 405 intron positions (224 in phase zero) arise in genes sequenced from the vertebrates. There are 238 positions (150 in phase zero) in genes from the plants, 287 positions (163 in phase zero) in genes from the invertebrates, and 149 positions (72 in phase zero) in the genes from the fungi. When we look at these groups, we observe that vertebrate introns lack a correlation with protein modules. Fig. 3 shows as a percentage the excess of phase zero introns over the expectation for vertebrate and nonvertebrate introns. Vertebrate introns do not show much of an excess of phase zero introns, although there is a small excess around 33 Å. Fig. 4 shows that the overall χ^2 values improve for the set of nonvertebrate phase zero introns. The curve identifies patterns of modules at diameters of 21, 28, and 33 Å with P values < 0.0001.

DISCUSSION

This finding that the correlation of intron positions with the modular structure of ancient proteins is carried primarily by the phase zero introns, along with the interpretation that about 65% of the present introns are added or moved, while about 35% of the introns are correlated with the three-dimensional structure and are candidates for introns left over from exon shuffling in the progenote, provides a resolution of many of the arguments about introns early versus introns late. We identify one fraction of the introns as candidates for introns-late and another fraction as specific candidates for introns-early. This compromise does not, however, satisfy an introns-late view because it argues that a fraction of the introns are early, which has the implication that exon shuffling was involved in the construction of the first genes (5).

The lack of a general signal (there is a small excess of intron positions around 33 Å) for correlation in the vertebrates is interesting and puzzling. There is roughly the same excess of phase zero introns in the subset of vertebrate intron positions as there is in the other subsets. This would suggest that intron positions are subject to different dynamics in accordance with their phylogenetic distribution. The genes that we added to our original set of 32 proteins tended to weaken the statistical signal rather than strengthen it. The reason, we now understand, lies in that the bulk of the intron positions in those genes came from vertebrate sequences. We expect future work with other organisms will only strengthen the statistical signal.

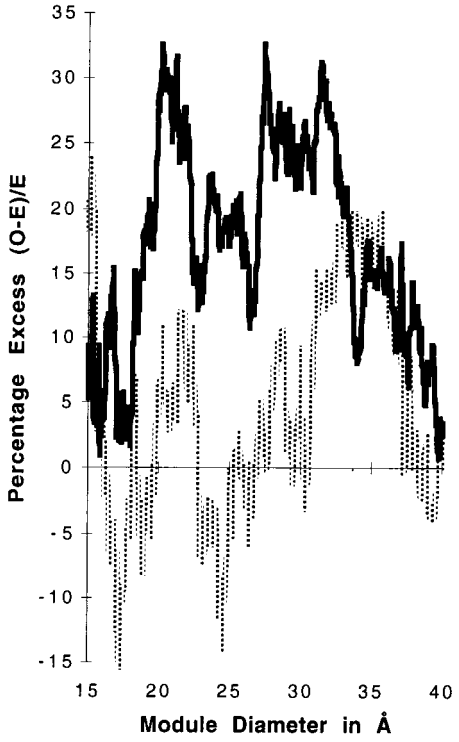

FIG. 3. The percentage excess of intron positions above expectation ((0-E)/E) for the datasets of nonvertebrate (black) and vertebrate (gray) phase zero intron positions.

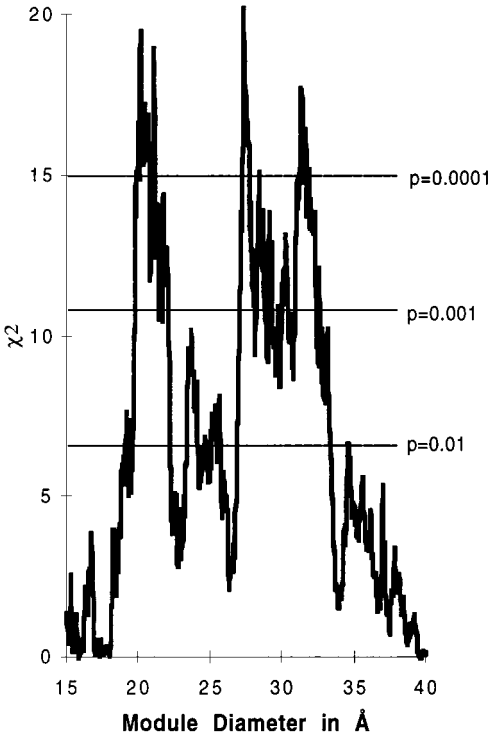

FIG. 4. χ^2 values for the excess of nonvertebrate phase zero intron positions (389 positions) inside the boundary regions as a function of module diameter.

How can this emphasis on phase zero intron positions be reconciled with the apparently greater excess of (1, 1) symmetric exons than (0, 0) symmetric exons? As we pointed out earlier, and as Tables 1 and 2 show, there is a much greater absolute number of symmetric (0, 0) exons than (1, 1) exons. In Table 1, there are 1,046 (0, 0) exons, 237 (1, 1) exons, and 165 (2, 2) exons. There are four times as many (0, 0) exons as (1, 1) exons and six times as many as (2, 2) exons, which represents, on our model, the "background" assumption. The greater background number of (1, 1) exons is part of the reason that we think there were a few phase one introns originally.

We have argued that about one-third of all introns are candidates for being ancient and fall mostly in phase zero (about 10% may be in phase one). The other two-thirds occur roughly equal in all three phases and are candidates for introns added in all three phases or moved randomly into all three phases. We can further use our data to estimate how many introns would have been lost during evolution. If the conjecture that the first genes were constructed of exons, in the length patterns that we infer, is correct (lengths ranging from roughly 15 to 30 amino acids), we might expect the 44 proteins, whose aggregate length is 15,400 amino acids, to have about 670 introns originally in phase zero. Since we now see about 330 introns in phase zero correlated with three-dimensional structure, we suggest that half of the original introns were lost. This figure is also consistent with the argument that half the introns were lost to generate the current exon spectrum from the original shorter spectrum.

Considerations on Theory. How does one test a theory? The general goodness of a theory lies in how well satisfied and how varied are its predictions. Does a theory encompass many aspects of the observed world or is it simply an *ad hoc* restatement of an observation? The theory that the intron/exon structure of genes is a consequence of the first genes being assembled by exon shuffling at the beginning of evolution makes a variety of different predictions, some of which have now been tested and shown to hold with high statistical support. This theory predicts that introns should tend to be in the same phase, that intron phases should be correlated across groups of exons with a preference for symmetric patterns, and that introns should be related to three-dimensional structure of proteins. We showed here that these three predictions are met for phase zero intron positions in ACRs; the actual fraction of intron positions involved we estimate to be about 30 to 40% of all intron positions in the ACRs.

The theory of early exon shuffling also makes further predictions which have not yet been tested. One of these predictions is that the modules should show a pattern of reuse across protein structures that would follow from their having been used by exon shuffling as elements to assemble these proteins. This prediction is that a small set of modules were used over and over again. Still a further prediction is that a pattern of introns will be found to repeat at the boundaries of modules reused by shuffling. These two further predictions have not yet been tested and are independent of the three tests that have already been done.

Further Alternative Theories. Are there alternative theories that could explain the observations? We discuss some theories of intron addition below. These theories have an *ad hoc* character and can often be shown not to hold by a consideration of other properties.

The excess of introns in phase zero in the intron phase distribution might be explained by the introns adding to shadow sequences in previously continuous genes, sequences like AGGT or AGG that have been hypothesized to serve as targeting sites for the insertion of introns (12). Such a theory can be tested by examining the conservation of exon sequences at intron boundaries, to see how strong such "shadow sequences" actually are, and by examining the distribution of putative target sequences to see if they match the phase distribution. We have given a discussion (13) that shows that the current distribution of such sequences does not mimic the intron phase distribution.

A separate theory to explain intron phase correlations, that exons are often multiples of three bases, is to postulate that the splicing mechanism measures the size of an exon and tends to measure that size in multiples of three. That such a mechanism might possibly exist could be supported biochemically by such arguments as those by Robberson *et al.* (14), that splicing mechanisms can recognize both ends of the exon, and by the observation in some RNA viruses (15) of a preferred packaging of RNA by a physical process in multiples of six. In such a model, one might hypothesize that the nucleosome restricts exons to multiples of three (by an unknown mechanism). Although such a model would predict an excess of symmetric exons, it would not predict any further excesses of symmetric exon pairs, triples, etc.

Consider the alternative theory that there was a set of target sequences, present a billion years ago but which have mutated since, that were correlated with amino acid sequences in such a way that the inserted introns tend to lie between amino acids and the corresponding amino acid sequences lie in the regions between the modules, the boundary regions. Such a theory has a superficial plausibility. However, it would be untestable by the phase position, the symmetric exon, or the module data because it is hypothesized to agree with these findings. However, this example of an *ad hoc* theory does not predict the excess of symmetric sets of exons nor does it predict any reuse of modules in different proteins.

There are a variety of theories that attempt to correlate positions of added introns to the boundaries of modules by invoking some form of evolutionary selection pressure. One class of such theories suggests that introns are effectively mutagenic upon their addition to a previously existing gene and, hence, would tend to survive in loops in the proteins or in other regions of low conservation. However, the boundaries of modules, as we have described them, lie frequently within alpha helices or beta strands. Introns are often found in regions of extremely high conservation, and, in general, introns lie in regions of high conservation in proportion to the extent of such regions.

Another hypothesis is that, as introns add to pre-existing genes, if a pair of introns falls around a module, that module might be shuffled out and used in another gene and, hence, selected by evolution. Such models preserve the modules created by intron addition because they are used by shuffling. However, these models invoke a wrong view of evolution. The selection procedure that fixes in the population the shuffled module in the largest gene does not fix the original version of the donor gene, since the donor gene, in general, is genetically unlinked. The ancient conserved genes that we have studied here have to be donor genes on these models.

Another evolutionary argument suggests that when introns add to pre-existing genes, the increased homologous recombination that the intron creates, between the parts of the protein that lie outside its termini, is deleterious if the intron lies inside a module because recombination breaks up co-adapted sites inside the module (M. Meselson, unpublished manuscript). This model, however, does not explain why the exon/module correlation exists only for phase zero introns as well as the excess of symmetric exons.

Although these particular *ad hoc* theories fail, there may still be some alternative theory that accounts for all the data. The nature of the alternative is yet unknown, the best tests of the Exon Theory of Genes lie in its further predictions.

Overall, our final picture is that about 35% of today's intron positions in ancient proteins represent ancient introns, that over the course of evolution about half of the original introns

were lost, and that, again over evolutionary time, a number of introns have been added in all three phases corresponding roughly to 65% of the intron positions in today's databases.

S.J.d.S. was supported by Fundacao de Amparo a Pesquisa do Estado de Sao Paulo (Sao Paulo, Brasil) and the PEW-Latin American Fellows Program.

1. Cavalier-Smith, T. (1991) *Trends Genet.* **7,** 145–148.
2. Palmer, J. D. & Logsdon, J. M. J. (1991) *Curr. Opin. Genet. Dev.* **1,** 470–477.
3. Doolittle, W. F. (1978) *Nature (London)* **272,** 581–582.
4. Gilbert, W. (1979) *Introns and Exons: Playgrounds of Evolution.* (Academic Press, New York).
5. Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 901–905.
6. de Souza, S. J., Long, M. & Gilbert, W. (1996) *Genes Cells* **1,** 493–505.
7. Long, M., Rosenberg, C. & Gilbert, W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 12495–12499.
8. de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W. & Gilbert, W. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14632–14636.
9. Cho, G. & Doolittle, R. F. (1997) *J. Mol. Evol.* **44,** 573–584.
10. Rzhetsky, A., Ayala, F. J., Hsu, L. C., Chang, C. & Yoshida, A. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 6820–6825.
11. Gilbert, W., de Souza, S. J. & Long, M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 7698–7703.
12. Dibb, N. J. & Newman, A. J. (1989) *EMBO J.* **8,** 2015–2021.
13. Long, M., de Souza, S. J., Rosenberg, C. & Gilbert, W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 219–223.
14. Robberson, B. C., Cote, G. J. & Berget, S. M. (1990) *Mol. Cell. Biol.* **10,** 84–94.
15. Calain, P. & Roux, L. (1993) *J. Virol.* **67,** 4822–4830.