Nucleic Acids Research, 2005, Vol. 33, No. 6 ***1741–1748***
doi:10.1093/nar/gki316

Conservation versus parallel gains in intron evolution

Alexander V. Sverdlov, Igor B. Rogozin, Vladimir N. Babenko and Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bldg 38A, Bethesda, MD 20894, USA

Received January 31, 2005; Revised and Accepted March 2, 2005

ABSTRACT

Orthologous genes from distant eukaryotic species, e.g. animals and plants, share up to 25–30% intron positions. However, the relative contributions of evolutionary conservation and parallel gain of new introns into this pattern remain unknown. Here, the extent of independent insertion of introns in the same sites (parallel gain) in orthologous genes from phylogenetically distant eukaryotes is assessed within the framework of the protosplice site model. It is shown that protosplice sites are no more conserved during evolution of eukaryotic gene sequences than random sites. Simulation of intron insertion into protosplice sites with the observed protosplice site frequencies and intron densities shows that parallel gain can account but for a small fraction (5–10%) of shared intron positions in distantly related species. Thus, the presence of numerous introns in the same positions in orthologous genes from distant eukaryotes, such as animals, fungi and plants, appears to reflect mostly bona fide evolutionary conservation.

INTRODUCTION

Eukaryotic protein-coding genes typically are interrupted by multiple introns, which are excised at the donor and acceptor splice sites in a complex splicing reaction mediated by the spliceosome (1). The origin of spliceosomal introns remains a mystery, and the dynamics of their evolution is poorly understood (2,3). However, comparative genomics has the potential of opening a window on the deep past, including, perhaps, the exon–intron structure of protein-coding genes at the earliest stages of eukaryotic evolution. After multiple, complete sequences of eukaryotic genomes became available, comparative analyses revealed numerous introns that occupy the same position in orthologous genes from distant species (4,5). In particular, orthologous genes from humans and the green plant *Arabidopsis thaliana* share ~25% intron positions. Moreover, even genes from protists, which are thought to have diverged from common ancestors with multicellular forms early in eukaryotic evolution, such as those of the malaria plasmodium, share many introns with orthologous genes of animals and plants (4,5).

The straightforward interpretation of these observations is that the shared introns were inherited from the common ancestor of the respective species whereas lineage-specific introns were inserted into genes at later stages of evolution (4,5). Under this premise, parsimonious reconstructions indicate that, even in early eukaryotes, protein-coding genes already had a fairly high intron density, comparable to that in modern plant and animal genes. However, the inference that shared intron positions reflect evolutionary conservation is complicated by the fact that intron insertion does not seem to be a strictly random process. A simple form of such nonrandomness was taken into account in the parsimonious reconstruction of intron evolution by Monte Carlo simulation with intron insertion allowed only in a fraction of a gene's sequence (e.g. 10%). The results suggested a relatively small contribution of independent insertion in the same position in different lineage (parallel gain) to intron evolution (4,5).

Nevertheless, a greater role of parallel gain due to nonrandom intron insertion cannot be ruled out. The short sequence stretches flanking introns have significantly biased nucleotide frequencies which led to the notion of protosplice sites, the target sites for intron insertion (6,7). However, it remained unclear whether the consensus nucleotides flanking the splice junctions were remnants of the original protosplice sites or evolved convergently after intron insertion. The existence of protosplice sites was addressed directly by examining the context of introns inserted within codons, which encode amino acids conserved in all eukaryotes and, accordingly, are not subject to selection for splicing efficiency (8). It has been shown that introns either predominantly insert into specific protosplice sites, which have the consensus sequence (A/C)AG|Gt (Table 1), or insert randomly but are preferentially fixed at such sites (8).

At least two cases of apparent parallel gain of introns in orthologous genes from plants and animals have been reported (9,10). Moreover, recent probabilistic modeling of intron evolution suggested that many, if not most, introns shared by phylogenetically distant species were likely to originate

*To whom correspondence should be addressed. Tel: +1 301 435 5913; Fax: +1 301 435 7794; Email: koonin@ncbi.nlm.nih.gov

Table 1. Frequency of nucleotides in the reconstructed protosplice site (8)

Position	−3	−2	−1	+1	+2
Consensus sequence	M(A/C)	A	G	G	T
A	0.39	0.62	0.08	0.21	0.24
T	0.04	0.21	0.08	0.15	0.43
G	0.16	0.07	0.76	0.58	0.19
C	0.41	0.10	0.08	0.06	0.14

by parallel gain (11). The implications of this conclusion for our understanding of evolution of eukaryotic gene structure are substantial: it follows that intron distribution in extant organisms is largely determined by relatively recent insertions and cannot be used to infer exon–intron structure of ancestral genes.

Therefore, we decided to re-examine the problem of parallel intron gain, taking into account the accumulating information on protosplice sites. Here, we show that there is no preferential conservation of protosplice sites in eukaryotic genes and describe a simulation analysis of intron insertion process under the protosplice site model. The results show that parallel gain of introns is a rare event, and, accordingly, most of the introns shared by distantly related species (~90–95%) reflect long-term evolutionary conservation.

MATERIALS AND METHODS

The data set used for this analysis consisted of 684 clusters of eukaryotic orthologous groups of proteins (KOGs) including a single representative from each of eight eukaryotic species with completely sequenced genomes: *Anopheles gambiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Plasmodium falciparum*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (5). Protein sequences in each KOG were aligned using the MAP program (12). Intron positions were extracted from the feature tables of complete genome annotations in the GenBank database. Intron-flanking sequences ('shadow' sites), i.e. sequences surrounding introns and presumably representing remnants of the protosplice sites into which introns are thought to insert were extracted from the nucleotide sequences of the corresponding exons.

Functional sites in nucleotide sequences can be usefully abstracted into weight matrices (13–15) of the form $W = |\ln w(b,j))|$, where $w(b,j)$ is the frequency of the nucleotide b in position j. Given a weight matrix, the matching score $S(b_1, \ldots, b_L)$ of a sequence $b_1, \ldots, b_L$ is

$$S(b_1, \ldots, b_L) = \sum_{J=1}^{L} \ln w(b_j, j)$$

The weight matrix can be used to recognize the corresponding functional sites by defining a threshold matching score (cut-off) value. In particular, the weight matrix of the protosplice sites (Table 1) can be used for recognition of 'shadow' sequences. For the purposes of this work, we compared the protosplice site weight matrix to the set of shadow sites in coding regions with a series of stringency threshold values such that between 20% (the most stringent threshold) and 95% (the most liberal threshold) of the shadow sites had a weight greater than the threshold (values <20% were not used due to the small number of predicted protosplice sites, fewer than the actual number of introns, data not shown). The threshold values were calculated separately for each species.

In order to assess the likelihood of independent intron insertions in the same position, we simulated the process of intron insertion into genes by randomly scattering introns over the predicted protosplice sites. To incorporate the non-uniformity of intron distribution along gene sequences (16–20) into the model, introns were scattered separately for the 5′-terminal and 3′-terminal halves of each gene. The number and phase distribution (phase 0 introns are located between codons, and phase 1 and 2 introns are located after the first and second positions of codons, respectively) of introns used in the simulation were the same as the actually observed number and distribution in each respective gene, and the probability of intron insertion was made proportional to the weight of a protosplice site. For a predicted protosplice j, this probability is

$$P(j) = S_j(b_1, \ldots, b_L) / \sum_{k=1}^{M} S_k(b_1, \ldots, b_L)$$

where M is the total number of predicted protosplice sites in the respective gene. In cases when an intron could not be inserted into any protosplice site because all sites were already filled, the entire KOG was discarded from the simulation. This model is an extension of the statistical model originally proposed by Stoltzfus in the first theoretical analysis of intron conservation (21).

RESULTS

Absence of preferential conservation of protosplice sites

In order to identify protosplice sites, the coding sequences of eukaryotic genes were compared to the protosplice site weight matrix (Table 1) (8). In modern intron-flanking sequences, the protosplice sites, which have the consensus (A/C)AG|Gt (Table 1), survive in the form of 'shadow' sites with similar but not identical nucleotide frequencies (8). These shadow sites were used to define a series of stringency threshold values for protosplice site identification; the threshold was defined as the percentage of shadow sites recognized with a given weight (see Materials and Methods). This procedure produced the distribution of protosplice sites in the same position of orthologous genes in one, two or more species (Figure 1). The frequency of predicted protosplice sites in different species varied from ~1 protosplice site per 100 bases to ~70 protosplice sites per 100 bases depending on the threshold value (Table 2). Predictably, the stringently defined protosplice sites are only rarely 'conserved' but, with the relaxation of the threshold, occurrence of the sites in more than one species becomes common. However, the distribution of protosplice sites did not significantly differ from the distributions of random pentanucleotides which were derived by shuffling positions of the protosplice site weight matrix: the frequencies of protosplice sites co-occurrence in different species were well within the narrow range of frequencies of randomly generated sites (Figure 2). This observation indicates that there is no specific conservation of protosplice sites in eukaryotic genes.

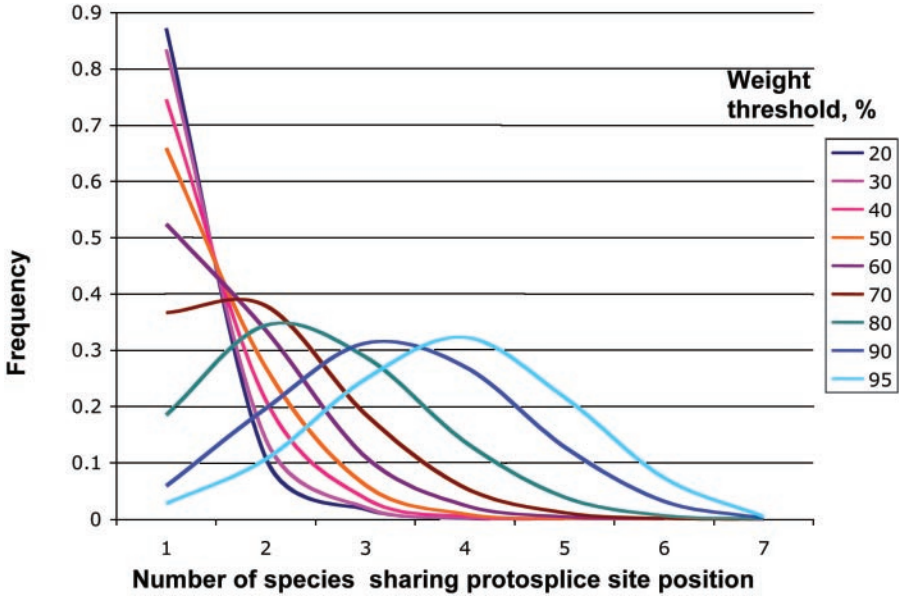

Figure 1. Frequency of protosplice site occurrence in the same positions of orthologous genes depending on the threshold.

Table 2. Number of detected protosplice sites per 100 bases depending on the threshold[a]

	20	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*95*
Ag	0.9	2.0	4.0	7.3	12.4	20.2	35.1	57.1	72.0
At	0.9	2.5	5.2	8.3	13.7	22.2	36.3	59.1	75.9
Ce	0.6	1.8	3.9	7.0	11.3	19.8	33.7	55.1	69.9
Dm	1.0	2.3	4.2	7.5	12.4	19.5	33.0	54.0	70.1
Hs	1.3	2.7	5.4	8.6	13.6	21.0	33.7	54.0	70.2
Pf	0.6	1.7	3.8	5.7	10.6	21.5	39.0	66.5	83.0
Sc	0.9	2.0	4.1	7.0	11.5	19.9	35.3	59.1	77.3
Sp	0.7	2.0	3.9	6.5	11.1	20.1	35.0	59.0	75.8

[a]The weight threshold is defined as the percentage of shadow sites recognized (see Materials and Methods). Abbreviations: Ag, A.*gambiae*; At, A.*thaliana*; Ce, C.*elegans*; Dm, *D.melanogaster*; Hs, *H.sapiens*; Pf, *P.falciparum*; Sc, *S.cerevisiae*; Sp, *S.pombe*.

It is well known that introns show an non-uniform phase distribution, with the ratios of, approximately, 5:3:2 between phases 0, 1 and 2 (5,11,22,23). Notably, these distributions are very different from the nearly uniform phase distribution predicted by the protosplice model (24) (Table 3). Therefore, as discussed previously, the observed excess of phase 0 introns is probably due to the preferential retention of these introns by natural selection (25).

Modeling independent intron gains

The observation that protosplice sites are not specifically conserved does not rule out the possibility of substantial parallel gain of introns. Indeed, with the relaxed protosplice site definition, such as the 90% threshold, there were many coinciding sites in different species (Figure 1), providing for the possibility of independent intron insertion in the same position. In order to assess the likelihood of such events, we simulated the process of intron insertion into genes by randomly scattering introns over the predicted protosplice sites (see Materials and Methods). This simulation, run for three pairs of distantly related, intron-rich genomes, those of *H.sapiens*, *S.pombe* and *A.thaliana*, yielded low frequencies of predicted independent intron gains (Figure 3), suggesting that almost all introns shared by each pair of species are inherited from their respective common ancestors. To rule out bias due to potential alignment artifacts, we performed the same simulations separately for unequivocally aligned, highly conserved regions of the alignments (5), with results nearly identical to those obtained with unfiltered alignments (Figure 3). The threshold value used to select protosplice sites had a negligible effect on the number of parallel gains; although more noticeable in the analysis limited to the most conserved regions, the differences between thresholds were minor compared to the difference between the simulation results and the actual counts of shared introns (Figure 3). In most of the simulations, the number of introns independently inserted in the same protosplice site in different species was <5–10% of the observed number of shared introns; accordingly, the positions of the rest of the shared introns (>90%) appear to have been conserved for more than a billion years which separate, e.g. animals and plants from their common ancestor (Figure 3 and Supplementary Table 1).

When the same simulations were run for all genomes together, even with the most stringent threshold used (20% of the shadow sites recognized), the number of intron positions that were 'conserved' in two species was several times lower than the observed number of shared introns, and the number of introns 'conserved' in three or more species was negligible (Figure 4 and Supplementary Table 2). It should be noticed that, in this bulk analysis of eight species, distantly related genomes were mixed with more closely related ones, such as fruit fly and mosquito. Closely related species have a greater chance of independent intron gains in the same position both in the simulations and, possibly, during evolution, due to the high similarity between orthologous protein sequences. Therefore, this type of analysis allows only a rough estimate of the fraction of independent intron gains (~10–20%) in the multiple

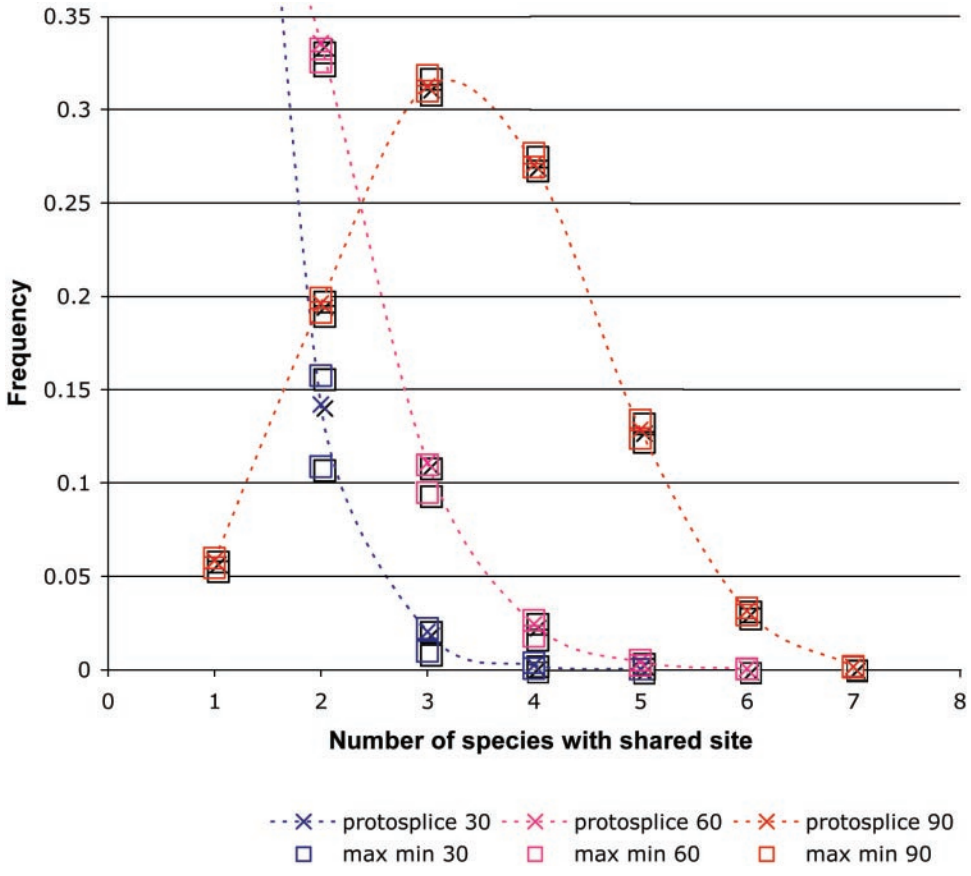

Figure 2. Frequency of protosplice site and random site occurrence in the same positions of orthologous genes for three threshold values. For each threshold, the range of frequencies obtained with 10 random sites is shown. Note that, in each case, the protosplice site frequency falls within the range of random site frequencies.

Table 3. Number of detected protosplice sites in different phases per 100 bases depending on the threshold[a]

	Phases for threshold = 30			Phases for threshold = 60			Phases for threshold = 90		
	0	1	2	0	1	2	0	1	2
Ag	0.66	0.68	0.74	4.0	4.1	4.2	18.9	19.0	19.2
At	0.85	0.81	0.83	4.6	4.5	4.6	19.8	19.6	19.8
Ce	0.6	0.59	0.60	3.8	3.8	3.8	18.4	18.4	18.4
Dm	0.72	0.74	0.79	4.0	4.1	4.3	17.9	18.0	18.1
Hs	0.92	0.88	0.93	4.5	4.5	4.6	17.9	18.1	17.5
Pf	0.63	0.51	0.53	3.9	3.3	3.4	22.8	21.7	22.0
Sc	0.70	0.56	0.77	4.1	3.3	4.1	19.8	18.8	20.6
Sp	0.65	0.61	0.69	3.7	3.6	3.8	19.5	19.2	20.0

[a]The weight threshold is defined as the percentage of shadow sites recognized (see main text). Species abbreviations: Ag, *A.gambiae*; At, *A.thaliana*; Ce, *C.elegans*; Dm, *D.melanogaster*; Hs, *H.sapiens*; Pf, *P.falciparum; Sc, S.cerevisiae*; Sp, *S.pombe*.

alignments of eight species; the above caveat suggests this is likely to be the upper bound of independent intron gain.

Non-local factors that potentially affect intron insertion

A formal possibility exists that some other sequence features distinct from the local nucleotide context and not accounted for in the simulations described here affect the probability of intron insertion in the protosplice sites. In the previous work, we effectively tested the influence of long-range factors on the frequency of independent intron gains by analyzing frequencies of introns separated by a short, fixed distance (1–5 nt) in two or more species (5). No excess of such introns was detected, which makes substantial long-range effects acting over long evolutionary spans unlikely (see Discussion).

Nevertheless, we specifically examined the potential effect of the observed distribution of exon lengths on the extent of independent intron gain. To this end, the predicted frequency of independent gains yielded by simulations for all genomes described above was plotted against the similarity (Spearman correlation coefficient) between the simulated and observed distributions of exon lengths. No significant correlation was found (Figure 5 and Supplementary Figure 1) suggesting that, at least within the framework of the present model, the distribution of exon length did not substantially affect the frequency of independent intron gains.

DISCUSSION

The simulation analysis based on the protosplice model described here shows that independent intron insertion in the same site in orthologous genes is rare, at least in distantly related genomes. Accordingly, the great majority (>90%) of introns shared by such distant orthologous genes, e.g. those

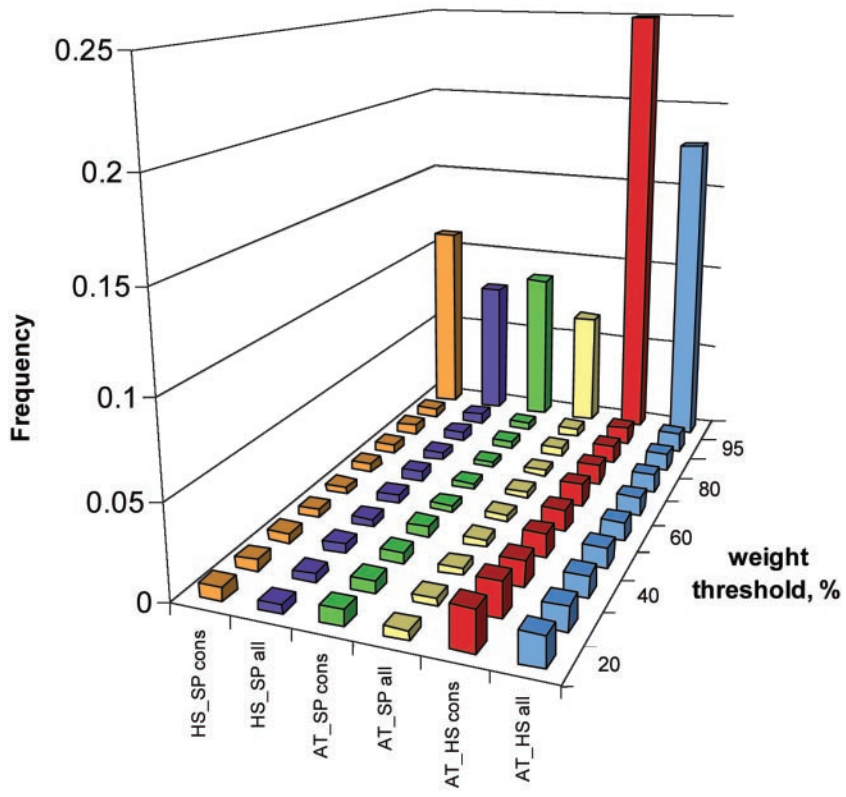

Figure 3. Frequency of independent insertions of introns into the same protosplice sites in orthologous genes from *H.sapiens*, *S.pombe* and *A.thaliana* obtained in simulations compared with the observed frequency of shared intron positions (the last row). The results are shown for a series of increasingly stringent weight thresholds used for the identification of protosplice sites. For each threshold, the mean of 1000 simulations is shown. 'all' indicates that complete alignments of 684 orthologous genes were analyzed, 'cons' stands for the analyses of unambiguously aligned, conserved regions from the same genes (5). Species abbreviations: AT, *A.thaliana*; HS, *H.sapiens*; and SP, *S.pombe*.

from animals and plants, seem to be evolutionarily conserved, i.e. form an uninterrupted lineage of vertical descent from the common ancestor of these organisms. These observations are consistent with the results of an earlier analysis of intron distribution in 20 ancient paralogous families which appear to have accumulated introns independently (26). In this study, Cho and Doolittle detected only two possible parallel gains among 239 analyzed intron positions (~1%) (26). This value, which was obtained with an approach completely different from the one employed in the present work, is comparable to the frequency of independent gains inferred from the simulation results discussed here (5–10%) (Figures 3 and 4, Supplementary Tables 1 and 2); the slightly greater proportion of parallel gains in the present analysis could be due to a higher intron density in our data set (one intron per 68 bases as compared to one intron per 132 bases in the work of Cho and Doolittle). The analysis of Cho and Doolittle was designed as a test of the introns-early hypothesis (27). In the process, however, not only was one of the predictions of this hypothesis (intron sliding, the relocation of exon–intron boundary over a short distance) refuted but, in addition, evidence was obtained against frequent, independent parallel gains of introns, a staple of an extreme intron-late view (10,11).

In a recent probabilistic analysis of intron evolution in 10 large families of eukaryotic genes, Qiu and co-workers arrived to the conclusion that the majority of shared introns were gained independently rather than inherited from the last common ancestor of the respective genes (11). The difficulty with this type of analysis is that alignments of numerous sequences inevitably have a high density of intron positions which may result in artificial high frequency of independent gain in a model. Qiu and co-workers attempted to resolve this issue using a Bayesian model that allowed ancestral inheritance of introns, gain of introns and loss of introns (intron gains and losses were assumed to be completely reversible). However, this model used the unrealistic assumption that the sites actually occupied by an intron in at least one family member comprised the total set of protosplice sites in which multiple independent gains of introns could occur without restriction. It seems likely that this approach led to a significant over-estimate of parallel gains.

Together with the previously reported analysis of protosplice sites (8), the results presented here seem to clarify the mode of intron gain and loss during evolution of eukaryotes. Although it has been shown that introns are, indeed, inserted into protosplice sites, these sites are no more conserved than any random sequence of the same length. Analysis of gene alignments revealed cases of extremely high density of introns, e.g. one intron per 9.7 bases in small G proteins (17) which means that >10% of nucleotide positions in a gene were accessible for intron insertion, at least for some period of time during evolution. As the number of sequenced members of gene families increases, this bound is expected to move even further upward. Taken together, these observations argue

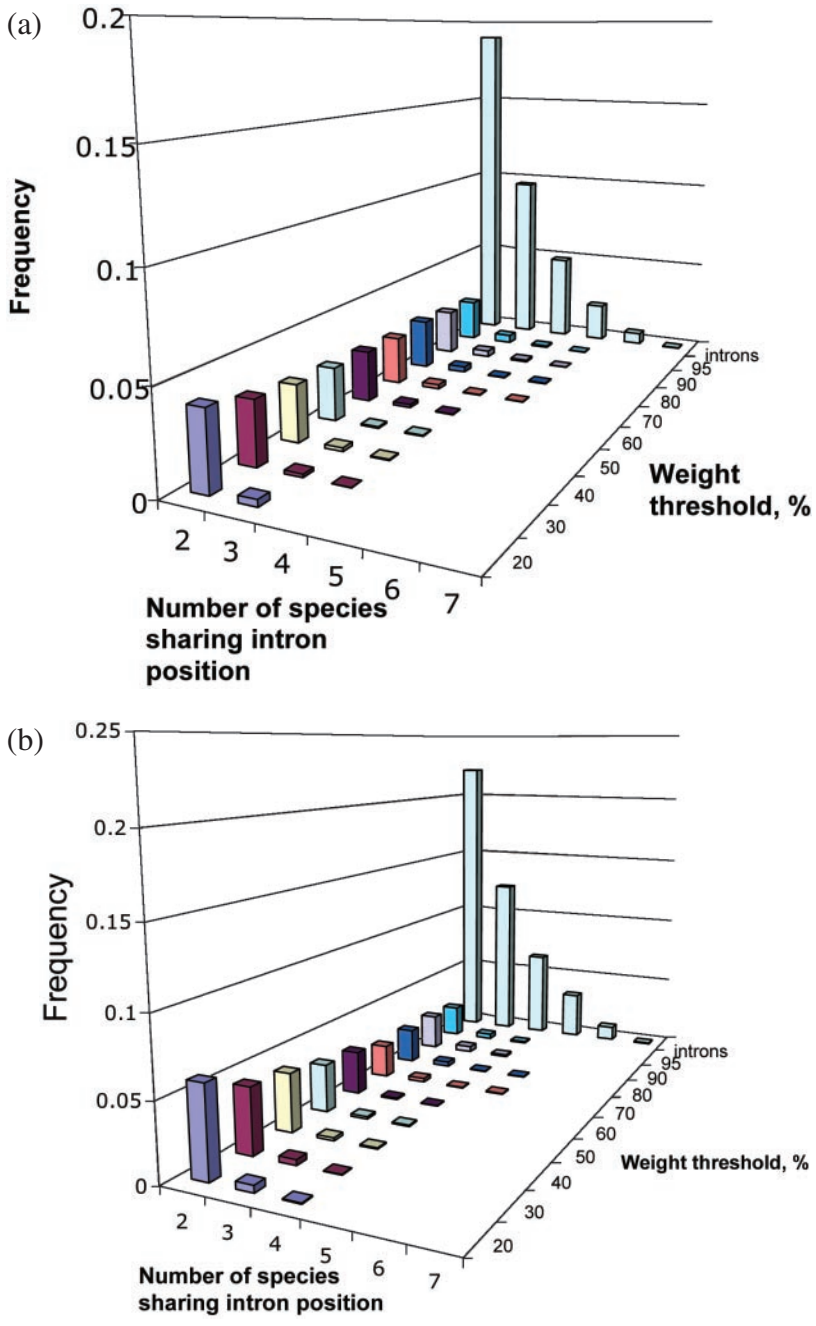

Figure 4. Frequency of independent insertions of introns into the same protosplice sites in orthologous genes from eight species obtained in simulations compared with the observed frequency of shared intron positions (the last row). (**a**) Data for complete alignments of 684 orthologous genes. (**b**) Data for conserved, unambiguously aligned portions only. The results are shown for a series of increasingly stringent weight thresholds used for the identification of protosplice sites. For each threshold, the mean of 1000 simulations is shown; the standard error is given in Supplementary Table 1. The row marked 'introns' shows the actually observed frequencies of shared intron positions.

against widespread existence of rare, conserved cryptic splice sites which could serve as attractors for intron insertion (28).

Clustering of intron positions, which would be best compatible with substantial amount of independent intron gain, was analyzed in several studies which aimed at revealing intron sliding (5,17,18,26). No evidence of intron clustering was found in any of these studies except for an excess of introns separated by one nucleotide which appears to be the only type of sliding occurring with appreciable frequency (17,18). Anecdotal evidence of apparent intron clustering has been recently reported (29). However, statistical analysis (17,18) of the data presented in this work showed that the clustering was not significant (data not shown).

In principle, non-local features of gene organization could affect intron insertions and, if such effects remained undetected, could bias conclusions on the frequency of independent gains. Such features include the avoidance of short exons and the non-uniform distribution of introns across the length of

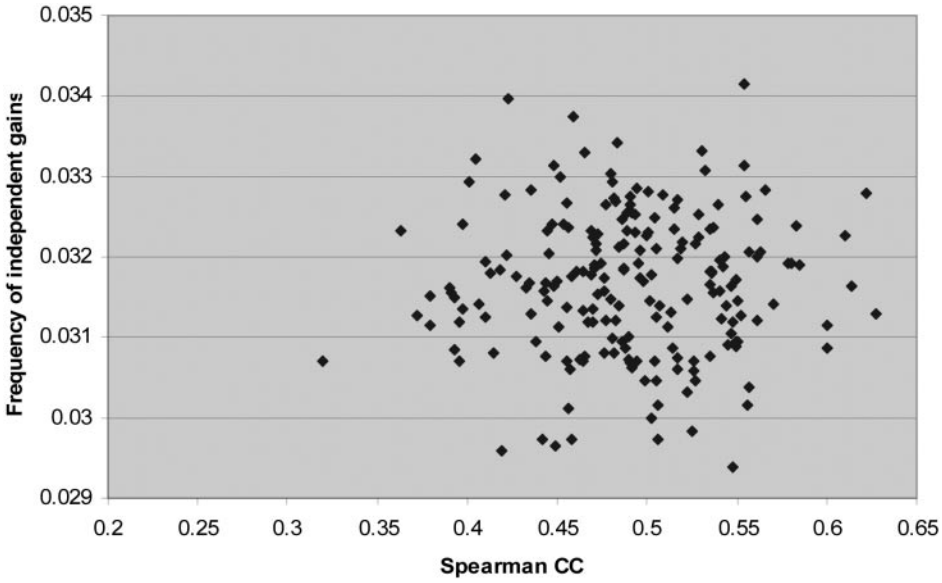

Figure 5. Correlation between the frequency of independent intron gains (complete alignments of 684 orthologous genes) and similarity between observed and simulated distributions of exon lengths (Spearman correlation coefficient). The 60% threshold was used for the identification of protosplice sites (see Results for other threshold values in the Supplementary Material). There was no significant correlation between the two variables: Pearson linear correlation coefficient = 0.027, $P = 0.70$.

genes, i.e. preferential location of introns in the 5′ portions of genes in many species (16,17,19,20). However, these features do not appear to be strongly conserved in evolution. In particular, we observed dramatic differences between intron distributions in animal genomes (20). Therefore, it seems unlikely that such features had a substantial impact on the long-term evolution of introns. This conjecture is compatible with the absence of statistically significant clustering of intron positions in alignments of gene families (5,17,18,26).

A comparison of the nucleotide sequences around the splice junctions that flank introns inferred to be 'old' (shared by two or more major lineages of eukaryotes) or 'new' (lineage-specific) revealed substantial differences between the two classes in the distribution of information between introns and exons (25). Old introns have a lower information content in the exon regions adjacent to the splice sites than new introns but have a corresponding higher information content in the intron itself. This suggests that introns insert into protosplice sites but, during the evolution of an intron after insertion, the splice signal shifts from the flanking exon regions to the ends of the intron (25). Very similar results have been reported for the pairwise comparison of orthologous introns in the nematodes *C.elegans* and *C.briggsae* (30). However, if independent intron gain in the same position was the main reason for the high frequency of introns shared by distantly related species, the opposite trend would be expected because independent gains are more likely to occur in evolutionarily conserved protosplice sites with a high similarity to the consensus sequence (10). Thus, the above analyses of the information content of intron–exon junctions suggest that at least a significant fraction of shared introns result from evolutionary conservation rather than independent intron gains in the same position.

Collectively, all these observations are consistent with the simulation results described here and suggest that parallel, independent gains account but for a small fraction (5–10%) of the shared intron positions in orthologous eukaryotic genes from distantly related species. In other words, the great majority (>90%) of intron positions that are shared by phylogenetically distant eukaryotes, e.g. plants, fungi and animals, seem to reflect bona fide evolutionary conservation. Methodologically, this means that the Dollo principle, i.e. the assumption that the likelihood of parallel gain is negligible, is legitimate for analysis of intron evolution, at least when intron density is relatively low (31).

An important point to be kept in mind is that all estimates of the fraction of conserved introns presented here, by definition of parsimony, pertain solely to shared intron positions. All lineage-specific introns are automatically treated as newly gained (5,31). In a very recent article, which appeared when the present work was under review, Roy and Gilbert developed a maximum likelihood model of intron evolution which led them to the drastic conclusion that many of the lineage-specific introns were likely to be ancient, their evolution having involved multiple, independent losses (32). Accordingly, they estimated that eukaryotic ancestral forms could have had extremely intron-rich genes, with the last common ancestor of plants and animals, perhaps, having almost as many introns as humans. The optimal parameters for maximum likelihood modeling of intron evolution remain to be explored; it seems plausible that the most accurate estimates of the intron density of ancestral eukaryotic forms will fall somewhere between the inherently conservative values presented here and the generous numbers of Roy and Gilbert.

In terms of more general evolutionary scenarios, the obtained results are compatible with the 'many introns early in eukaryotic evolution' view (3,5,33). The biological underpinning of the conservation of numerous intron positions over billions of years of eukaryotic evolution remains a major subject for further studies.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Arlin Stoltzfus, Scott Roy, Masatoshi Nei, Nicholas Dibb, Liran Carmel and Yuri Wolf for useful discussions. Funding to pay the Open Access publication charges for this article was provided by The National Institutes of Health, USA.

Conflict of interest statement. None declared.

REFERENCES

1. Lamond,A.I. (1999) RNA splicing. Running rings around RNA. *Nature*, **397**, 655–656.
2. Logsdon,J.M.,Jr (1998) The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.*, **8**, 637–648.
3. Lynch,M. and Richardson,A.O. (2002) The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.*, **12**, 701–710.
4. Fedorov,A., Merican,A.F. and Gilbert,W. (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl Acad. Sci. USA*, **99**, 16128–16133.
5. Rogozin,I.B., Wolf,Y.I., Sorokin,A.V., Mirkin,B.G. and Koonin,E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.*, **13**, 1512–1517.
6. Dibb,N.J. (1991) Proto-splice site model of intron origin. *J. Theor. Biol.*, **151**, 405–416.
7. Dibb,N.J. and Newman,A.J. (1989) Evidence that introns arose at proto-splice sites. *EMBO J.*, **8**, 2015–2021.
8. Sverdlov,A.V., Rogozin,I.B., Babenko,V.N. and Koonin,E.V. (2004) Reconstruction of ancestral protosplice sites. *Curr. Biol.*, **14**, 1505–1508.
9. Hankeln,T., Friedl,H., Ebersberger,I., Martin,J. and Schmidt,E.R. (1997) A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain. *Gene*, **205**, 151–160.
10. Tarrio,R., Rodriguez-Trelles,F. and Ayala,F.J. (2003) A new Drosophila spliceosomal intron position is common in plants. *Proc. Natl Acad. Sci. USA*, **100**, 6580–6583.
11. Qiu,W.G., Schisler,N. and Stoltzfus,A. (2004) The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol. Biol. Evol.*, **21**, 1252–1263.
12. Huang,X. (1994) On global sequence alignment. *Comput. Appl. Biosci.*, **10**, 227–235.
13. Gelfand,M.S. (1995) Prediction of function in DNA sequence analysis. *J. Comput. Biol.*, **2**, 87–115.
14. Rogozin,I.B. and Milanesi,L. (1997) Analysis of donor splice sites in different eukaryotic organisms. *J. Mol. Evol.*, **45**, 50–59.
15. Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.
16. Smith,M.W. (1988) Structure of vertebrate genes: a statistical analysis implicating selection. *J. Mol. Evol.*, **27**, 45–55.
17. Stoltzfus,A., Logsdon,J.M.,Jr, Palmer,J.D. and Doolittle,W.F. (1997) Intron 'sliding' and the diversity of intron positions. *Proc. Natl Acad. Sci. USA*, **94**, 10739–10744.
18. Rogozin,I.B., Lyons-Weiler,J. and Koonin,E.V. (2000) Intron sliding in conserved gene families. *Trends Genet.*, **16**, 430–432.
19. Mourier,T. and Jeffares,D.C. (2003) Eukaryotic intron loss. *Science*, **300**, 1393.
20. Sverdlov,A.V., Babenko,V.N., Rogozin,I.B. and Koonin,E.V. (2004) Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene*, **338**, 85–91.
21. Stoltzfus,A. (1994) Origin of introns—early or late. *Nature*, **369**, 526–527author reply 527–528.
22. Fedorov,A., Suboch,G., Bujakov,M. and Fedorova,L. (1992) Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res.*, **20**, 2553–2557.
23. Roy,S.W. and Gilbert,W. (2005) The pattern of intron loss. *Proc. Natl Acad. Sci. USA*, **102**, 713–718.
24. Long,M., de Souza,S.J., Rosenberg,C. and Gilbert,W. (1998) Relationship between 'proto-splice sites' and intron phases: evidence from dicodon analysis. *Proc. Natl Acad. Sci. USA*, **95**, 219–223.
25. Sverdlov,A.V., Rogozin,I.B., Babenko,V.N. and Koonin,E.V. (2003) Evidence of splice signal migration from exon to intron during intron evolution. *Curr. Biol.*, **13**, 2170–2174.
26. Cho,G. and Doolittle,R.F. (1997) Intron distribution in ancient paralogs supports random insertion and not random loss. *J. Mol. Evol.*, **44**, 573–584.
27. Gilbert,W. (1987) The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 901–905.
28. Sadusky,T., Newman,A.J. and Dibb,N.J. (2004) Exon junction sequences as cryptic splice sites: implications for intron origin. *Curr. Biol.*, **14**, 505–509.
29. Krauss,V., Pecyna,M., Kurz,K. and Sass,H. (2005) Phylogenetic mapping of intron positions: a case study of translation initiation factor eIF2γ. *Mol. Biol. Evol.*, **22**, 74–84.
30. Coghlan,A. and Wolfe,K.H. (2004) Origins of recently gained introns in *Caenorhabditis*. *Proc. Natl Acad. Sci. USA*, **101**, 11362–11367.
31. Rogozin,I.B., Babenko,V.N., Wolf,Y.I. and Koonin,E.V. (2005) Dollo parsimony and reconstruction of genome evolution. In Albert,V.A. (ed.), *Parsimony, Phylogeny, and Genomics*. Oxford University Press, Oxford, pp. 190–200.
32. Roy,S.W. and Gilbert,W. (2005) Complex early genes. *Proc. Natl Acad. Sci. USA*, **102**, 1986–1991.
33. Mattick,J.S. (1994) Introns: evolution and function. *Curr. Opin. Genet. Dev.*, **4**, 823–831.