Large-scale comparison of intron positions among animal, plant, and fungal genes

Alexei Fedorov*, Amir Feisal Merican*†, and Walter Gilbert*‡

*Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138; and †Institute of Biological Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia

Contributed by Walter Gilbert, October 15, 2002

We purge large databases of animal, plant, and fungal intron-containing genes to a 20% similarity level and then identify the most similar animal–plant, animal–fungal, and plant–fungal protein pairs. We identify the introns in each BLAST 2.0 alignment and score matched intron positions and slid (near-matched, within six nucleotides) intron positions automatically. Overall we find that 10% of the animal introns match plant positions, and a further 7% are ''slides.'' Fifteen percent of fungal introns match animal positions, and 13% match plant positions. Furthermore, the number of alignments with high numbers of matches deviates greatly from the Poisson expectation. The 30 animal–plant alignments with the highest matches (for which 44% of animal introns match plant positions) when aligned with fungal genes are also highly enriched for triple matches: 39% of the fungal introns match both animal and plant positions. This is strong evidence for ancestral introns predating the animal–plant–fungal divergence, and in complete opposition to any expectations based on random insertion. In examining the slid introns, we show that at least half are caused by imperfections in the alignments, and are most likely to be actual matches at common positions. Thus, our final estimates are that ≈14% of animal introns match plant positions, and that ≈17–18% of fungal introns match animal or plant positions, all of these being likely to be ancestral in the eukaryotes.

exon | phase distribution | evolution | eukaryote | prokaryote

Introns are prevalent in the complex eukaryotes but rare in the simple ones. Are these introns ancestral in all of the eukaryotes or do they arise as the organisms become more complex? Introns can be acquired by or eliminated from a gene during evolution, but what is the balance?

An introns-late view argues that introns arise as "selfish" elements that play no constructive role in evolution. On this picture, introns appear relatively late in the evolution of eukaryotes (1–3) and spread as mobile elements that invade genes by insertion into short ≈4- to 5-nt-long "proto-splice sites" (4) (although the notion of proto-splice sites has been challenged; refs. 5 and 6).

An introns-early theory suggests that introns made an essential contribution to the evolution of genes via "exon shuffling," which created genes from exon "pieces" by recombination within the introns (7–12). In this view, introns existed before any eukaryote–prokaryote divergence, and since that time, the prokaryotic lineage completely lost its introns, whereas they were retained in the eukaryotes.

The sequences within the introns change during evolution, far more rapidly than those of the exons. The only conserved elements are the short sequences at the 5′ and 3′ termini, which are very similar for all introns. The rest of the intron sequence appears neutral to selection, and the length of the intron sequence can change by orders of magnitude. However, the position of an intron in a gene's coding sequence is well conserved. If one compares the exon-intron structure of orthologous genes, for example, those of human and fruit fly, one cannot establish any relationships between introns based on their sequences, but about half of the fruit fly introns have the same positions as human introns on the aligned sequences of the orthologous gene pairs.

In the 1980s, the first gene comparisons between animals and plants suggested that there were conserved introns that would have descended from a common ancestor (13–16). However, this view of early introns was strongly challenged by analyses of the intron distribution in different branches of eukaryotes, supporting an introns-late scenario (1, 3, 17).

The conjecture that there were deep branching, a-mitochondrial eukaryotes that lacked introns (1), such as *Giardia*, has now been questioned. Recent work has shown that possibly all these protists branch very much later, with the fungi, and are the offspring of mitochondrial-bearing ancestors. Furthermore, there are signs of the splicing apparatus, and even introns, in many of these species (18–22), suggesting that the splicing apparatus and introns are ubiquitous in all eukaryotic species.

However, are the introns added separately down each lineage or are they ancestral? There is clearly a large amount of intron loss and intron gain. How large a signal is there from introns putatively ancestral in the eukaryotes? To attack this question, we have carried out a large-scale comparison among animal, plant, and fungal genes. We obtained three large samples of animal, plant, and fungal genes with known exon–intron structures, performed pair-wise comparisons of each gene from one sample with every gene from the other two samples, selected the best matched pairs, and automatically marked intron positions on the alignments. Our final estimates are that ≈14% of animal introns match plant positions, and that ≈17–18% of fungal introns match animal or plant positions, all of these being likely to be ancestral in the eukaryotes. The 30 animal–plant alignments with the highest matches (for which 44% of animal introns match plant positions) when aligned with fungal genes are also highly enriched for triple matches: 39% of the fungal introns match both animal and plant positions. This is strong evidence for ancestral introns predating the animal–plant–fungal divergence, and in complete opposition to any expectations based on random insertion.

Materials and Methods

Sample of Genes. Our source of genes with known intron positions was the Exon–Intron Database (EID, www.mcb.harvard.edu/gilbert/EID) (23), derived from GenBank, release 121 (24). This EID database was filtered to remove all noncanonical introns (introns without canonical dinucleotides at their termini: gt . . . ag, gc . . . ag, at . . . ac), since noncanonical junctions are a main indicator of intron-position errors in GenBank. Then, a plant sample (32,234 entries), an animal sample (54,671 entries), and fungal sample (7,478 entries) were extracted from this filtered database. The plant sample is composed of those EID entries that have both "PLN" and "Viridiplantae" in the species-description line. The animal sample is composed of those EID entries that have "INV," "PRI," "ROD," "MAM," or "VRT," as well as "Metazoa" in the species-description line. The fungal

Abbreviation: EID, Exon–Intron Database.

‡To whom correspondence should be addressed. E-mail: gilbert@nucleus.harvard.edu.

Table 1. Comparison of intron positions and phases

Introns	Type of comparison	No. of alignments	Length of alignments, aa	No. of introns				% Common/all introns	% Common/all random exp	Phase distribution, %					
										All introns			Common introns		
				Common	Sliding	Unique	Total			p0	p1	p2	p0	p1	p2
Animal	ANM–PLN	1,514	5.8×10^5	855	623	7,259	8,737	9.8	0.68	51	24	25	56	23	21
Plant	ANM–PLN	1,514	5.8×10^5	855	623	10,345	11,823	7.2	0.50	63	17	21	56	23	21
Fungus	ANM–FNG	684	2.6×10^5	288	99	1,564	1,951	14.8	0.54	42	31	27	42	32	26
Animal	ANM–FNG	684	2.6×10^5	288	99	3,836	4,223	6.8	0.26	51	25	24	42	32	26
Fungus	PLN–FNG	674	2.7×10^5	253	158	1,595	2,006	12.6	0.69	42	32	27	47	29	23
Plant	PLN–FNG	674	2.7×10^5	253	158	5,193	5,604	4.5	0.25	62	17	21	47	29	23

The source of the introns is given in the first column. The order of the comparison of genes, which determines which gene regions are selected from the database during the calculation, is given in the second column. The databases used were purged to the 20% similarity level. The random expectation is calculated based on the total nucleotide length of the alignments. The phase zero, one, and two introns are shown as p0, p1, and p2.

sample contains those entries labeled "PLN" and "Fungi." The samples were then purged to 20% homology level to remove gene duplicates using the program GBPURGE (www.fallingrain.com/publicserver), yielding an animal sample of 9,456 entries, a plant sample of 5,455 entries, and a fungal sample of 1,956 entries.

Comparison of Intron Positions. To compare intron positions in homologous genes we used flat FASTA-formatted files containing the protein sequences as well as information about the positions and phases of all introns in the description line. These files were derived from the protein form of EID, the "gb121.pEID" file (23). Stand-alone gapped BLAST 2.0 binaries compared pairwise all proteins in one sample with all proteins in another using the BLAST option (−v1–b1) producing only the single best match. The BLAST outputs were automatically processed by a PERL program CIP.pl (Comparison of Intron Positions), which takes a BLAST protein alignment as input, marks the introns of the genes for the aligned protein pair onto the alignment, and compares these intron positions. Further, CIP.pl counts the number of common introns (those introns which have identical positions in each aligned protein), sliding introns (those locations that are within 6 nt of each other), and unique introns (neither common nor sliding). In addition, CIP.pl calculates the distribution of common, sliding, and unique introns along the genes. We used the following CIP.pl parameters: only alignments with BLAST scores of 55 bits or higher are accepted (*e* value of $\approx 10^{-6}$), and (*ii*) each gene is counted only once (if there are several alignments for the same protein, only the alignment with the maximal homology score is used). All of the alignments with marked introns and the entire data set is available on our web site, www.mcb.harvard.edu/gilbert/CIP.

Results

Distribution of Common, Sliding, and Unique Intron Positions. Table 1 presents the results of the comparison of intron positions between animal and plant genes, animal and fungal genes, and plant and fungal genes. Nearly 10% of the 8,737 animal introns match plant positions in the 1,514 alignments of gene products purged to the 20% level, 14-fold higher than the expectation for random matches. Because there are 35% more plant introns in these alignments, only 7% of the plant introns match animal positions. In ≈700 alignments, 15% of fungal introns match animal positions, whereas 13% match plant positions. The ratio of common introns to sliding introns (nonidentical positions within 6 nt) ranges from 1.4 to 3, again suggesting that these matches are not caused by chance (because then one expects 12-fold more sliding than identical matches).

This analysis of the 20%-purged data eliminates much of the bias in gene representations in the database. However, the proportion of paralogous pairs will be high. To examine this issue, we also calculated the matches in data purged only to a 95% threshold (data not shown) to maximize the orthologous pairs in all these comparisons, so that, for example, animal α-tubulins will be compared with plant α-tubulins (rather than possibly comparing animal α- to plant β-tubulin in the 20%-purged sample.) The 95% threshold set, however, has large numbers of repeated genes from families that are widely studied. Nonetheless, in that animal–plant comparison, 11% of the 16,091 animal introns in 3,901 alignments match plant positions. Likewise, all of the other matching fractions were very similar (data not shown). Furthermore, we ran all comparisons the other way around, plant–animal for instance, to see if any lack of symmetry in the BLAST procedures would affect the outcome. The results were very similar (data not shown).

The distribution of matches is not what one would find from

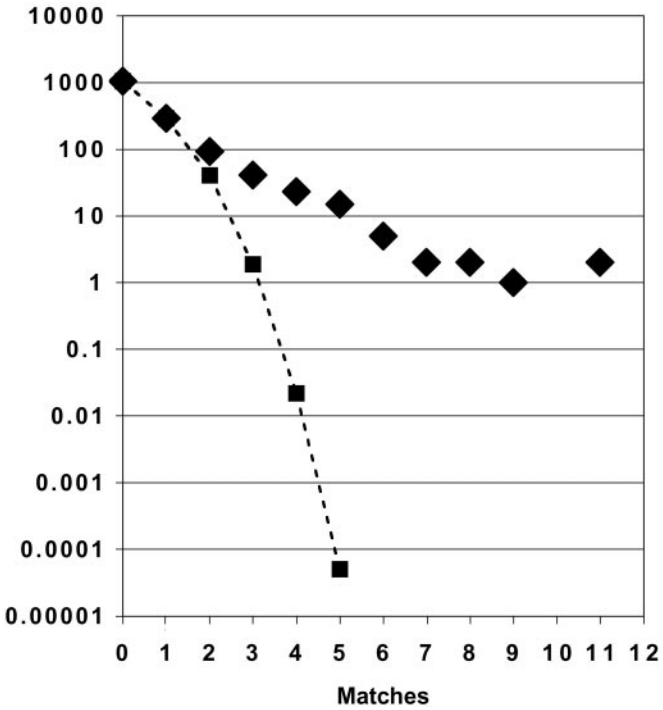

Fig. 1. Comparison of the animal–plant matches to the Poisson expectation. The diamonds show the number of alignments at each set of matches. The lower curve, with the squares, is the Poisson estimate fitted to the zero and single hit numbers.

Table 2. Animal–plant gene matches with the highest number of common intron positions

Gene no.	Protein description	Length of animal–plant alignments, aa	No. of introns							
			Animal–plant comparison				Fungal introns			
			Common	Sliding	unANM	unPLN	comA + P	comANM	comPLN	unFNG
1	RAN binding protein (importin 7)	1,009	11	2	10	8	0	1	0	0
2	Suppressor of actin 1 (LIM-domains protein)	507	11	1	5	4	0	0	0	1
3	DNA mismatch repair protein	690	9	0	8	6	2	0	0	1
4	Ubiquitine thiolesterase 5	674	8	1	6	5	0	1	0	0
5	Glutathione synthase	462	8	0	3	2	0	0	1	1
6	Importin (karyopherin)	484	7	1	4	0	1	0	0	1
7	Serine palmitoyltransferase	473	7	0	3	4	0	1	0	0
8	DNA polymerase epsilon subunit	2,237	6	4	37	38	0	0	1	1
9	Phenylalanine tRNA synthetase	450	6	1	5	7	1	1	0	0
10	*N*-myc downstream-regulated gene	283	6	1	4	3	0	0	0	0
11	Protein kinase C substrate 80K-H	289	6	0	3	2	3	0	0	2
12	Deoxyhypusine synthase	352	6	0	2	0	2	0	0	0
13	Xanthine dehydrogenase	1,310	5	3	26	5	0	0	0	3
14	N_arginine dibasic convertase	823	5	3	18	14	0	0	0	0
15	DNA polymerase delta subunit	406	5	1	3	6	1	0	0	0
16	Membrane-bound aminopeptidase P	562	5	1	12	8	0	0	0	2
17	DNA transestrate (topoisomerase homolog)	348	5	0	6	8	2	3	1	5
18	G1→S phase transition protein	427	5	0	6	7	0	0	0	1
19	CDP-diacylglycerol synthase	377	5	0	6	4	2	0	0	0
20	Eukaryotic translation initiation factor 3	415	5	0	6	2	0	0	0	0
21	Novel protein CGI-09	423	5	0	5	7	0	0	0	0
22	Cyclin C	248	5	0	5	3	3	0	0	0
23	Nucleolar protein 5A	422	5	0	3	1	0	0	0	1
24	Phosphomannomutase	241	5	0	2	4	0	1	0	0
25	Hypothetical protein	396	5	0	2	1	0	0	0	0
26	Triose phosphate isomerase	247	5	0	1	3	1	0	0	4
27	Hypothetical protein	291	5	0	0	2	2	0	0	0
28	Folylpolyglutamate synthetase	480	4	2	6	7	0	1	0	0
29	Mitogen-activated protein kinase	277	4	2	5	5	3	0	0	1
30	Brain-enriched WG-repeat protein	503	4	1	5	9	1	0	1	1
Total			178	24	207	175	24	9	4	25

The 30 genes with the highest animal–plant matches. The table lists the name of the gene, the length of the alignment, and the common, sliding, and unique animal (unANM) and plant (unPLN) introns within the alignment. The table then lists the fungal introns common to all three (comA + P), common to animal (comANM), common to plant (comPLN), and those not common to either (unFNG). A total of 44% of the animal introns are common; 47% of the plant introns are common; 53% of the fungal introns match animal positions, 45% match plant positions, and 39% match both.

a random process. Fig. 1 compares the observed number of matches in each alignment with the Poisson expectation: there are far more genes with high numbers of matches than would arise from the Poisson curve.

Table 2 lists the top 30 genes with the greatest common animal–plant intron positions; (the complete set of data are on our web site). The top match is the RAN-binding protein (importin 7). Fig. 2 shows the alignment with marked intron positions. Within the region of the alignment, the animal gene has 23 introns, whereas the plant has 21. Eleven introns have common positions, and two intron pairs are marked as slid. Ten animal and eight plant introns have no matches.

Common Versus Sliding Intron Positions. The two animal–plant sliding-intron positions in the RAN-binding protein (blue in Fig. 2) lie close to gaps in the machine alignment. Fig. 2 *B* and *C* show that these regions of the alignment can easily be corrected to make these two introns common. Thus, this animal–plant gene pair has in reality 13 common intron positions. Visual inspection of the sliding intron positions for the 30 best animal–plant matches showed that in 50% of the cases the alignment regions in the vicinity of sliding intron positions can be corrected to transform sliding introns to common ones.

This notion, that many of the "sliding" introns are actually common ones misaligned is supported by the data in Table 3, which shows the ratio of common to sliding intron positions as a function of the alignment stringency. When the homology between the animal and plant protein pairs is high (fraction of identical amino acids >65%), there is a 7.5-fold excess of common intron positions over sliding ones. The weaker the homology, the smaller is the excess of common intron positions over sliding. For the weakest homology (when the fraction of identical amino acids in the alignments is <25%) there is even a 5% excess of sliding over common positions. All of these data testify that the majority of the sliding intron positions arise because of the imperfection of the protein alignments and that they are, in reality, common.

Common Introns Among Animal, Plant, and Fungal Genes. Table 2 shows the 30 genes with the highest number of common animal–plant intron positions. For these 30 genes, 44% of the animal intron positions match to plant positions, and 47% of the plant

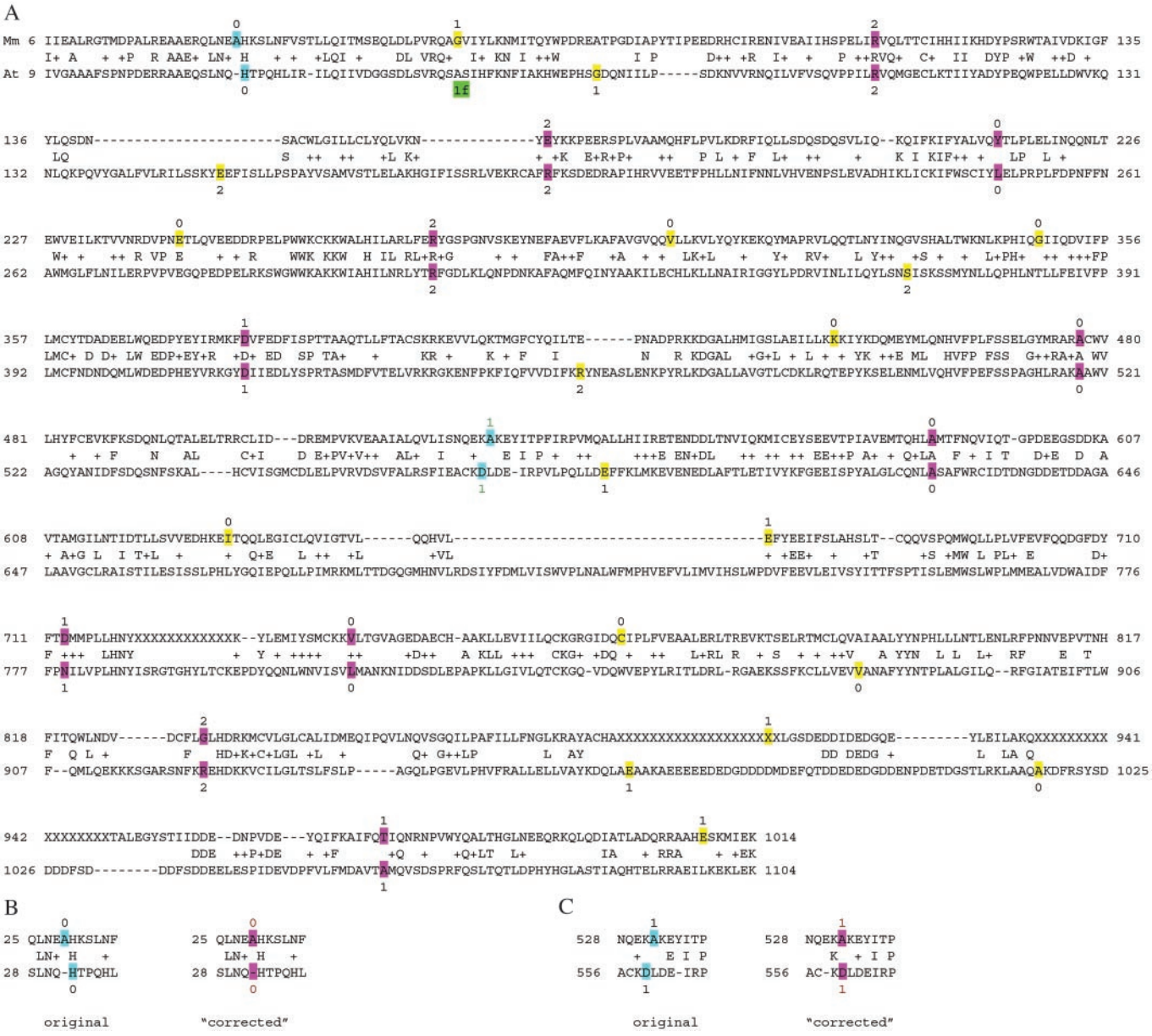

Fig. 2. BLAST 2.0 alignment of the RAN-binding protein. The upper sequence is the *Mus musculus* MMU278435 gene; the lower sequence is the *Arabidopsis thaliana* ATF17J16 gene. (*A*) The positions of common introns on the protein sequences are marked in red, sliding introns are marked in blue, and unique introns are marked in yellow. The phases of animal introns are shown above their marked positions and the phases of plant introns are shown below. The position and phase of one mapped fungal intron is shown below the aligned sequence on the green background; the digit followed by ''f'' indicates the phase of the fungal intron. (*B* and *C*) Segments of the alignment containing sliding intron positions are followed by the corrections to the alignment that transform sliding intron positions to common positions.

intron positions match animal ones. We used the BLAST pairwise alignment to compare these 30 genes to fungal-intron-containing genes and then mapped the fungal introns from the best fungal–animal match onto the 30 animal–plant pairs. Fig. 2 shows an example of such an animal–plant–fungal comparison, and Table 2 presents the full set of data on matched intron positions. Of the 62 fungal introns mapped onto the 30 best animal–plant pairs, 33 (53%) match animal introns, 28 (45%) match plant introns, and 24 (39%) match both animal and plant introns. These numbers are considerably higher than the average match between fungal–animal and fungal–plant pairs (Table 1) and, thus, testify to the antiquity of these common introns.

Intron Phase Distribution. The phase distributions of the animal, fungal, and plant introns (Table 1) are in complete accordance with the data of Long *et al.* (5). Interestingly, the phase distribution of the common animal–plant introns is intermediate between that of animal introns and that of plant introns. Likewise, the phase distribution of the common plant–fungal introns lies between the parental distributions.

Discussion

This large-scale examination of animal, plant, and fungal genes shows that 10% of animal introns have the same positions as plant introns, and a further 7% lie within six nucleotides from plant introns (sliding positions). Additionally, we have shown that the majority of sliding introns are caused by the imperfection of machine alignments, appearing more and more as the sequence similarity becomes weaker and weaker, and are most

Table 3. As the alignment becomes less sure, the fraction of sliding introns increases

Range of alignment homology, % identity	No. of common introns	No. of sliding introns	Ratio common/sliding
15–25	106	112	0.95
25–35	280	257	1.09
35–45	248	160	1.55
45–55	134	63	2.13
55–65	60	21	2.86
65–100	15	2	7.50

likely common positions in the animal–plant genes. This conclusion about sliding agrees with that of Stoltzfus *et al.* (25), who claimed that most sliding introns "are found to be artefactual" and caused by alignment errors, database errors, and errors in the precise determination of intron positions. These findings give us a reason to pool the common and half of the sliding groups to assert that overall 14% of animal-intron positions match plant ones and that 17–18% of fungal introns match animal or plant positions.

The matching of intron positions is high above expectation if all of the introns had entered the sequences randomly at the nucleotide level. Our calculations show that for the two genes at the top of the Table 2, even for the most favorable scenario of intron insertion into a restricted number of proto-splice sites, the probability of so many matches is 10^{-3}. Furthermore, the excess of common over "sliding" positions (observed as a 1.4- to 3-fold excess, but correctable to a 3- to 6-fold excess) is not expected; random insertion would predict a 12-fold excess of "sliding" positions (± 6 nt).

Presently, there are two different explanations for the common intron positions. The introns-late hypothesis postulates that all introns were inserted recently into "proto-splice sites" of genes. On this view, the common intron positions between animal and plant genes occur by a chance coincidence of nonrelated and independently inserted introns. The alternative hypothesis explains the common intron positions as ancestral: ancient introns, which existed before the divergence of the animal and plant kingdoms. If the second scenario were true, then the common animal–plant intron positions should be also enriched in fungal genes, because the evolutionary separation of fungi and animals occurred after the separation of animals and plants. Indeed, the results in Table 2 demonstrate that to be the case. For the 30 nonrelated genes with the highest numbers of common animal–plant introns, 60% of the fungal introns have positions common to animal and/or plant introns, and 39% of fungal introns are common simultaneously to both plant and animal introns. This exceptionally high abundance of introns with positions common to all three taxa of animals, plants, and fungi strongly supports the antiquity of these common intron positions.

Even though the raw numbers are not high compared with the assumption that introns might have entered a limited number of "proto-splice sites," the details of the distribution do not agree with a proto-splice model. Recent data by Endo *et al.* (26) shows that intron insertion is not always restricted to proto-splice sites. Moreover, because coding sequences undergo continual evolutionary changes, proto-splice sites for intron insertions will not have invariable positions throughout evolution. The considerable differences in preferences for specific nucleotides in the third position of codons (so-called codon bias) of animal, fungal, and plant genes, will magnify the change in distribution of proto-splice sites between the taxa. Species variability in the genomic nucleotide composition (27) also diversifies the distribution of proto-splice sites inside the coding sequences of different species. So, a preference for intron insertion into proto-splice sites should correspond to a model of completely random intron insertion into a limited number of sites.

However, the distribution of intron matches is not at all what one expects for a random coincidence model. The number of genes with high animal–plant matches is far above the Poisson estimate. Furthermore, these genes, enriched for animal–plant matches (44% of animal introns match) are also enriched for fungal–animal–plant matches (39% of fungal introns match animal and plant positions). On any random model, the coincidences in the fungal genes should be independent. This high intron matching in separate taxa is in flat contradiction to the expectation of random independent insertion.

Could these common intron positions arise from a recent horizontal gene transfer between the taxa? We inspected the phylogenic relationships of the 30 animal, plant, and fungal proteins and concluded that this is not the case. At best, only for a few proteins an ancient horizontal gene transfer might be arguable. Therefore, recent horizontal gene transfer is not the main reason for the common intron positions in Table 2.

The majority of contemporary introns have unique positions in the animal, fungal, or plant taxa, and, likely, a considerable portion of them may be recently acquired. Even among common intron positions there will be some fraction of recently inserted ones, which have the same positions by chance. On the other hand, intron loss has been occurring throughout evolution, which decreases the number of common introns. Thus, even knowing the proportion of common intron positions today, it is impossible to realistically estimate the number of ancient introns. A lower approximation would be the figure of 14% of the animal introns given above, but we think it likely that 60–80% of the original introns have been lost with a similar number of additions.

The observed phase distribution of common introns is similar to that for all introns in that all three phases are represented, with phase zero being intermediate between the parental taxa. The common (ancestral) introns appearing in animals, plants, and fungi, are not restricted to phase zero. Thus all three phases would have been populated early in eukaryotic evolution. However, we cannot exclude the possibility that some dramatic changes in intron phase distribution had occurred much earlier around the time of the eukaryotic origin, as we have proposed elsewhere (9–11).

In summary, we have identified a large number of matched positions among animal, plant, and fungal introns. The 30 genes with the highest number of matches show a 44% match between animals and plants. Furthermore, the corresponding fungal introns for these genes also show a high matching, 39%, to both animals and plants simultaneously. This set of introns is most likely to be ancestral, lying within these genes before the separation of animals, plants, and fungi.

We thank Drs. W. F. Doolittle and M. Long for critical comments and valuable suggestions on the manuscript. A.F.M. was supported by Fulbright Grant 24953.

1. Logsdon, J. M., Jr. (1998) *Curr. Opin. Genet. Dev.* **8,** 637–648.
2. Cavalier-Smith, T. (1985) *Nature* **315,** 283–284.
3. Palmer, J. D. & Logsdon, J. M., Jr. (1991) *Curr. Opin. Genet. Dev.* **1,** 470–477.
4. Dibb, N. J. & Newman, A. J. (1989) *EMBO J.* **8,** 2015–2021.
5. Long, M., de Souza, S. J., Rosenberg, C. & Gilbert, W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 219–223.
6. Long, M. & Rosenberg, C. (2000) *Mol. Biol. Evol.* **17,** 1789–1796.
7. Gilbert, W. (1978) *Nature* **271,** 501.
8. Doolittle, W. F. (1978) *Nature* **272,** 581–582.
9. de Souza, S. J., Long, M., Klein, R. J., Roy, S., Lin, S. & Gilbert, W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5094–5099.
10. Roy, S. W., Nosaka, M., de Souza, S. J. & Gilbert, W. (1999) *Gene* **238,** 85–91.

11. Fedorov, A., Cao X., Saxonov S., de Souza S., Roy, S. W. & Gilbert, W. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 13177–13182.
12. Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 901–905.
13. Kersanach, R., Brinkmann, H., Liaud, M.-F., Zhang, D.-X., Martin, W. & Cerff, R. (1994) *Nature* **367,** 387–389.
14. Shah, D. M., Hightower, R. C. & Meagher, R. B. (1983) *J. Mol. Appl. Genet.* **2,** 111–126.
15. Marchionni, M. & Gilbert, W. (1986) *Cell* **46,** 133–141.
16. Gilbert, W., Marchionni, M. & McKnight, G. (1986) *Cell* **46,** 151–153.
17. Logsdon, J. M., Jr., Tyshenko, M. G., Dixon, C., D.-Jafari, J., Walker, V. K. & Palmer, J. D. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8507–8511.
18. Nixon, J. E. J., Wang, A., Morrison, H. G., McArthur, A. G., Sogin, M. L., Loftus, B. J. & Samuelson, J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 3701–3705.
19. Fast, N. M., Roger, A. J., Richardson, C. A. & Doolittle, W. F. (1998) *Nucleic Acids Res.* **26,** 3202–3207.
20. Fast, N. M. & Doolittle, W. F. (1999) *Mol. Biochem. Parasitol.* **99,** 275–278.
21. Archibald, J. M., O'Kelly, C. J. & Doolittle, W. F. (2002) *Mol. Biol. Evol.* **19,** 422–431.
22. Katinka, M. D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., *et al.* (2001) *Nature* **414,** 450–453.
23. Saxonov, S., Daizadeh, I., Fedorov, A. & Gilbert, W. (2000) *Nucleic Acids Res.* **28,** 185–190.
24. Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A. & Wheeler, D. L. (1999) *Nucleic Acids Res.* **27,** 12–17.
25. Stoltzfus, A., Logsdon, J. M., Palmer, J. D. & Doolittle, W. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 10739–10744.
26. Endo, T., Fedorov, A., de Souza, S. J. & Gilbert, W. (2002) *Mol. Biol. Evol.* **19,** 521–525.
27. Karlin, S. & Burge, C. (1995) *Trends Genet.* **11,** 283–290.