Spliceosomal introns as tools for genomic and evolutionary analysis

Manuel Irimia¹ and Scott William Roy²,*

¹Departament de Genètica, Universitat de Barcelona, Barcelona, Spain and ²National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Received December 1, 2007; Revised January 7, 2008; Accepted January 8, 2008

ABSTRACT

Over the past 5 years, the availability of dozens of whole genomic sequences from a wide variety of eukaryotic lineages has revealed a very large amount of information about the dynamics of intron loss and gain through eukaryotic history, as well as the evolution of intron sequences. Implicit in these advances is a great deal of information about the structure and evolution of surrounding sequences. Here, we review the wealth of ways in which structures of spliceosomal introns as well as their conservation and change through evolution may be harnessed for evolutionary and genomic analysis. First, we discuss uses of intron length distributions and positions in sequence assembly and annotation, and for improving alignment of homologous regions. Second, we review uses of introns in evolutionary studies, including the utility of introns as indicators of rates of sequence evolution, for inferences about molecular evolution, as signatures of orthology and paralogy, and for estimating rates of nucleotide substitution. We conclude with a discussion of phylogenetic methods utilizing intron sequences and positions.

INTRODUCTION

Patterns and evolution of intron–exon structures

Spliceosomal introns are sequences that interrupt eukaryotic genes and are removed from RNA transcripts by the spliceosome, a complex cellular RNA–protein machine incorporating five RNAs and hundreds of proteins (1). Our understanding of the evolution of spliceosomal introns has increased exponentially over the past few years due to the release of many genome sequences from most major eukaryotic lineages, both about intron loss and gain dynamics (2–8), as well as the evolution of intron sequences and splicing (9–13).

Several reviews have recently tackled the question of the evolution of introns in eukaryotes (14–19). During the first 25 years after their discovery in 1977 (20), much of the study of spliceosomal introns focused on a debate about the timing of origin of the first introns (21), whether before the divergence of eukaryotes from prokaryotes (which lack spliceosomal introns) (22–25) or within the evolutionary history of eukaryotes (26–31). Although this debate continues, the momentum has clearly tipped towards the perspective that introns appeared once in early (or pre-) eukaryotic evolution by the proliferation and transformation of type II self-splicing introns (26,27,32,33), possibly transferred from the mitochondrion.

Over the past 5 years, the focus has shifted away from the question of the ultimate origin of introns to attempts to track the history of intron loss/gain and intron sequence evolution during eukaryotic history. We can now be confident that large numbers of introns were present by early eukaryotic history (14,34–40) and that many or even most modern introns date to the times of early eukaryotic ancestors. Over at least recent eukaryotic evolution (say, the last ~100 My), intron gain has been a very rare event, with most lineages experiencing rates of gain corresponding to <0.0002 gains per gene per million years (7,8,41–46). Rates of intron loss have been more variable: in some lineages, rates of loss are perhaps 10% per 100 My, whereas other lineages have experienced almost no intron loss over tens or hundreds of millions of years (2,6,7,42,44–48). Figure 1 shows the example of metazoans, where the majority of intron positions have been retained between vertebrates and basal animals.

Moreover, intron positions have been shown to be very constant over time—i.e. that intron ‘sliding’, in which an intron would migrate a few base pairs along a gene, is a very rare occurrence (4,49,50). Several studies have also documented the mechanisms of intron loss: patterns of intron loss including 3'-biased intron loss (7,51–53), exact
removal of intron sequences (2), apparently coincident loss of adjacent introns (7,47,54) and possibly germline-biased intron loss (unpublished data), all seem to indicate that intron loss proceeds via reverse transcription of RNA intermediates (55,56). In addition, comparative analyses have uncovered an apparent (though incomplete) correspondence between rates of intron loss and rates of sequence evolution—degree of loss of ancestral introns appears directly correlated with degree of sequence change [(57) and unpublished data].

Introns as the repository of information about gene structure

The focus of this article is not on these patterns of evolution themselves, but on their implications for analysis of genome structures and eukaryotic evolution in general. Although introns have largely been regarded as a hindrance for genome analysis given the difficulties associated with gene annotation in the presence of introns, intron positions and sequences are potentially very useful in addressing a wide variety of important genomic and evolutionary problems. In particular, intron loss/gain has been shown to be a very slow process in many lineages relative to other genetic characters (sequence evolution of proteins, genes and non-coding DNA, insertion and deletion of transposable elements and even genome rearrangement), thus intron positions contain (and retain) a large amount of information about genome structure and deep evolutionary history.

Over the past few years, a variety of researchers from disparate fields have developed methods that harness this information for purposes ranging from reconstructing evolutionary phylogenies to improving gene prediction, from alignment of homologous protein sequences to assignment of orthology in large protein families. Many of these methods are already quite powerful, though often known primarily to those working on introns themselves, rather than to those working on the problems addressed by the methods. Other methods are still largely undeveloped, and represent promising future lines of work. Here, we review these approaches, and delineate possible uses in genomic and evolutionary study.

Figure 1. Intron positions are often conserved over long evolutionary times. Protein-level alignments of the translation initiation factor 4A gene (TIF4A) from a variety of metazoan species are shown. Intron positions are indicated by digits corresponding to the phase of the intron relative to the surrounding codons (phases 0, 1 and 2 introns fall before the first, second and third bases of a codon, respectively). Most intron positions are conserved at the exact homologous position and phase over all of animal history within these species, all the way from the placozoan Trichoplax adhaerens to chordates. Abbreviations: Hsa, Homo sapiens; Mmu, Mus musculus; Cin, Ciona intestinalis; Bfl, Branchiostoma floridae; Cap, Capitella sp; Lgi, Lottia gigantea; Nve, Nematostella vectensis; Tad, Trichoplax adhaerens.
USES OF SPLICEOSOMAL INTRONS FOR GENOME SEQUENCING, ANNOTATION AND ALIGNMENT

Intron length distributions and genome assembly and annotation

A potential utility of introns for large-scale sequencing efforts involves the distribution of intron lengths. Since introns are removed from protein-coding transcripts, intron lengths are not expected to respect coding frame: across the genome, we expect roughly equal proportions of introns that are multiple of three bases (‘3n’ introns), one more than a multiple of three bases (3n + 1) and two more (3n + 2). However, one of us and David Penny (58) recently reported a survey of predicted genes from genome annotations across 29 different species in which we found common deviations from this expectation. In some cases, the number of 3n introns was much larger than the numbers of 3n + 1 or 3n + 2 introns (Figure 2A), in other cases less. Further investigation indicated that many such cases seemed to be due to genome-wide problems in annotation. Such an internal check for genome annotations could constitute an important step in improving genome assemblies before their public release.

Intron position conservation and improved gene annotation

Intron positions are very often conserved over very long evolutionary distances (Figure 1). In some lineages, this reaches extremes. In *Theileria* apicomplexans, 99.7% of intron positions are conserved between *T. parva* and *T. annulata*, diverged roughly 82 Mya (2). In mammals, 99.9% of intron positions are conserved between human and dog, diverged around 100 Mya (42). Lineages with significant numbers of introns are particularly difficult to annotate in the absence of exhaustive transcript sequence information. Here, comparison with species which are known to have very similar intron–exon structures in orthologous regions could vastly improve uncertain annotations (59). When predicted intron positions are mapped onto protein-level alignments of predicted orthologs, the results are often very clear—protein-level sequence similarity will cease abruptly at the boundary of a species-specific intron position [(2,60,61) and unpublished observations]. While this could in fact reflect
biological reality, it seems very likely that this often reflects misprediction of the intron in one species—either there is an intron present in both species, which has gone unpredicted in one, or there is no intron at that position in either species, and truly exonic sequence in one species has been predicted as an intron.

Figure 3 shows a pair of orthologs from *Plasmodium falciparum* and *P. yoelii*. Clear protein sequence similarity continues through conserved intron positions, and then ends abruptly at a predicted *P. yoelii*-specific intron. Alignment of the genes at the DNA level clearly shows sequence similarity between the predicted *P. yoelii* intron and the predicted *P. falciparum* coding sequence, strongly suggesting that the predicted *P. yoelii* intron instead is coding.

Reconciliation between protein models in species pairs or clusters could greatly improve gene predictions, particularly in species where highly skewed sequence composition or frequent repetitive sequence renders accurate gene prediction most difficult. For example, Coghlan and Durbin (62) recently presented a new method to combine predictions from different gene finders used for one species by comparing intron/exon structures of the different predictions and between gene models of closely related species, building gene structures based on the most conserved exons. They applied this methodology to the nematodes *Caenorhabditis briggsae* and *C. remanei*, obtaining increases of >10% in exon-level specificity and almost 3% in sensitivity, compared to the best outputs from previous gene finders.

**Figure 3.** Reconciling orthologous intron-exon structures to improve gene predictions. (A) Clear protein similarity between the MAL7P1.99 and PY05856 genes of *P. falciparum* and *P. yoelii* continues through two conserved intron positions, and then ends abruptly at a predicted *P. yoelii*-specific intron. (B) Alignment of the genes at the DNA level shows strong sequence similarity between the predicted *P. yoelii* intron and the predicted *P. falciparum* coding sequence, strongly suggesting that the predicted *P. yoelii* intron instead is coding.

**Improved protein sequence alignments**

Since introns often maintain their positions over very long evolutionary timescales, intron positions can often retain information about gene homology after protein sequences have experienced enough change as to render alignment difficult. As such, intron positions can be used as check points in protein alignments, improving their quality. Recently, Csuros and coauthors (63) developed a method to use intron positions to improve protein-level alignments in regions of questionable alignment (Figure 4). They introduced alignment penalties and rewards for intron positions into the alignment matrixes, considering intron position matches/mismatches scoring alternative alignments. The use of these algorithms significantly improved the quality of some gapped alignments.
INTRONS AS TOOLS TO STUDY MOLECULAR EVOLUTION

Intron density as a surrogate for rates of sequence evolution

A central goal of full genome sequencing is understanding the evolutionary history of ourselves and other organisms. Identification of slow evolving lineages (the so-called ‘living fossils’) is thus of central interest, both for what they can reveal about organismal complexity and genome structure evolution. In this context, intron density may serve as an important indicator for branch length. Since rates of intron gain across a wide variety of lineages have been very low (2,6,7,42,44–47), intron numbers in modern species often largely reflect the extent of intron loss since intron-rich ancestors. Intron number is therefore inversely correlated to ‘branch length’, in terms of intron loss.

Interestingly, and somewhat surprisingly given the very different sets of mutation and presumed evolutionary forces controlling intron loss and sequence evolution, a correspondence between degree of intron loss and degree of sequence change has been found in some eukaryotic lineages. The most straightforward case involves a genomic survey across metazoan genomes (57). The genomes which had experienced the least intron loss since the metazoan ancestor (vertebrates and the marine annelid Platynereis) also had experienced less sequence change since the ancestor than other studied lineages (from among insects, urochordates and nematodes). A second example concerns the multitude of nearly intronless protists that also exhibit very high degrees of sequence change (29,64). If in fact a relationship between intron number and branch lengths holds generally, intron density as estimated from small-scale genomic sequencing efforts could be useful in identifying short-branch taxa.

Inferences about molecular evolution

Patterns of intron distribution and evolution can also provide insights into other aspects of molecular evolution. First, intron presence/absence is useful in inferring the mechanism of gene duplication, since intron absence is a hallmark of gene duplication by recombination (65). Thousands of intron-less copies of the so-called ‘processed pseudogenes’ are present in many eukaryotic genomes (66,67), originated by retrotranscription of processed mRNAs and subsequent insertion into the genome (65). This mechanism can be easily distinguished from segmental genome duplication by the absence of introns (65,68). Second, the correspondence of intron positions with the boundaries of domains whose reshuffling contributed to the origin of new proteins in metazoan lineages allows the reconstruction of the mechanism of origin of multi-domain protein-encoding genes (69–73).

Given the apparent dependency of intron loss on reverse transcriptase, rates of intron loss could also provide information about the presence and activity of retroelements through evolutionary history. For instance, we recently showed that rates of intron loss have varied by orders of magnitude in the history of apicomplexan evolution (74). Given the apparent dependence of intron loss on retroelement activity (52), the lack of known active retroelements in modern Plasmodium and Theileria species is consistent with the lack of recent intron loss (74). If so, the much more extensive loss in both the Plasmodium and Theileria ancestors since the genera’s divergence suggests retroelement activity. Thus the pattern of intron loss through time may provide information about the activity of retroelements over evolutionary depths where the actual retroelement insertion history has been erased by subsequent mutation.

Intron positions as signatures of orthology and paralogy

Since the rate of intron loss in modern organisms is often very low, the pattern of intron positions can be used as an indication for orthology among paralogous groups (75–78). These studies usually complement others such as classical phylogenetic analysis of gene families or synteny comparisons, and may be especially useful in the annotation of newly sequenced genomes in assigning orthology among large gene families with very similar domains (such as kinases, TGF-βs, immunoglobulins, etc.) that are hard to distinguish by traditional phylogenetic methods. Some evidence suggests that patterns of intron gain and loss might be different among paralogous groups [although see (80)]. If so, orthology inferences could be hampered by the higher rate of intron change. On the other hand, increased rates of loss and gain could increase intron positions’ usefulness in identifying orthology even over short evolutionary times (with little sequence differentiation).

Estimation of neutral rates of nucleotide substitution

Intron sequences themselves appear to tolerate sequence changes quite easily. Putatively neutrally evolving (portions of) intron sequences are thus a key tool in estimating neutral rates of mutation. Hoffman and Birney (81) recently published a new method to estimate neutral rates of nucleotide substitution based on the study of the substitutions occurring on the alignable introns sequences. They compared their method to a previous method also based on intron sequences (82) and more classic methods based on substitutions in synonymous coding sites, finding a strong correlation between estimates from the two types of methods on different species comparisons. Interestingly, synonymous sites have been shown to be under purifying selection [reviewed in (83,84)], or even under positive selection (85–87). However, introns are also known to contain different types of functional elements (88–90) and thus selection of regions of estimation of neutral rates requires caution.

INTRONS IN PHYLOGENETIC ANALYSIS

Intron sequences and positions contain a record of the evolutionary history of a species or group of species, which presumably contains valuable phylogenetic information. Two phylogenetic strategies have utilized introns as phylogenetic markers at two very different evolutionary depths. Intron sequences, which are relatively fast evolving, have been commonly used to resolve relationships between closely related species. At the other end,
evolution of gene structures by intron loss and gain, which can be very slow in some lineages, has been used in order to resolve deeper nodes over which our confidence in traditional sequence methods can be reduced by the large amount of change over the studied species.

Intron sequences as phylogenetic markers

The use of intron sequences for resolution of relationships between closely related species was established more than 15 years ago (91,92), and it is so common as to barely require comment (93–98). The appeal of intronic sequences for phylogenetics of recent divergences owes to their plausibly being both more rapidly and more neutrally evolving than protein coding and other clearly functional sequences (Figure 5A). As such, relatively simple phylogenetic methods are thought to be of use in utilizing intronic sequences over depths where multiple mutation is unlikely, and over which protein sequences may have experienced too little change to yield sufficient signal. Moreover, obtaining informative intron sequence data sets from non-model organisms is relatively easy and fast using PCR-based methods (91), and the global lack of functional constraints and high potential phylogenetic information content make introns a good complement to mtDNA-based phylogenies for poorly studied groups (99–101).

Intron positions as phylogenetic markers

As mentioned above, intron presence/absence is a relatively very slowly evolving character in most lineages studied to date. Whereas the average number of changes per nucleotide site in putatively unconstrained nucleotide sequence between mouse and human is estimated to be $K_0 = 0.6$ (102), and the degree of protein sequence change around 21.5% (102), a survey of more than 150 000 intron positions between the species found only 120 changes (0.08%), a degree of change three orders of magnitude lower (42). Stajich and Dietrich (47) found <1% intron loss/gain change across a clade of four Cryptococcus species, compared to 35% change in nucleotide sites across the same species (47). Median $dS$ is estimated to be around 0.49 for the Plasmodium parasites P. falciparum and P. yoelii (103), but fewer than 1.5% of intron sites have experienced a loss/gain event (45).

Relative to sequence-based studies, phylogenetics using intron presence/absence is truly in its infancy. Among the very few published studies, there are essentially three strategies. First, some authors have used one or a few intron loss/gain patterns to group species into a clade (78,104,105) (Figure 5B), since as rare genome change (RGC) losses/gains could be theoretically highly parsimonious (106). As RGCs, introns have the advantage of having a very wide taxonomic resolution, potentially low homoplasy and applicability to a broad range of eukaryotic groups (106). Here, however, significant caution is necessary. First, while such ‘magic bullet’ approaches may work well for groups with very few intron changes, the possibility of homoplasy (in particular, of multiple loss of the same intron) is likely in cases where there is more change, for instance, nematodes and dipterans (107); second, individual cases attest to the recurrent loss of the same intron, even while flanking introns remain intact (107,108). Until the degree of such variation across sites is better known, individual cases of intron loss/gain as phylogenetic markers should in our opinion be viewed as non-conclusive in many cases.

A second strategy involves using explicit phylogenetic models to analyze large collections of intron presence/absence data across species. One of us and Walter Gilbert (109) used intron presence/absence data for 684 sets of eukaryotic orthologs across eight animal, fungus, plant, and apicomplexan species (4) to develop a method for resolving deep nodes in metazoan phylogeny. Nguyen and coauthors (38) then developed a more general phylogenetic method for analysis of the same kind of data. These data were used to address the relationship between arthropods, nematodes and deuterostomes, with both analyses placing deuterostomes as the outgroup.

Again, there is reason for caution here. Both analyses assumed constancy of intron loss rates across intron sites, an assumption which is unlikely to hold generally, and which might bias methods towards long branch attraction and other recurrent problems in phylogenetics, especially important in this particular phylogeny (110–114). Moreover, the small absolute number of characters obtainable from such an exhaustive comparative genomic effort may make it difficult to obtain sufficient knowledge of the shape of the distribution of rates across sites, which may make these shortcomings difficult to correct for. On the other hand, the availability of representatives of the groups in question that have experienced less intron loss will presumably allow for more confident reconstruction.

The third and perhaps most promising (as well as intriguing) strategy was developed by Krauss and coauthors (115), and aims to overcome these problems by restricting the analysis to cases in which ancestral and derived states can be more confidently inferred. Very short exons (e.g. shorter than 50 bp) are very rare across most characterized species, likely due to problems associated with accurate splicing of such regions. Therefore, introns found at nearby positions in orthologous coding regions are unlikely to have coexisted. Assuming that multiple insertions into the same site are very rare, one can then confidently infer that there is an edge on the phylogenetic tree (along which both intron loss and gain has occurred) separating those species that share one of the positions from those that share the other (Figure 5C). In cases in which a known outgroup shares one of the two positions, one can furthermore infer the directionality of this change, and therefore place all derived species into a clade. However, it remains to be seen whether (and in what cases) sufficient numbers of such intron pairs will be obtainable, thus it is not yet clear to what extent this model will be generalizable.

A fourth possibility that has not to our knowledge been explored would utilize shifts of intron boundaries (i.e. shortening or lengthening of adjacent coding sequence) (Figure 5D). Analysis of large numbers of alignments in various lineages shows that such changes are very rare indeed, and that such changes might represent useful
Figure 5. Introns as tools for phylogenetics. (A) Introns sequences are often used to resolve phylogenetics relationships between closely related species. Here, four phylogenetically informative sites in an alignment of sequences of orthologous introns from five Drosophila species (the 1st intron from the CG10050 locus) are consistent with the accepted consensus relationship of the species (left). (B) Intron loss/gain and position conservation as phylogenetic characters. Intron position conservation across subsets of species suggests phylogenetic groupings. Under many circumstances, the high degree of intron loss/gain change requires correction for the possibility of multiple changes at a site (3). (C) Nearby pairs of homoplasies. Since short exons are rare across various genomes, introns at nearby positions in orthologous genes are unlikely to have coexisted, suggesting multiple rare changes (an intron loss followed by a nearby gain) between putatively ancestral and derived gene structures. Here, two pairs of nearby intron positions in the CG10050 locus are consistent with the accepted consensus relationship of the species (left).
genomic changes. A major concern in developing methods using these changes will be exclusion of incorrect genome annotation as an explanation.

An important caveat to the seeming usefulness of intron loss/gain across species is the possibility of highly skewed distributions of rates across sites (alluded to above). In a few cases, careful study of closely related species with known phylogenies has indicated that a single intron position has been subject to striking recurrent intron loss while other intron positions have remained intact (107,108). In such cases, the observed phylogenetic position will give support to inaccurate phylogenetic groups. It is not currently known how general this pattern of recurrent loss of the same introns is, and so it is not yet clear how much of a problem this may constitute in large-scale or genome-level comparisons. Strong confidence in use of intron losses/gains as phylogenetic characters awaits a better understanding of the causes and generality of large differences in loss rates across sites.

CONCLUSION

Problems associated with signal-to-noise ratios are ubiquitous in bioinformatic, genomic and evolutionary analyses. As slowly evolving characters, intron positions provide useful and otherwise scarce information. We have reviewed well-developed, early-stage and potential uses of introns as tools for addressing a wide range of problems. We look forward to development of further methods.

ACKNOWLEDGEMENTS

M.I. was funded by the Spanish Ministerio de Educación y Ciencia, through the FPI grant (BFU2005-00252), and S.W.R. by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS. Funding to pay the Open Access publication charges for this article has been provided by the National Institutes of Health. We thank Eugene Koonin and Jordi Garcia-Fernandez and their groups for intellectual support and stimulation, for financial support, and for fostering environments of open intellectual exploration in their respective groups. SWR thanks Rogério da Silva and all of Marcelo Ferreira’s group for heroes of hospitality and friendship during the course of this project.

Conflict of interest statement. None declared.

REFERENCES


