

# The Evolutionary Gain of Spliceosomal Introns: Sequence and Phase Preferences

Wei-Gang Qiu,<sup>1</sup> Nick Schisler,<sup>2</sup> and Arlin Stoltzfus<sup>3</sup>

Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland

Theories regarding the evolution of spliceosomal introns differ in the extent to which the distribution of introns reflects either a formative role in the evolution of protein-coding genes or the adventitious gain of genetic elements. Here, systematic methods are used to assess the causes of the present-day distribution of introns in 10 families of eukaryotic protein-coding genes comprising 1,868 introns in 488 distinct alignment positions. The history of intron evolution inferred using a probabilistic model that allows ancestral inheritance of introns, gain of introns, and loss of introns reveals that the vast majority of introns in these eukaryotic gene families were not inherited from the most recent common ancestral genes, but were gained subsequently. Furthermore, among inferred events of intron gain that meet strict criteria of reliability, the distribution of sites of gain with respect to reading-frame phase shows a 5:3:2 ratio of phases 0, 1 and 2, respectively, and exhibits a nucleotide preference for MAG GT (positions  $-3$  to  $+2$  relative to the site of gain). The nucleotide preferences of intron gain may prove to be the ultimate cause for the phase bias. The phase bias of intron gain is sufficient to account quantitatively for the well-known 5:3:2 bias in phase frequencies among extant introns, a conclusion that holds even when taxonomic heterogeneity in phase patterns is considered. Thus, intron gain accounts for the vast majority of extant introns and for the bias toward phase 0 introns that previously was interpreted as evidence for ancient formative introns.

## Introduction

Eukaryotic protein-coding genes often are interrupted by sequences that are excised from the RNA transcript of the gene by a ribonucleoprotein spliceosome (Newman 1994) and that typically conform to the sequence GT...AG. Soon after their discovery, these “spliceosomal” introns became the subject of debate between those who attribute their pervasive presence to a primordial period of gene formation (Doolittle 1978; Darnell and Doolittle 1986; Gilbert, Marchionni, and McKnight 1986) and those who attribute it to widespread gain of an invasive element (Cavalier-Smith 1991; Palmer and Logsdon 1991).

In the past, this dispute focused largely on generalizations from limited examples. More recently, the capacity to analyze massive amounts of sequence data has brought a new focus on the evolutionary causes of broadly observed patterns, such as the tendency for introns to occur in “phase 0” positions, that is, between codons (Smith 1988); the tendency for phases of nearby introns to match (Fedorov et al. 1992; Long, Rosenberg, and Gilbert 1995; Fedorov et al. 1998); and unusual amino acid and protein properties near intron sites (Craik, Rutter, and Fletterick 1983; Fichant 1992; de Souza et al. 1996, 1998).

Such patterns have not been explained satisfactorily. On the one hand, advocates of the “introns-late” view tend to assume—on the basis of phylogenetic analyses indicating that most introns have restricted distributions,

suggesting a gain event long after the origin of the gene (Cho and Doolittle 1997; Rzhetsky et al. 1997; Stoltzfus et al. 1997)—that any widespread patterns of nonrandomness cannot be caused by formative introns, but must be the result of ongoing processes of intron gain and loss. As has long been recognized (Hickey, Benkel, and Abukashawa 1989; Dibb 1993; de Souza, Long, and Gilbert 1996), a process of intron gain with nucleotide sequence preferences for sites of gain would lead to other preferences: for the reading-frame phase in which the target sites are most common (Long et al. 1998), for intron phase autocorrelation (Long and Rosenberg 2000), for amino acids whose codons are favored by nucleotide preferences (Hickey, Benkel, and Abukashawa 1989), and for protein structural features favored by such amino acid preferences (Stoltzfus et al. 1994). Nevertheless, this chain of reasoning is speculative and remains largely unexplored. One simply cannot invoke a principle of “parsimony” to argue in favor of widespread intron gain, because the principle of parsimony does not apply if loss and gain events cannot be treated equally. Furthermore, available evidence on nucleotide preferences for intron gain consists mainly of anecdotes (Lee, Stapleton, and Huang 1991; Frugoli et al. 1998; Logsdon, Stoltzfus, and Doolittle 1998; Funke et al. 1999; Torki et al. 2000) consistent with the original study of Dibb and Newman (1989), along with more systematic results for the special case of rDNA introns (Bhattacharya et al. 2000). Phase distributions predicted from crude versions of an intron target sequence do not account quantitatively for known phase distributions (Long et al. 1998; Long and Rosenberg 2000).

On the other hand, although advocates of formative introns argue that the formation of new genes by exon shuffling is a pervasive mechanism that accounts for many or most protein-coding genes (de Souza et al. 1998), the patterns observed in genes do not conform easily to what is known about exon shuffling. The common features of well-known cases of exon shuffling (Patthy 1991) are that (1) introns tend to occur near chimeric junctions identifiable from protein sequence alignments, (2) nearby introns tend

<sup>1</sup> Present address: Department of Biological Sciences, Hunter College, City University of New York, NY.

<sup>2</sup> Present address: Department of Biology, Pomona College, Claremont, Calif., and Department of Biology, Furman University, Greenville, SC.

<sup>3</sup> Present address: Biotechnology Division, National Institute of Standards and Technology, Gaithersburg, MD.

Key words: intron gain, evolution, intron phase, proto-splice site, target site.

E-mail: arlin.stoltzfus@nist.gov.

*Mol. Biol. Evol.* 21(7):1252–1263. 2004

doi:10.1093/molbev/msh120

Advance Access publication March 10, 2004

to have the same phase (i.e., phases are positively autocorrelated), and (3) the intron sites (and chimaeric junctions) tend to correspond to the boundaries of protein domains (spatially discrete protein segments that tend to maintain a native structure when separated). Of these hallmarks of exon shuffling, the only one that is a general feature of intron-containing genes is the phase autocorrelation (Long, Rosenberg, and Gilbert 1995), which has other possible explanations, including intragenic duplications (Fedorov et al. 1998) and local biases in intron gain preferences (Long and Rosenberg 2000). Claims of widespread protein sequence chimaerism (Dorit and Gilbert 1991) have not withstood scrutiny (Doolittle 1991), and claims of weak correlations between intron locations and protein structure (de Souza et al. 1996), although perhaps statistically significant, invoke “modules” that have no known significance in protein biology (i.e., they are known only by the computer algorithms designed to identify them). Taken by itself, the excess of phase 0 introns (i.e., introns that fall between codons) can be interpreted as evidence of an ancient population of introns that evolved from primordial intergenic spacers, according to the classic “introns-early” view (Doolittle 1978). Yet, the argument that intron-protein correlations betray a formative role for introns is not helped by the suggestion of de Souza, et al. (1998) that such correlations apply only to phase 0 introns, because this fits neither the pattern seen in undisputed cases of exon shuffling, which is a preference for phase 1 (Patthy 1991), nor the pattern of phase autocorrelation seen among genes in general, which involves all three intron phases (Long, de Souza, and Gilbert 1995).

The most direct approach to resolving the role of intron gain in accounting for the distribution of present-day introns would be a systematic analysis of intron gain and loss in a large set of eukaryotic protein-coding genes, based on a model that does not treat gain and loss events equally. The individual steps required for such an analysis—family assignment, multiple sequence alignment, collation of intron data, inference of a phylogenetic tree, and reconstruction of loss and gain events—are relatively obvious, yet such an approach represents a formidable technical challenge, not only of integrating diverse procedures but also of effectively replacing the expert judgement commonly used in key steps of multiple sequence alignment and phylogenetic inference (e.g., Barton [2001]).

Here we present a systematic evolutionary analysis of the inheritance, gain, and loss of introns in a set of 10 families of protein-coding genes comprising 677 genes, 1,868 introns, and 488 distinct intron sites, carried out in a semiautomated fashion. The need for expert supervision is minimized by restricting the scope of sequence families so that they include only readily alignable members and by computing quality or reliability scores (for alignment columns, tree branches, and inferred evolutionary events) that are then used to select the aspects of a reconstructed evolutionary history that are the most reliable.

This analysis yields three major conclusions. First, using a Bayesian approach to estimate the probability of occupancy of ancestral intron sites, it is shown that, in the overwhelming majority of cases, intron sites occupied in present-day eukaryotic gene families are highly unlikely to

have been occupied in the most recent common ancestors of these gene families. That is, the presence of introns in extant genes is overwhelmingly caused by the gain of introns subsequent to the origin of the gene, rather than by the inheritance of formative introns. Second, when 336 of the most reliably inferred events of intron gain are separated by the reading frame phase of the intron site, the frequency distribution seen for phases {0, 1, 2}, namely, {0.496, 0.296, 0.207}, is significantly different from a uniform distribution but not significantly different from the frequency distribution of phases of extant introns {0.503, 0.292, 0.204}. Thus, phase-biased intron gain is sufficient, by itself, to account for extant phase frequencies. Third, for the same set of 336 most-reliable events of intron gain, outgroup sequences reveal a highly significant bias toward the nucleotides MAG GT (M = A or C; sequence positions from -3 to +2 relative to the site of gain), demonstrating that intron gain exhibits preferences for flanking exonic nucleotides.

## Materials and Methods

### Sequence Data and Family Assignments

Data for individual genes were processed using procedures developed for IDB (Schisler and Palmer 2000), a nonredundant database that contains only genomic sequences of eukaryotic protein-coding genes. Entries were assigned to provisional families by way of IDB cross-references to Pfam release 5.3 (Bateman et al. 2000). Ten such Pfam families were chosen based on (1) taxonomic diversity of members, and (2) use in prior studies that purport to find evidence for an ancient formative role of introns (de Souza et al. 1998). Entries with incomplete sequences or with non-GT-AG introns (indicating either a noncanonical intron or, more commonly, an error) were discarded.

On the basis of the finding that unsupervised alignments typically achieve high accuracy if the sequences are at least 30% similar (Briffeuil et al. 1998; Thompson, Plewniak, and Poch 1999), the initial Pfam families were subdivided into less inclusive clusters whose members have at least 30% pairwise amino acid similarity. Using software implemented in R (Ihaka and Gentleman 1996), clusters were identified visually based on multidimensional scaling of pairwise similarity scores computed with ClustalW (Thompson, Higgins, and Gibson 1994). Clusters with fewer than 20 members were discarded as being too small to warrant further analysis.

The remaining clusters are designated as families for the present purposes. To each such family is added a manually chosen outgroup (a prokaryote homolog, if available), for the purpose of rooting the family tree. The outgroup is included solely for the purpose of polarizing the tree and is not used in the inference of ancestral intron states.

### Multiple Alignment and Phylogenetic Analysis

For each family, protein sequences were multiply aligned using ClustalW (Thompson, Higgins, and Gibson 1994) with all combinations of a gap-open penalty in the set



**Table 1**  
**Gene Families**

Gene Name	Pfam ID	Member Sequences	Intron Positions
Actin	PF00022	163	64
Alcohol dehydrogenase (cluster 1)	PF00107	76	29
Alcohol dehydrogenase (cluster 2)	PF00107	40	19
Aldehyde dehydrogenase	PF00171	43	109
Cu/Zn superoxide dismutase	PF00080	32	33
Elongation factor 1 $\alpha$	PF00009	43	34
Glyceraldehyde 3-phosphate dehydrogenase	PF00044	71	62
Heat shock protein 70	PF00012	155	87
Mn/Fe superoxide dismutase	PF00081	27	31
Triosephosphate isomerase	PF00121	27	20
Totals		677	488

called “seqlogo” by G. E. Crooks, G. Hon, J.-M. Chandonia, and S.E. Brenner (<http://weblogo.berkeley.edu>).

#### Bayesian Inference of Intron Evolution

A matrix of intron presence/absence data is determined for each family from the multiple alignment of coding regions. This matrix, along with the family tree, is used to estimate ancestral states of introns, as well as rates of loss and gain. To allow an explicit approach to uncertainty, we applied a Bayesian network approach (Spiegelhalter et al. 1993; Jensen 1996; Cowell et al. 1999) to infer posterior distributions of parameters of a two-state Markov model of presence/absence transitions along branches of a known tree. In such a model, the probabilities of a change from state  $i$  to  $j$ , either gain ( $i=0, j=1$ ) or loss ( $i=1, j=0$ ), for an interval of time  $t$  are  $\Pr(j=1|i=0) = [\alpha/(\alpha + \beta)] [1 - e^{-(\alpha+\beta)t}]$  and  $\Pr(j=0|i=1) = [\beta/(\alpha + \beta)] [1 - e^{-(\alpha+\beta)t}]$ ,  $\alpha$  being the rate of gain and  $\beta$  the rate of loss (Harvey and Pagel 1991). Within a family, rates are assumed to be constant across the tree and across intron sites. The prior probability of presence of an intron at the root node is defined as  $\Pr(\text{root}=1) = \alpha/(\alpha + \beta)$ , which, in principle, implies a steady-state assumption regarding intron density, although in practice, the data are sufficiently abundant in the present case that the posterior probability distribution is dominated much more by how well the data are explained (i.e., by the likelihood) than by the prior distribution. This approach allows for estimation of posterior distributions of  $\alpha$ ,  $\beta$ , the ratio  $\alpha/\beta$ , and intron states at internal nodes, based on a given tree and a set of observed intron states.

This approach is implemented as follows. Intron data for a single family (excluding the outgroup sequence used to root the tree), along with the phylogenetic tree, are extracted from a local database and written in NEXUS format (Maddison, Swofford, and Maddison 1997). A separate program reads the NEXUS file and generates task-specific code for the BUGS (Bayesian Analysis Using Gibbs Sampling) command interpreter (Spiegelhalter et al. 1996). The BUGS command interpreter carries out Metropolis-Coupled Markov chain Monte Carlo estimation

of posterior probability distributions of parameters of the specified model, producing output that is imported into SPAN for further analysis. This implementation was validated by comparison with analytically determined results for small test cases.

## Results

### Sequence Families

Of the 10 initial groups of sequences chosen for analysis, each corresponding to a Pfam family, eight yielded single families along with small numbers of outliers that were discarded. The Pfam family for alcohol dehydrogenase (PF00107) yielded two families. No suitable family (i.e., a cluster of at least 20 members with at least 30% pairwise similarity; see *Materials and Methods*) was found for the highly diverse glutathione-S-transferases (PF00043). The resulting 10 families, described in table 1, comprise 677 sequences with introns at 488 distinct locations.

For each family, a multiple alignment (with column-wise stability scores) was computed, and used to infer a phylogenetic tree (with posterior probabilities for each branch), as described in *Materials and Methods*. Intron data for each family were collated by means of the multiple alignment. Figure 1 shows the phylogeny and intron data for one of the smallest families, the manganese-iron superoxide dismutases (Mn-Fe SODs).

### Statistical Inference of Intron Gain and Loss

Bayesian posterior probabilities for the presence of each intron at each node of a family tree, as well as posterior distributions for the rates of gain and loss, and for their ratio are estimated as described in *Materials and Methods*. The rate of loss is significantly greater than the rate of gain. For example, the most-probable ratio of loss to gain for the Mn-Fe SOD family is  $r = 6.4$  (95% confidence interval, 3.7 to 9.9), where one assumes that the set of all possible sites is the set of all observed sites. The ratio of loss to gain would be higher still if the model were to allow for an unknown number of intron sites that are not occupied in any included gene.

A distinctly useful result of applying this approach is that it yields, for each intron, a probability that the intron was present in the most recent common ancestor of the gene family. The distribution of these probabilities is shown in figure 2. For the vast majority of intron sites, the chance of occupancy at the base of the tree is extremely low: for 75% of introns, the chance is less than 0.0060; for 95%, it is less than 0.041; and for 99%, it is less than 0.15.

The inferred history of an individual intron site can be represented graphically in the manner shown in figure 3. The same graphical convention is used in figure 4 to show inferred histories for the subset of reliably aligned introns in the Mn-Fe SOD family (the manganese-iron superoxide dismutases family shown in figure 1). For most introns, a single event of gain can be assigned to a unique branch with high reliability. Even when multiple events are indicated, there is often a single scheme of high probability,

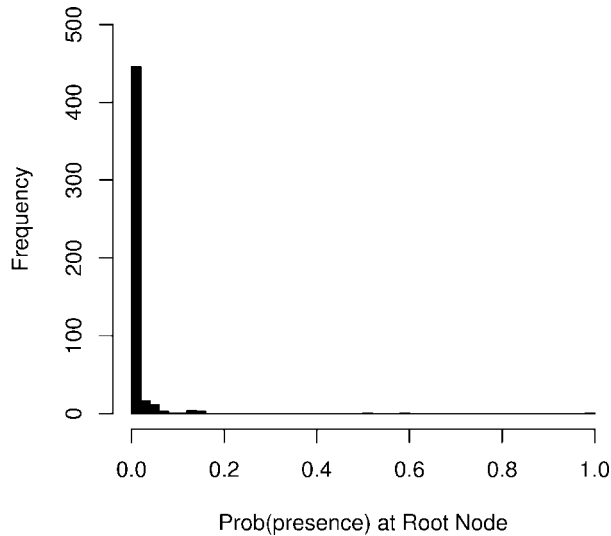


FIG. 2.—Distribution of the probability of presence of introns at the root node. The probability that an intron is present in the common ancestor of all of the eukaryotic sequences in its family is shown for all 488 intron sites in the 10 sequence families. The median probability of presence is 0.0020; the much higher mean value of 0.012 reflects three introns whose probability of presence in the common ancestor is 0.5 or more.

such as the double events of gain apparent for introns at 265-0 and 280-0 (fig. 4, panels 9 and 10). Loss events are suggested in three cases: 92-1 and 275-1 (panel 4) and 250-0 (panel 8).

#### Identification of a Most-Reliable Set of Gain Events

In the absence of a general theoretical framework that allows probabilities to be assigned to inferences dependent on both an uncertain phylogeny and an uncertain alignment, a heuristic approach may be used instead. In particular, one may use indicators of reliability or quality to divide results and associated inferences into categories of greater and lesser quality. This approach depends only on the rank order of a given quality indicator and not on the absolute values. Figure 5 shows the distributions of three indicators of reliability or quality: the alignment stability, the phylogenetic branch support, and the probability of events of intron gain and loss. The alignment stability score assigned to an intron site is the SOAP score for the relevant column in the multiple alignment of protein sequences, where the relevant column either contains the intron site (for phase 1 and 2 introns) or follows it (for phase 0 introns). The branch support value (analogous to a bootstrap value) is the posterior probability of the branch to which the intron gain or loss event has been assigned.

For each quality indicator, one may choose a threshold value that separates the best from the rest, so as to allow for a “high-grading” approach in which the most trusted conclusions rest on the most trusted results. By applying thresholds of alignment stability of 0.95 or higher, branch support of 0.99 or higher, and probability of change of 0.8 or higher, a set of 336 most reliable events of intron gain

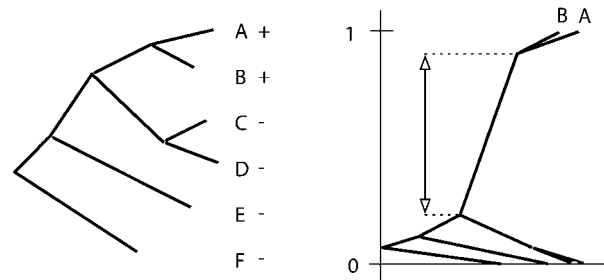


FIG. 3.—A graphical means to display probabilities of binary states on a phylogenetic tree. At left is shown a phylogenetic tree for six genes (A to F), next to a column indicating the presence or absence of an intron. At right is a figure with the same information displayed on Cartesian coordinates, including the tree topology, the branch lengths (in the horizontal dimension), and the presence and absence of introns in extant genes, represented in the vertical dimension. The terminal nodes representing genes A and B are positioned at  $Y = 1$  to indicate that the intron is present, and the remaining terminal nodes are at  $Y = 0$ . Internal nodes can be positioned so that the vertical position represents the probability that the intron was present in the given ancestor. Thus, each branch has a vertical component that represents the change in probability of presence of the intron: the larger the change, the greater the chance that an event of gain or loss occurred on that branch.

were identified. For comparison, the total number of gain events is 579 when a gain event is defined as a change in probability of 0.5 or higher, and given no restriction on alignment stability or branch support.

The manner in which the use of reliability indicators is expected to increase the quality of the analysis can be illustrated by reference to specific aspects of the Mn-Fe SOD results in figures 1 and 3. Figure 1 reveals a highly nonrandom set of introns at 33-2, 41-2, 42-2, and 43-2, which are identical in phase, closely spaced, and phylogenetically clustered. This pattern suggests a process, real or artifactual, that disperses the same intron to a local range of alignment positions. One candidate for a real process is the often discussed but poorly evidenced possibility of intron “sliding” (Stoltzfus et al. 1997; Rogozin, Lyons-Weiler, and Koonin 2000), but the facts that the phases coincide and that the introns are positioned near the upstream end of the alignment (typically the most uncertain part) suggest misalignment (i.e., an artifactual explanation [cf., Stoltzfus et al. 1997; Fedorov, Merican, and Gilbert 2002]). Misalignment would result in the same intron appearing in more than one alignment column, and this would lead to further errors in inferences of intron-gain events. Whereas a human expert carrying out a “hands-on” analysis might choose to exclude the region of the alignment with these problematic sites, in this analysis, the same region is excluded automatically on the basis of its low alignment stability scores (fig. 1, histogram).

As another example, one may consider the set of three genes in rows 10 to 12 of figure 1 that form a clade with the structure ((*Entamoeba*, *Babesia*) *Toxoplasma*). This branching order demands multiple events (gain-and-loss or double-gain [see figure 4, panel 4]) to account for the three introns (92-1, 116-1, and 275-1) shared between *Babesia* and *Toxoplasma* to the exclusion of *Entamoeba*. However, this branching order is suspect, because *Toxo-*

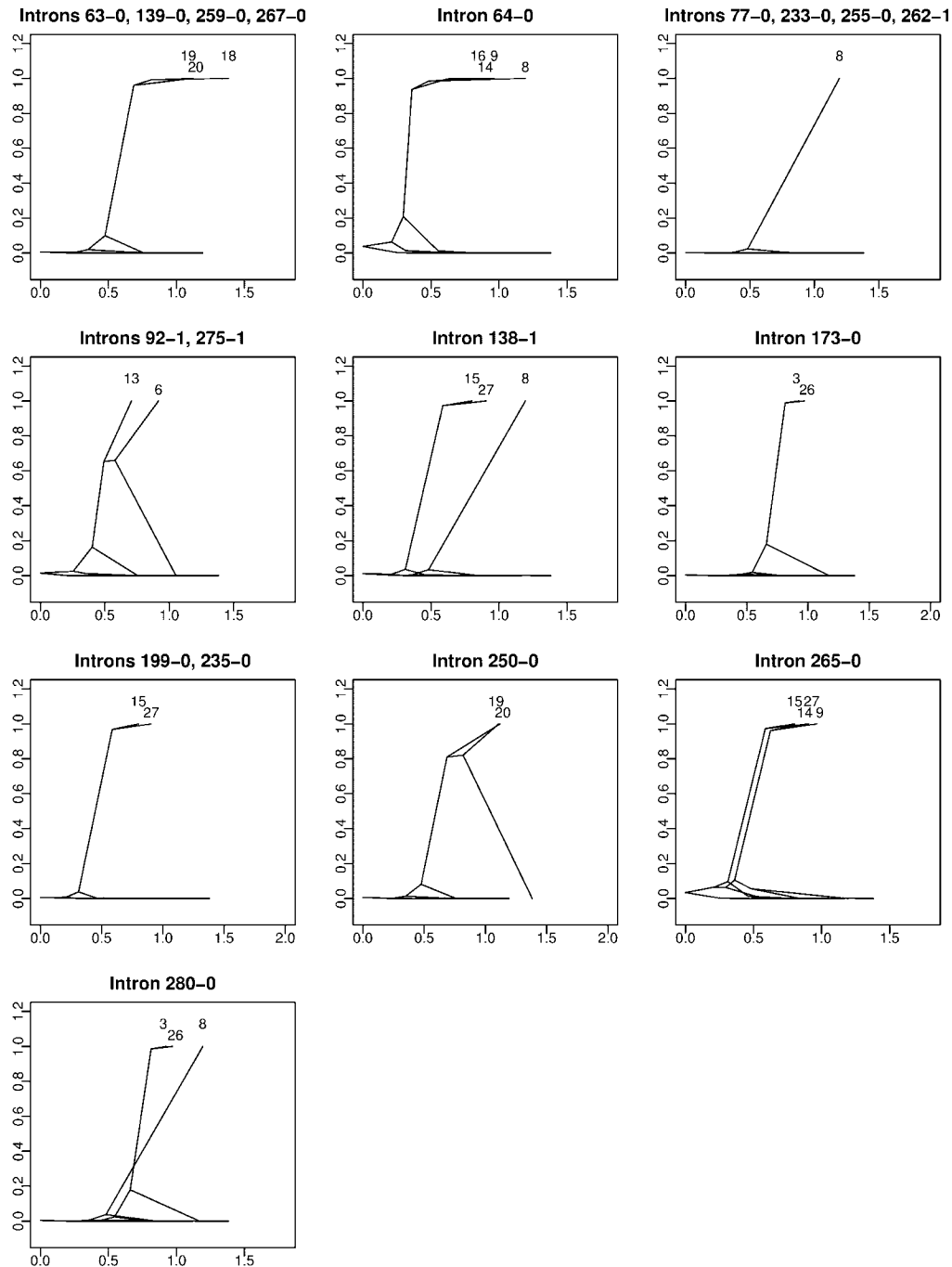


FIG. 4.—Evolution of intron characters in the Mn-Fe SOD family. The graphical convention illustrated in figure 3 is used to represent the evolutionary history of 18 intron sites in the Mn-Fe SOD family that have high alignment reliability scores (SOAP  $\geq 0.95$ ). The probabilities at internal nodes are Bayesian posterior probabilities (see *Materials and Methods*). Because some introns have the same distribution (see figure 1), only 10 plots are needed for 18 introns. Numbers at the tips of the trees represent genes as follows: (3) *Caenorhabditis elegans* AAB53822.1; (5) *Saccharomyces cerevisiae* AAB68939.1; (6) *Babesia bovis* AAB69755.1; (8) *Pneumocystis carinii* AAC24764.1; (9) *Penicillium chrysogenum* AAC36585.1; (13) *Toxoplasma gondii* AAC63943.1; (14) *Neurospora crassa* AAD28503.1; (15) *Arabidopsis thaliana* AAF01529.1; (16) *Schizosaccharomyces pombe* AAF19051.1; (18) *Oryza sativa* BAA92737.1; (19) *Arabidopsis thaliana* BAA97372.1; (20) *Arabidopsis thaliana* BAB11186.1; (26) *Caenorhabditis elegans* CAB02913.1; and (27) *Arabidopsis thaliana* CAB87434.1. Features of interest include the low probability of presence at the root in all cases, the appearance of double-gain events (panels 5, 9, and 10), a clear loss event (panel 8), and an ambiguous case of gain-loss or double-gain (panel 4).

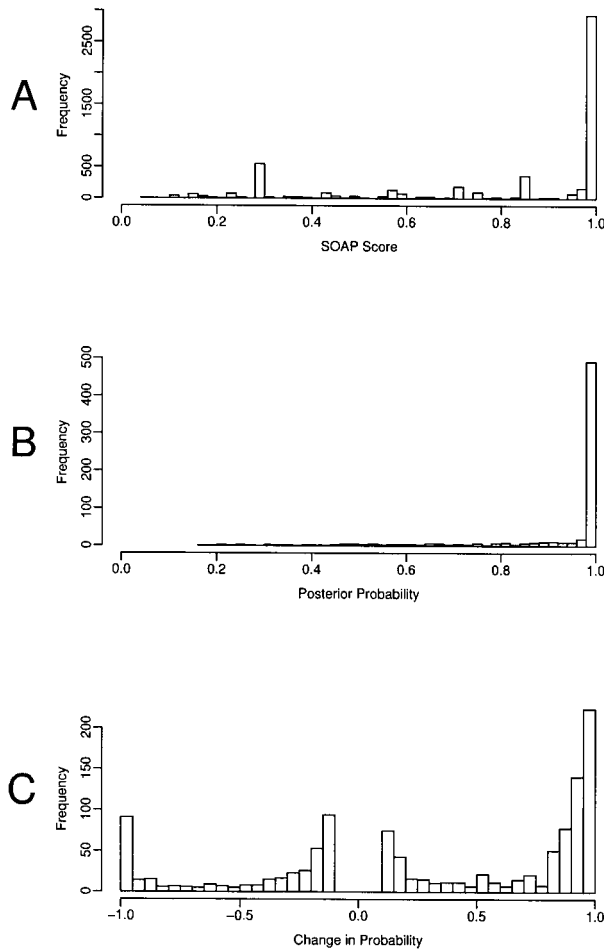


FIG. 5.—Distributions of quality scores used to identify a subset of “most reliable” results for analysis. (A) Distribution of alignment stability scores, which are columnwise SOAP scores (Löytynoja and Milinkovitch 2001), for all alignment columns in all 10 families. The unevenness in the distribution is caused by the coarse discretization of SOAP scores (caused ultimately by a coarsely discretized search of alignment parameter space [see *Materials and Methods*]). In this study, a threshold of 0.95 is used to distinguish higher scores from lower ones. (B) Branch support values, which are Bayesian posterior probabilities from phylogenetic inference with MrBayes (Huelsenbeck and Ronquist 2001), for all branches in all 10 trees. A threshold of 0.99 is used to distinguish higher scores from lower ones. (C) Reliability of inferred events of gain and loss, expressed as the increase in probability of presence of an intron, for all introns and all tree branches in all families, excluding values between  $-0.1$  and  $0.1$  (the majority of which are very close to 0). Values close to  $-1$  indicate loss events (a threshold of  $-0.8$  is used to distinguish stronger from weaker inferences); values close to 1 indicate gain events (a threshold of 0.8 is used to distinguish stronger from weaker inferences); and values close to zero indicate no event.

*plasma* and *Babesia* are apicomplexans and would be expected to branch together to the exclusion of *Entamoeba* and *Trypanosoma* (Fast et al. 2002), in which a single event of intron gain can account for each shared *Toxoplasma-Babesia* intron. A human expert might seek means to rectify the tree or exclude the introns at this site. By comparison, the approach used here automatically excludes events of gain for these introns because the branch support value is not sufficiently high to be included in the “most reliable” subset.

**Table 2**  
Phase Distributions of Extant Introns and Inferred Events of Gain and Loss

	Phase 0	Phase 1	Phase 2	Total
All intron sites	245 (50.2%)	143 (29.3%)	100 (20.5%)	488
All introns	932 (49.9%)	613 (32.8%)	323 (17.3%)	1,868
Inferred losses	43 (49.4%)	32 (36.8%)	12 (13.8%)	87
Inferred gains	173 (51.5%)	95 (28.3%)	68 (20.2%)	336

#### Phases of Introns and Inferred Events of Gain and Loss

Table 2 shows the phase distribution of introns, intron sites, reliable-gain events, and reliable-loss events. For the reliable subset of 336 inferred events of gain, the ratio of counts for phases 0, 1, and 2, namely 173:95:68, is heavily biased towards phase 0. Importantly, this distribution does not differ significantly from the phase distribution of all introns, or of all intron sites, which is about 5:3:2 in either case. Likewise, the phase ratio for reliable events of intron loss, namely 43:32:12, does not differ significantly from that of all intron sites (the  $\chi^2$  values for the  $2 \times 3$  contingency test with 2 degrees of freedom are 0.14, 3.37, 3.07, and 0.99, respectively, for gains by sites, gains by introns, losses by sites, and losses by introns, thus  $P > 0.05$  in every case).

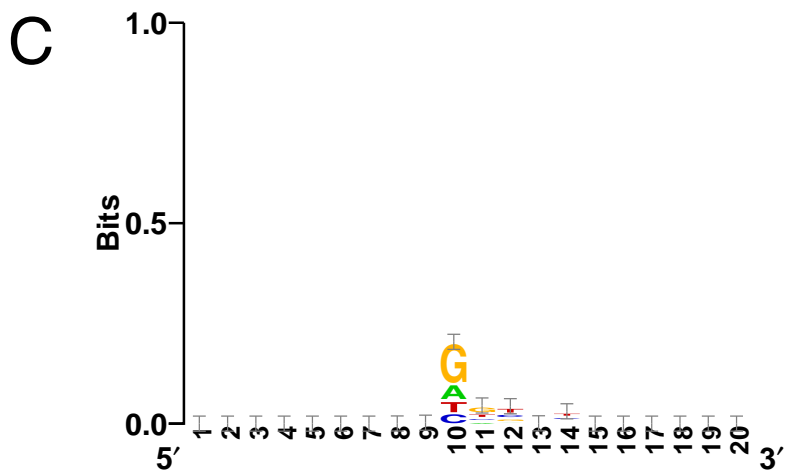
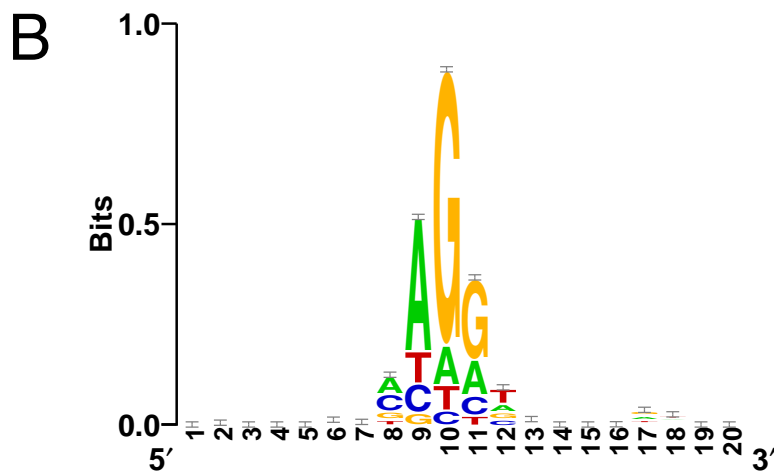
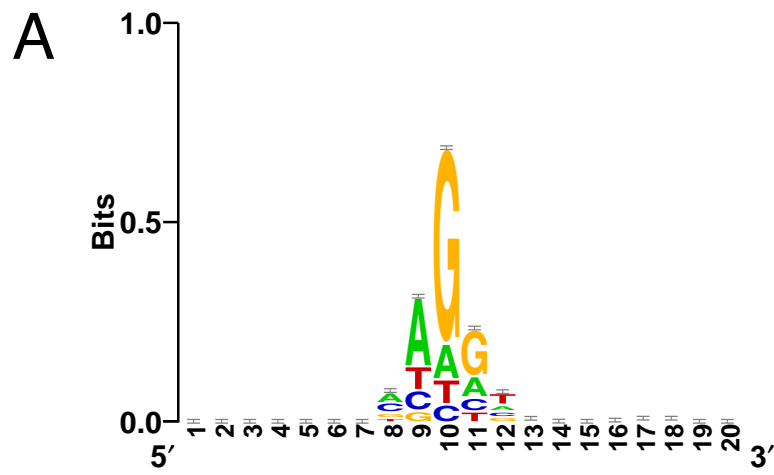
Such generalities conceal significant taxonomic heterogeneity at the kingdom level. Phase frequencies for introns in fungal genes do not differ significantly from a uniform distribution (by site,  $\chi^2 = 2.2$  with 2 df,  $P > 0.05$ ). By contrast, phase frequencies in plant and animal genes are heavily biased toward phase 0 (for phase uniformity among sites,  $\chi^2 = 45$ ,  $P = 2 \times 10^{-10}$  for animals and  $\chi^2 = 43$ ,  $P = 3 \times 10^{-12}$  for plants). Such differences have been known for several years (e.g., Deutsch and Long [1999]).

Remarkably, an analogous pattern applies when inferred events of intron gain are assigned to taxonomic kingdoms, as shown by the results in table 3 (an event is assigned to a given kingdom if and only if all of the genes descendant from the branch on which the event occurs are in the given kingdom). For reliable events of intron gain in fungi, the phase distribution does not differ significantly from uniformity ( $\chi^2 = 0.20$  with 2 df, not significant). For reliable events of intron gain in animals and plants, the phase distribution differs significantly from uniformity (e.g., for animals,  $\chi^2 = 30$  with 2 df,  $P = 3 \times 10^{-7}$ ) and from fungal phase frequencies (e.g., for animals,  $\chi^2 = 11$  for the  $2 \times 3$  contingency test with 2 df,  $P = 4 \times 10^{-3}$ ) but does not differ significantly from the phase distribution of extant introns in the same kingdom (for animals,  $\chi^2 = 1.6$  with 2 df; for plants,  $\chi^2 = 2.2$  with 2 df, not significant). Thus, kingdom-specific patterns of phase-biased intron gain account quantitatively for kingdom-specific patterns of phase bias among extant introns.

#### Nucleotide Preferences of Intron Sites and Intron Gain Sites

For many years, it has been known that the exonic nucleotides flanking introns are nonrandom, with a pattern that tends to favor MAG|GT, where | indicates the location of the intron (Mount 1982; Stephens and Schneider 1992).

Corrected version of Fig. 6 (note order of panels):



**Table 3**  
**Taxonomical Heterogeneity with Respect to Phase**

	Phase 0	Phase 1	Phase 2	Total
Fungal sites	49 (37.1%)	47 (35.6%)	36 (27.3%)	132
Fungal gains	33 (33.7%)	36 (36.7%)	29 (29.6%)	98
Animal sites	112 (55.4%)	50 (24.8%)	40 (19.8%)	202
Animal gains	72 (61.0%)	29 (24.6%)	17 (14.4%)	118
Plant sites	109 (57.1%)	53 (27.7%)	29 (15.2%)	191
Plant gains	64 (64.0%)	18 (18.0%)	18 (18.0%)	100

This pattern is sometimes called the “shadow sequence,” for the metaphorical shadow that the intron casts on the surrounding exonic sequence. The results shown in figure 6 confirm the prior observation that the shadow sequence for fungi is different from that for animals and plants (Deutsch and Long 1999).

The sequences at sites of historical events of intron gain cannot be observed directly. Of course, the sequences flanking extant introns gained in the past will tend to reflect any nucleotide preferences of intron gain, but they also will reflect any bias in post-gain evolution caused by the presence of the intron. Indeed, the shadow sequence might be explained entirely by a postgain convergence of flanking sequences. Therefore, an unbiased method for detecting gain-site preferences must ignore the extant intron-containing genes and, instead, look at their close relatives that have not been exposed to the intron.

To implement this approach systematically, for each of the 336 reliably inferred events of gain, a sequence of 20 nucleotides flanking the site of gain is extracted from an intronless gene drawn from the closest phylogenetic outgroup. Formally, these outgroup sequences are evolutionary relatives of the true gain site, and they will reflect preferences of gain as well as an unknown amount of evolutionary noise resulting from divergence. The pattern shown by these sequences in figure 7A indicates a clear preference for the nucleotides {AC}AG GT at positions  $-3$  to  $+2$ . Although the logos in figure 7 use all nucleotide positions, masking out third codon positions (which tend to diverge at a higher rate, decreasing the signal-to-noise ratio) typically increases the total information content significantly (not shown). The much lower information content for the outgroup sequences in figure 7 relative to the shadow sequences in figure 6 should not be taken to mean that biased intron gain is insufficient to account for the shadow sequence because of the noise included in the outgroup sequence sample. Finally, sequence preferences for intron gain, like phase preferences, show taxonomic heterogeneity (fig. 7B, C).

## Discussion

A systematic evolutionary analysis of data on 10 taxonomically diverse sequence families comprising 677 eukaryotic protein-coding genes reveals that, for the families studied, the process of evolutionary intron gain (1) accounts almost completely for extant introns, (2) is distinctly nonrandom with respect to nucleotide context, and (3) is distinctly nonrandom with respect to reading-frame phase.

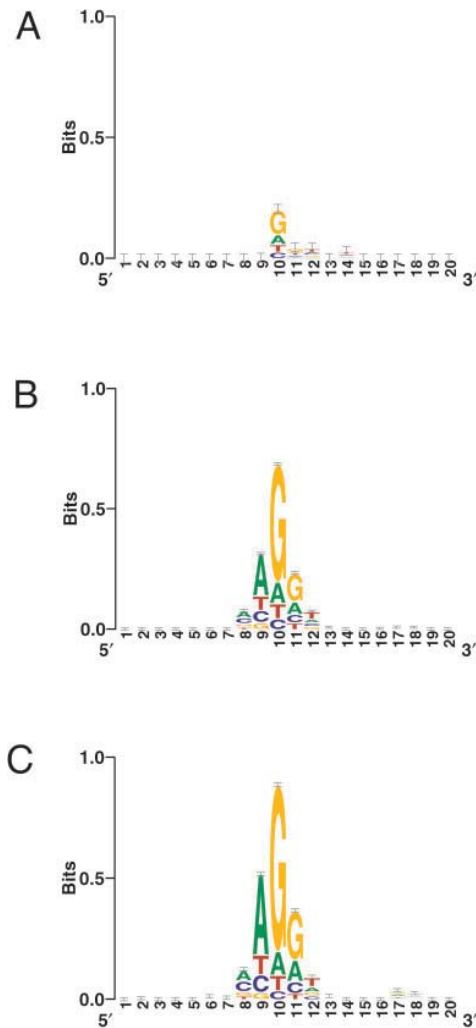


FIG. 6.—Shadow sequences. The three figures show sequence logos (Schneider and Stephens 1990) indicating the pattern at exonic sites flanking introns in (A) all genes used in this study, (B) plant and animal genes, and (C) fungal genes. The positions numbered 1 through 20 along the horizontal axis correspond to positions  $-10$  through  $-1$  upstream of the intron and 1 through 10 downstream of the intron. The vertical axis is in bits. The height of each stack of letters is the information content for that position (in bits), and the relative heights of letters within a stack reflect their frequencies at that position. The data from plant and animal genes are pooled because they do not differ substantially in their pattern, whereas fungal genes are distinctly different.

## The Origin of Introns by a Process of Gain

Because the distribution of introns among members of a gene family is a valuable source of information, the proper interpretation of such data has been a subject of debate. For many years, this debate took the form of a dispute over the relative importance of “differential loss” of introns versus separate gain of introns, or in its most extreme form, a debate between “introns-early” advocates who attributed all differences in intron patterns to differential loss and “sliding” of primordial introns (Liaud, Zhang, and Cerff 1990; Gilbert and Glynias 1993), and “introns-late” advocates who allowed both gain and loss, but denied

“sliding” (Palmer and Logsdon 1991). A fundamental source of ambiguity in this debate has been that any particular distribution of introns in a gene family can be explained in a variety of ways; for example, allowing only loss or gain and loss, loss and sliding, and so on. In the absence of prior restrictions on the types of events that are possible, counting up events of one type or another is inconclusive.

Nevertheless, opinions on the importance of intron gain have shifted in recent years, largely because of a revised assessment of the importance of intron “sliding.” As has long been recognized, any theory of intron evolution relying solely on the inheritance and loss of primordial formative introns would be inadequate because (1) there are just too many introns in too many different positions for them all to be formative introns, and (2) the putative signals of formative processes are sufficiently weak that, at best, only a minority of introns could have been involved directly. Initially, the “sliding” of ancient introns to new positions was a popular means of resolving both issues, but it has become clear subsequently that “sliding” cannot account for observed spatial and phylogenetic patterns of intron diversity (Rzhetsky et al. 1997; Stoltzfus et al. 1997; Fedorov, Merican, and Gilbert 2002). With “sliding” thus discounted, it is now tacitly accepted, even by former advocates of the “introns-early” view (e.g., de Souza et al. [1998]), that most, if not all, extant spliceosomal introns were added to genes during eukaryotic evolution, long after any putative primordial period of gene assembly.

The present study provides a rigorous basis for a stronger set of conclusions. The use of a probabilistic model of character transformation allows the evolutionary history of introns to be inferred from the data, without relying on restrictive assumptions or preconceptions about the antiquity of introns (see Roger [1996], for a precursor of this approach). The result that, for introns as a whole, the probability of presence in a common ancestral gene is roughly 1% indicates that the vast majority of extant introns owe their presence to an evolutionary process of intron gain.

This finding appears to contradict the conclusions of Fedorov, Merican, and Gilbert (2002), who carried out a massive between-kingdom comparison of homologous genes from plants, animals, and fungi, with the result that, depending on the comparison, 4.5% to 14.8% of intron sites found in genes from one kingdom can be found in genes of another kingdom. They interpret this finding to mean that some substantial fraction of extant intron sites, perhaps 14% or more, were occupied in a eukaryotic common ancestor. In fact, the contradiction with the present study is only in terms of conclusions and not in terms of results. For the 10 families used in this analysis (table 1), between-kingdom sharing of intron sites ranges from 4.95% to 17.8% (depending on the comparison), consistent with the values reported by Fedorov, Merican, and Gilbert (2002).

However, one cannot deduce the ancestral occupancy of these shared intron sites without some method for partitioning them into (1) ancestrally occupied sites and (2) sites of separately gained introns. In the present study, this partitioning is accomplished in the context of a probabilistic model of intron evolution, which indicates that most cases of between-kingdom sharing are the result of parallel gain.

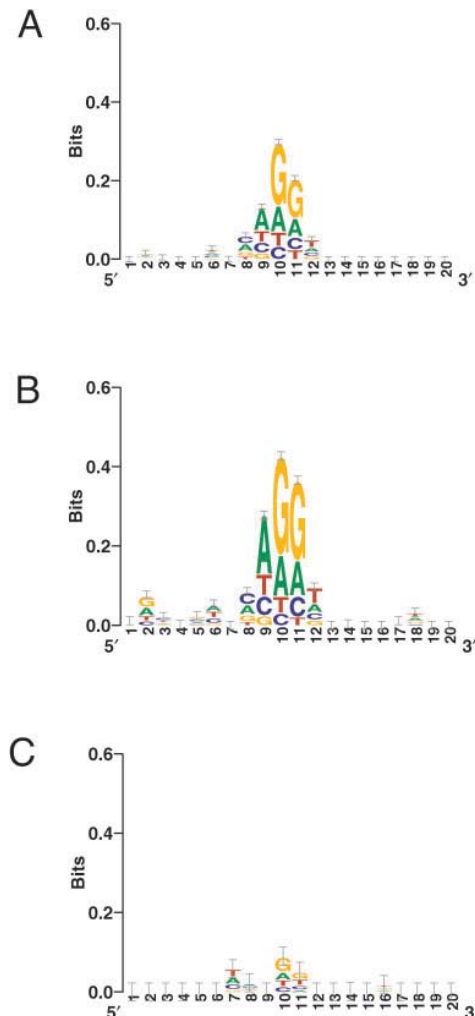


FIG. 7.—Nucleotide preferences for intron gain as reflected in outgroup sequences. The three figures show sequence logos (Schneider and Stephens 1990) of outgroup sample sequences for the “most reliable” intron gain events in (A) all genes, (B) plant and animal genes, and (C) fungal genes. A single outgroup sequence is chosen at random for each inferred gain event. Conventions are as given in the legend to figure 6, but note that the vertical scale (in bits) is smaller. The data from plant and animal genes are pooled in (B) because they do not differ substantially in their pattern, whereas fungal genes (C) are distinctly different.

An example would be intron 138-1 in the Mn-Fe SOD family, which is present in the *Pneumocystis* gene but not in genes from other fungi and is present in two closely related *Arabidopsis* genes but not in related genes from other plants (see figure 1 and figure 4, panel 5; other multiple gain events are shown in panels 9 and 10 of figure 4).

#### Nucleotide Preferences for Intron Gain

The results in figure 7 indicate that intron gain is nonrandom with respect to flanking nucleotides. These results require cautious interpretation because they represent, not the gain-site preference for introns, but the gain-

site preference obscured by an undetermined amount of noise caused by evolutionary divergence. Nevertheless, these results support three conclusions, with implications outlined below. First, although the absolute strengths of gain preferences are subject to interpretation, the most strongly preferred nucleotides from  $-3$  to  $+2$  clearly are MAG GT, where  $M = \{C, A\}$ . Second, these preferred nucleotides show a highly significant match to the nucleotides found most commonly in the “shadow sequence,” namely MAG on the upstream side and GT on the downstream side (e.g., see figure 6 or Stephens and Schneider [1992] and Long et al. [1998]). Third, although some of the indeterminacy of the signal shown in figure 7 is caused by noise, some of it reflects indeterminacy of the gain site itself.

The last point is reasoned as follows. If every intron were gained at the sequence AAG GT, then the varying information content at different positions in figure 7 would reflect, not differences in the strength of intron gain preferences at different flanking sites, but varying contributions of noise at different flanking sites. Yet, varying levels of sequence divergence do not account for major features of the observed pattern. For instance, the strongest signal is the G at  $-1$ , which should have the most noise, because—as a result of the phase 0 preference—the  $-1$  position is the most likely to be a third codon position and, thus, a rapidly diverging position (e.g., for all G-ending codons except ATG and TGG, the nucleotide transition from G to A would be a synonymous change). Thus, given the observation of different degrees of preference at different sites, intron gain sites cannot conform to a single pattern, but must represent a mixture, presumably including not only AAG GT and CAG GT but also other sequences with four or fewer preferred nucleotides.

Such nucleotide preferences for intron gain, whatever their exact nature, must reflect some combination of fitness factors and mutational factors. A hypothesis based only on fitness effects would be that the mutational process of intron gain exhibits no preferences, but mutant alleles with intron insertions face negative selection to the extent that they lack flanking exonic nucleotides that favor splicing, presumably MAG | GT, an idea suggested previously by Hickey, Benkel, and Abukashawa (1989). If so, then the intron gain site would be, operationally, a selective retention site. A hypothesis based solely on mutational factors would be that introns transpose preferentially into sites similar to MAG GT and that allele fixation, although it may systematically bias intron gain with respect to other criteria (e.g., distance to the nearest neighboring intron), does not do so with respect to flanking nucleotides. If so, the intron gain site would be a mutational target site, analogous to the target sites of transposable elements.

Interestingly, if introns transpose by reverse splicing (Eickbush 2000), the selective and mutational factors may overlap considerably. Natural selection would favor whatever flanking nucleotides increase the speed and accuracy of the forward reaction of splicing, but these same nucleotides would tend to increase the speed and accuracy of the reverse reaction necessary for retrotransposition; that is, they would represent target-site preferences via the

mutational reversal of splicing and retention-site preferences via the efficacy of the forward reaction of splicing.

The observed sequence preferences tend to rule out the possibility of widespread gain of introns by the “proto-splice site duplication” model of Dibb (1991). Dibb noted that duplication of a sequence containing xxxAGGTyyy creates a sequence xxxAGGTyyy...xxxAGGTyyy that has most of the *cis*-acting signals for splicing out a GTyyy...xxxAG segment, which would leave a spliced mRNA with the sequence xxxAGGTyyy, as before. However, because the upstream GT and downstream AG nucleotides are so strongly preferred, present in over 99 % of introns (Mount 1996), gain sites resulting from this mechanism would nearly always have the sequence AG GT. Yet, as explained above, a unique gain-site sequence is irreconcilable with the complexity of the pattern in figure 7. Thus, while the duplication mechanism might contribute somewhat to intron gain, it does not appear to be the predominant mechanism of gain.

#### Phase Preferences for Intron Gain

The results presented here indicate that the phase ratio of intron gain events is approximately 5:3:2 (for phases 0, 1, and 2, respectively) for eukaryotes in general. Importantly, this distribution is not significantly different from the phase distribution of extant introns. Thus, the phase bias of evolutionary intron gain is sufficient to account for the phase bias of extant introns. When intron gain events are assigned to taxonomic kingdoms (taxa in this case being given a phyletic definition that includes ancestors), it is apparent that the phase distribution for intron gain in fungi is different from that in animals and plants. In particular, the phase distribution of gain events in fungi is approximately uniform (i.e., one-third of each phase), and the same is true of the phase distribution of extant introns in fungi. This suggests, again, at an even finer level, that phase-biased gain accounts for the phase bias in extant introns.

Various mutational and selective factors could be responsible for the observed phase preferences. In the past, a variety of authors claimed, on the grounds that intron gain is a so-called “random” process, that intron gain could not result in any phase preferences or phase autocorrelation (Smith 1988; Long, Rosenberg, and Gilbert 1995; de Souza, Long, and Gilbert 1996). The legitimacy of such claims is called into question by the work of Long and Rosenberg (2000), who have shown that some possible intron target sites show a nonuniform distribution of phases (because of the phasic nonuniformity in the nucleotide composition of genes) and may show autocorrelations in phase because of regional differences in nucleotide composition. The AG G pattern, in particular, shows a strikingly nonuniform phase distribution. For example,  $\{0.57, 0.26, 0.17\}$  is the expected phase distribution in a dicodon model of human coding regions (Long and Rosenberg 2000). If phase biases arise from nucleotide sequence preferences, then it should be possible to predict phase preferences from a more precise description of the preferences, along with a model of coding regions, and such predictions should be able to account for taxon-specific

phase patterns by invoking taxon-specific nucleotide preferences.

Finally, these results greatly restrict the extent to which the retention of primordial formative introns may be invoked to account for the phase 0 excess (de Souza et al. 1998). The strong phase 0 preference of intron gain and the predominance of intron gain as the explanation for extant introns means that any contribution of ancient phase 0 introns must be quantitatively minor. For example, if the 10 sequence families used here are at all representative of eukaryotic sequence families in general, one may rule out the suggestion that 30% or more of present-day introns represent primordial formative introns (de Souza et al. 1998) on the grounds that if 30% of introns are primordial phase 0 introns and the remaining 70% arose by a process of gain, then given the phase bias of gain (table 2), the expected distribution of intron phases would be {0.66, 0.20, 0.14}, which differs significantly from the observed distribution of phases (by sites,  $\chi^2 = 54$ , with 2 df,  $P = 10^{-12}$ ). Instead, the results on phases are consistent with the results on the probability distribution of presence-at-the-root (fig. 2), both of which indicate that the contribution of putative primordial introns is of negligible importance in accounting for present-day patterns in gene structure.

### Acknowledgments

The authors thank Ari Löytynoja for modifications to the SOAP program, Gavin Crooks for modifications to Weblogo, Eric Nawrocki for tests of data quality, and Eugene Melamud and Peter Yang for Perl code integrated into SPAN. This work was supported by the Center for Advanced Research in Biotechnology and by the National Library of Medicine (NIH Grant R01-LM007218–01A1 to A.S.). The identification of specific commercial software products in this paper is for the purpose of specifying a protocol and does not imply a recommendation or endorsement by the National Institute of Standards and Technology.

### Literature Cited

- Barton, G. J. 2001. Creation and analysis of protein multiple sequence alignments. Pp. 215–232 in A. D. Baxevanis and B. F. F. Ouellette, eds. *Bioinformatics: a practical guide to the analysis of genes and proteins*. John Wiley & Sons, New York.
- Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. Sonnhammer. 2000. The Pfam protein families database. *Nucleic Acids Res.* **28**:263–266.
- Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2003. GenBank. *Nucleic Acids Res.* **31**:23–27.
- Bhattacharya, D., F. Lutzoni, V. Reeb, D. Simon, J. Nason, and F. Fernandez. 2000. Widespread occurrence of spliceosomal introns in the rDNA genes of *Ascomycetes*. *Mol. Biol. Evol.* **17**:1971–1984.
- Briffeuil, P., G. Baudoux, C. Lambert, X. De Bolle, C. Vinals, E. Feytmans, and E. Depiereux. 1998. Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions. *Bioinformatics* **14**:357–366.
- Cavalier-Smith, T. 1991. Intron phylogeny: a new hypothesis. *Trends Genet.* **7**:145–148.
- Cho, G., and R. F. Doolittle. 1997. Intron distribution in ancient paralogs supports random insertion and not random loss. *J. Mol. Evol.* **44**:573–584.
- Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. 1999. *Probabilistic networks and expert systems*. Springer, New York.
- Craik, C. S., W. J. Rutter, and R. Fletterick. 1983. Splice junctions: association with variation in protein structure. *Science* **220**:1125–1129.
- Darnell, J. E., and W. F. Doolittle. 1986. Speculations on the early course of evolution. *Proc. Natl. Acad. Sci. USA* **83**:1271–1275.
- de Souza, S., M. Long, and W. Gilbert. 1996. Introns and gene evolution. *Genes Cells* **1**:493–505.
- de Souza, S. J., M. Long, R. J. Klein, S. Roy, S. Lin, and W. Gilbert. 1998. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. USA* **95**:5094–5099.
- de Souza, S. J., M. Long, L. Schoenbach, S. W. Roy, and W. Gilbert. 1996. Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl. Acad. Sci. USA* **93**:14632–14636.
- Descartes, A., and T. Bunce. 2000. *Programming the Perl DBI*. O'Reilly & Assoc., Cambridge, MA.
- Deutsch, M., and M. Long. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27**:3219–3228.
- Dibb, N. J. 1991. Proto-splice site model of intron origin. *J. Theor. Biol.* **151**:405–416.
- . 1993. Why do genes have introns? *FEBS Lett.* **325**:135–139.
- Dibb, N. J., and A. J. Newman. 1989. Evidence that introns arose at proto-splice sites. *EMBO J.* **8**:2015–2021.
- Doolittle, R. F. 1991. Counting and discounting the universe of exons. *Science* **253**:677–679.
- . 1978. Genes in pieces: were they ever together? *Nature* **272**:581–582.
- Dorit, R. L., and W. Gilbert. 1991. The limited universe of exons. *Curr. Opin. Genet. Dev.* **1**:464–469.
- Eickbush, T. H. 2000. Molecular biology: introns gain ground. *Nature* **404**:940–941, 943.
- Fast, N. M., L. Xue, S. Bingham, and P. J. Keeling. 2002. Re-examining alveolate evolution using multiple protein molecular phylogenies. *J. Eukaryot. Microbiol.* **49**:30–37.
- 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.
- Fedorov, A., A. F. Merican, and W. Gilbert. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl. Acad. Sci. USA* **99**:16128–16133.
- Fedorov, A., G. Suboch, M. Bujakov, and L. Federova. 1992. Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res.* **120**:2553–2557.
- Fichant, G. A. 1992. Constraints acting on the exon positions of the splice site sequences and local amino acid composition of the protein. *Hum. Mol. Genet.* **1**:259–267.
- Frugoli, J. A., M. A. McPeck, T. L. Thomas, and C. R. McClung. 1998. Intron loss and gain during evolution of the catalase gene family in angiosperms. *Genetics* **149**:355–365.
- Funke, R. P., J. L. Kovar, J. M. Logsdon, Jr., J. C. Corrette-Bennett, D. R. Straus, and D. P. Weeks. 1999. Nucleus-encoded, plastid-targeted acetolactate synthase genes in two closely related chlorophytes, *Chlamydomonas reihardtii* and *Volvox carteri*: phylogenetic origins and recent insertion of introns. *Mol. Gen. Genet.* **262**:12–21.

- Gelman, A., and D. B. Rubin. 1996. Markov chain Monte Carlo methods in biostatistics. *Stat. Methods Med. Res.* **5**:339–355.
- Gilbert, W., and M. Glynias. 1993. On the ancient nature of introns. *Gene* **135**:137–144.
- Gilbert, W., M. Marchionni, and G. McKnight. 1986. On the antiquity of introns. *Cell* **46**:151–153.
- Harvey, P. H., and M. D. Pagel. 1991. *The comparative method in evolutionary biology*. Oxford University Press, Oxford, UK.
- Hickey, D. A., B. F. Benkel, and S. M. Abukashawa. 1989. A general model for the evolution of nuclear pre-mRNA introns. *J. Theor. Biol.* **137**:41–53.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754–755.
- Ihaka, R., and R. Gentleman. 1996. R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**:299–314.
- Jensen, F. V. 1996. *Introduction to Bayesian networks*. Springer, New York.
- 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. Biol.* **221**:175–191.
- Liaud, M. F., D. X. Zhang, and R. Cerff. 1990. Differential intron loss and endosymbiotic transfer of chloroplast glyceraldehyde-3-phosphate dehydrogenase genes to the nucleus. *Proc. Natl. Acad. Sci. USA* **87**:8918–8922.
- Logsdon, J. M. Jr., A. Stoltzfus, and W. F. Doolittle. 1998. Molecular evolution: recent cases of spliceosomal intron gain? *Curr. Biol.* **8**:R560–R563.
- Long, M., S. J. de Souza, and W. Gilbert. 1995. Evolution of the intron-exon structure of eukaryotic genes. *Curr. Opin. Genet. Dev.* **5**:774–778.
- Long, M., S. J. de Souza, C. Rosenberg, and W. Gilbert. 1998. Relationship between “proto-splice sites” and intron phases: evidence from dicodon analysis. *Proc. Natl. Acad. Sci. USA* **95**:219–223.
- Long, M., and C. Rosenberg. 2000. Testing the “proto-splice sites” model of intron origin: evidence from analysis of intron phase correlations. *Mol. Biol. Evol.* **17**:1789–1796.
- 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.
- Löytynoja, A., and M. C. Milinkovitch. 2001. SOAP, cleaning multiple alignments from unstable blocks. *Bioinformatics* **17**:573–574.
- Maddison, D. R., D. L. Swofford, and W. P. Maddison. 1997. NEXUS: an extendible file format for systematic information. *Syst. Biol.* **46**:590–621.
- Mount, S. M. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* **10**:459–472.
- . 1996. AT-AC introns: an ATtACK on dogma [comment]. *Science* **271**:1690–1692.
- Newman, A. J. 1994. Pre-mRNA splicing. *Curr. Opin. Genet. Dev.* **4**:298–304.
- Palmer, J. D., and J. M. Logsdon, Jr. 1991. The recent origins of introns. *Curr. Opin. Genet. Dev.* **1**:470–477.
- Patthy, L. 1991. Modular exchange principles in proteins. *Curr. Opin. Struct. Biol.* **1**:351–361.
- Plummer, M., N. Best, K. Cowles, and K. Vines. 2002. CODA: Output analysis and diagnostics for MCMC. Martyn Plummer, International Agency for Research on Cancer, Lyon, France.
- Roger, A. J. 1996. *Studies on the phylogeny and gene structure of early-branching eukaryotes*. Doctoral thesis, Department of Biochemistry, Dalhousie University, Halifax, Canada.
- Rogozin, I. B., J. Lyons-Weiler, and E. V. Koonin. 2000. Intron sliding in conserved gene families. *Trends Genet.* **16**:430–432.
- Rzhetsky, A., F. J. Ayala, L. C. Hsu, C. Chang, and A. Yoshida. 1997. Exon/intron structure of aldehyde dehydrogenase genes supports the “introns-late” theory. *Proc. Natl. Acad. Sci. USA* **94**:6820–6825.
- Schisler, N. J., and J. D. Palmer. 2000. The IDB and IEDB: intron sequence and evolution databases. *Nucleic Acids Res.* **28**:181–184.
- Schneider, T. D., and R. M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**:6097–6100.
- Smith, M. W. 1988. Structure of vertebrate genes: a statistical analysis implicating selection. *J. Mol. Evol.* **27**:45–55.
- Spiegelhalter, D. J., A. P. Dawid, S. L. Lauritzen, and R. G. Cowell. 1993. Bayesian analysis in expert systems. *Stat. Sci.* **8**:219–247.
- Spiegelhalter, D. J., A. Thomas, N. G. Best, and W. R. Gilks. 1996. BUGS: Bayesian inference using Gibbs sampling. Version 0.5, MRC Biostatistics Unit, Cambridge, UK.
- Stephens, R. M., and T. D. Schneider. 1992. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* **228**:1124–1136.
- Stoltzfus, A., J. M. Logsdon, Jr., J. D. Palmer, and W. F. Doolittle. 1997. Intron “sliding” and the diversity of intron positions. *Proc. Natl. Acad. Sci. USA* **94**:10739–10744.
- Stoltzfus, A., D. F. Spencer, M. Zuker, J. M. Logsdon, Jr., and W. F. Doolittle. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science* **265**:202–207.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- Thompson, J. D., F. Plewniak, and O. Poch. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* **27**:2682–2690.
- Torki, M., P. Mandaron, R. Mache, and D. Falconet. 2000. Characterization of a ubiquitous expressed gene family encoding polygalacturonase in *Arabidopsis thaliana*. *Gene* **242**:427–436.
- Wheeler, D. L., C. Chappay, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp. 2000. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **28**:10–14.

Edward Holmes, Associate Editor

Accepted February 16, 2004