Protein-coding Segments: Evolution of Exon–Intron Gene Structure

Manyuan Long, *The University of Chicago, Chicago, Illinois, USA*

Many genes are disrupted by noncoding regions of DNA of variable sizes called introns, giving the genes an exon–intron structure. Most introns, particularly spliceosomal introns found in the nuclei of higher eukaryotes, have no obvious functions.

Introductory article

Article Contents

- Introduction
- Discovery of Split Genes in Adenovirus and Other Systems
- Theories of Intron Origin
- Relationship Between Exons and Protein Domains and Modules
- Evolution of Spliceosome-catalysed mRNA Splicing versus Self-splicing
- Possible Relationship Between Splicing and Stop Codons
- Nonsplicing Information in Introns

Introduction

Many genes encoding proteins and other molecular elements are disrupted by noncoding DNA regions called introns. This split structure of genes, the exon–intron gene structure, is typical of most higher eukaryotic genes, some lower eukaryotic genes (of protists), and some prokaryotic genes. In general, most introns, particularly those in the nucleus of higher eukaryotes, called spliceosomal introns, have no obvious functions. The sizes and numbers of introns in genes are variable from gene to gene and from organism to organism, suggesting that there is little functional constraint on introns. That large genomes have gigantic introns (e.g. the introns in human dystrophin genes) and small genomes can host only small introns (e.g. the introns in nucleomorph, a vestigial nuclear genome in algal cells), however, indicates the constraint of genome sizes to some extent. These simple phenomena have raised a number of issues around the significance of the existence of introns and stimulated interest in the study of the evolution of exon–intron structures.

Discovery of Split Genes in Adenovirus and Other Systems

The conventional idea that a gene was a continuous genetic unit was overturned in the late 1970s when the existence of split genes was discovered. To establish the complementarity between DNA sequences of genes and cytoplasmic messenger RNAs (mRNAs) in eukaryotic cells, a number of experiments were conducted in which the mRNA encoding an adenovirus 2 (Ad2) protein hexon and other Ad2 mRNAs were hybridized to restricted DNA fragments of adenovirus 2 genome. Under the electron microscope, it was observed that the hybridization generated loops of viral DNA of different lengths, because the viral DNA in the loops contained no counterparts in the mRNAs as a consequence of RNA splicing. These loops were the intron sequences.

Soon after the discovery of adenovirus introns, several other experiments showed that cellular genes also contain introns. In a sequence comparison of mRNA and genomic DNA using restriction-mapping analysis, it was found that the gene encoding ovalbumin was interrupted by introns. In another experiment, by directly comparing the DNA sequence to the protein sequence, it was found that the gene for immunoglobulin light chain consists of the precursor region followed by a short intron, followed by the variable region, separated by a long intron from the constant region. Using both restriction mapping and electron microscopic analysis, as used in the determination of introns in the Ad2 genes, two introns were found to interrupt the β-globin genes in mammals.

Theories of Intron Origin

There are two competing theories to account for the origin of introns: the introns-late theory and the introns-early theory, with different hypotheses about the date of the origin. Both theories have observational support. Walter Gilbert at Harvard University extended the introns-early theory to the exon theory of genes by elaborating broader issues of gene evolution. Compared to the introns-late theory, the exon theory of genes has been supported by more independent lines of evidence. It has more explanatory power in gene evolution and invokes fewer and simpler assumptions in the interpretation of a number of important observations.

The exon theory of genes has three major components that address the molecular mechanisms of the origin of new genes, the age of introns, and the subfunctional property of exons in a panoramic description of the entire process of gene evolution. These are the concept of exon shuffling, the antiquity of introns, and the modular comparison of genes.

1. Exon shuffling. Novel genes can be created by the juxtaposition of various preexisting exons in new combination, here called recombination. This process

of recombination is facilitated by introns. A clear example of exon shuffling is the origin of the plant cytochrome c_1 gene, where the function of the mitochondrial-targeting domain originated from the glyceraldehyde-3-phosphate dehydrogenase gene by exon recombination (**Figure 1**).

2. Ancient origin of introns. Introns existed before the emergence of eukaryotes, probably at the beginning of gene evolution. This idea was based on a genetic observation and on likely chemical properties inherent in the earliest self-replicative polynucleotide chains. Genetically, transposable elements inserted into protein-coding regions are usually deleterious or lethal. Thus, it is thought that the insertion of transposable elements or other genetic elements is unlikely to have been a process to create introns. Chemically, RNA probably preceded DNA as a genetic molecule (during a phase termed the 'RNA world'). In the RNA world, genes encoded in RNA might have evolved exon–intron structures processed by splicing.
3. Exons often represent independent functional units and thus genes possess a modular composition. The variety of possible combinations of independent modular units can create a huge diversity of proteins from a limited number of exons. Thus, the exon theory of genes not only encompasses eukaryotic genes but also elaborates the origin of genes in general.

Introns-late theory, in contrast, assumes that all spliceosomal introns are recent arrivals by the insertions of either group II introns or transposable elements into eukaryotic genomes. It is possible that some introns may have originated recently, although no clear cases of inserted introns from these two sources have been found. However, a phylogenetic analytic analysis of a large number of introns has revealed that the intron density (the number of introns per unit of protein-coding region) is uneven in the major branches of the eukaryotic tree, as Jeffery Palmer and John Logsdon found. This nonrandom distribution of intron density was interpreted as evidence for the recent origin of introns. Meanwhile, phylogenetic analysis of some gene families has indicated that some introns in these genes seem to be explained better as newly inserted introns.

A logical difficulty is apparent for these approaches, however, because proving the recent origin of some introns does not prove that all introns originated recently. A more likely picture is that some introns are ancient and some recent. This mixed model of introns-early theory, which incorporates extensive intron loss and intron gain, can generate a phylogenetic distribution of intron density that mimics the distributions observed by Palmer and Logsdon.

In addition to the exon theory of genes, there are also several other versions of the introns-early proposed by different groups of authors. One version is the minigene theory of Jeremy Knowles and his colleagues at Harvard University, which holds that each single exon represents a single original gene. Another version is the split-gene theory by Periannan Senapathy in National Institutes of Health, which proposes that the existence of exon–intron structures in the progenote is a consequence of the limited length distribution of coding frames in random sequences.

Relationship Between Exons and Protein Domains and Modules

A correspondence between exons and protein structures was first proposed in connection with the minimum conditions that a polypeptide chain has to meet. It was argued that a peptide newly created by exon shuffling has to

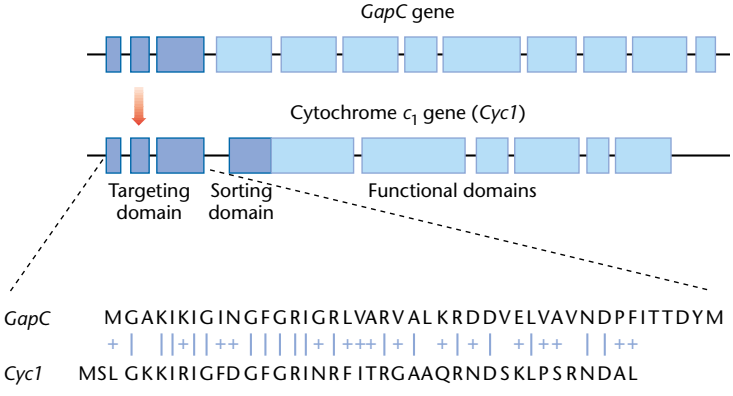

Figure 1 The mitochondrial-targeting presequence of the cytochrome c_1 (*Cyc1*) precursor in potato originated from the plant *GapC* gene for cytosolic glyceraldehyde-3-phosphate dehydrogenase. The boxes represent exons. The three duplicate exons from *GapC* genes recombined with the mitochondrial-orginated nuclear gene encoding cytochrome c_1 to form a complete cytochrome c_1 precursor gene (*Cyc1*). The presequence encoded by the three exons in *Cyc1* has a mitochondrial targeting function. The similarity between the donor peptide sequence (*GapC*) and the peptide sequences encoded by the shuffled exons in the acceptor gene (*Cyc1*) are significantly high (64% similarity), as shown in the alignment of the two amino acid sequences.

be properly folded (e.g. folded into a stable globular form with an active site) to carry out its essential function. Exon shuffling would work efficiently only when the exons correspond to intact folded protein fragments. A shuffled exon that represents a random piece of peptide would be less likely to change original protein into a new protein with novel function. When the protein structure data of haemoglobin was examined using a visualization tool called the Go plot, a correlation between protein structures and intron positions was observed. The peptide fragments that correspond to exons are called modules, suggesting a modular structure of proteins delineated by the exon–intron structures of genes.

In a world of ancient introns without recent addition or movement of introns, it is expected that all proteins would show a clear correspondence between exons and protein modules. The loss of introns would leave some gaps between intact modules in a particular protein. By examining such gaps in triose phosphate isomerase (TPI), Walter Gilbert and colleagues in 1986 predicted the existence of an intron; the predicted intron was observed 6 years later. However, any recent random insertion of an intron in a protein might blur the original correspondence. The extent of disruption to the correspondence depends on the number of inserted introns. This mixed model of ancient introns with recent intron gains is difficult to distinguish from the pure introns-late model. In the latter extreme model, all introns are assumed to arise recently; there is no correspondence between exons and protein structure. This difficulty in comparison of the two models becomes even more apparent in a small number of intron positions in some gene families, which would decrease statistical power for detection of signals from the early evolution of the exon–intron structure of genes.

DeSouza–Gilbert boundary

Although previous studies in TPI and other proteins did not resolve the issue of correspondence between intron positions and protein structure, these tests have stimulated interest in the development of a new approach to distinguish the two alternative models (a mixed model of introns-early and an extreme introns-late model where all introns are assumed to be recent arrivals). The new approach is to consider whole populations of ancient genes, instead of examining one or a few gene families with limited introns, and to develop criteria to define modules objectively, rather than arbitrarily.

Sandro DeSouza and Walter Gilbert proposed an objective way of defining boundary regions between modules. By superimposing the distances of residues of proteins in a two-dimensional space (plane), as in a Go plot, the boundary between two adjacent modules is defined as the region that corresponds to the dark region in which two residues are assigned to two different modules. This boundary is called the DeSouza–Gilbert boundary (DG boundary) (**Figure 2**). The boundary can be defined by continuous distance criteria, not by only one distance criterion as used previously. In each defined boundary, the excess of introns, if any, in the boundary regions can be measured by counting introns within and outside the boundary and comparing the results with the expected intron numbers based on the proportion of the total length of boundary regions over the entire protein.

Based on the definition of the DeSouza–Gilbert boundary, the distribution of 44 ancient genes with 988 introns was analysed using computational biological techniques. Unsurprisingly, some introns were found to interrupt modules. However, such inserted introns were far less frequent than random predictions would suggest, and a very significant proportion of introns was found to be within the DG boundaries (these introns are of the type of the intron located between two intact condons). Thus, a significant statistical correlation exists between exons and modular protein structures as defined by the DG boundary. This strongly supports the mixed model of introns-early that incorporates both ancient and recent introns, and rejects the extreme introns-late model.

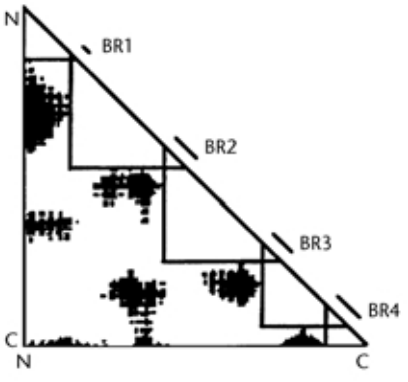

Figure 2 DeSouza–Gilbert boundaries are defined as the regions of overlap of the triangles in a Go plot. A Go plot is a graphic representation of the distances between each pair of α-carbons in a three-dimensional structure of a protein. In this plot, distances greater than the defined size (28 Å) are shaded black and any triangle drawn along the diagonal that does not contain a black region defines a module. The overlapping region of two adjacent triangles represents the boundary region (BR) between the two protein modules (DeSouza–Gilbert boundary), as the four boundary regions (BR1 to BR4) indicate. Reproduced with permission from DeSouza SJ *et al.* (1996).

Evolution of Spliceosome-catalysed mRNA Splicing versus Self-splicing

The spliceosome is the machinery that recognizes and splices introns from nuclear pre-mRNA. It is assembled sequentially from a number of small nuclear ribonucleoprotein particles (snRNPs) and proteins in a highly coordinated process that controls its operation. Based on the types of snRNPs, spliceosomes are classified into two types: the U2-type comprising U1, U2, U4/U6 and U5 snRNPs and the U12-type comprising U11, U12, U4atac/U6atac and U5 snRNPs. The U2-type spliceosome catalyses the splicing reaction of most nuclear introns with GT-AG dinucleotides at the ends (U2-type introns); the U12-type spliceosome catalyses the removal of the minor class of introns with AT-AC ends (the U12-type introns). It was also observed that some introns with a GT-AG boundary are spliced by the U12-type spliceosome and some introns with an AT-AC boundary are removed by the U2-type spliceosomes.

Spliceosomes are ancient, as shown by the many similar RNA base-pairing interactions among the snRNPs between distantly related organisms such as yeast and humans. U6 snRNPs in the U2-type spliceosomes of yeast and human are very similar in size, sequence and structure (the sequence similarity is $\sim 60\%$). Since the U12-type introns have been found in plants, *Drosophila* and vertebrates, the U12-type spliceosomes must also have coexisted with the dominant U2-type spliceosomes for more than a billion years. However, the ancestral relationship between the two types of spliceosomes is unclear as yet.

In addition to the nuclear introns whose splicing depends on spliceosomes, there are also some introns that do not depend on any protein enzymes for their splicing. These are the self-splicing group I introns. Group I introns are sporadically distributed in eukaryotes, eubacteria, bacteriophages and some organelles. Group II introns in organelle genomes and some bacteria can also be self-spliced *in vitro*. Both groups of introns form elaborate secondary structures resulting from intramolecular base pairing. Group II introns even create a lariat-like structure during the splicing process. Finally, transfer RNA (tRNA) introns are also found in both archeae and eukaryotes.

The splicing reactions for self-splicing introns and spliceosomal introns are strikingly similar, suggesting some evolutionary relationship between them. In particular, the similarity between spliceosomal introns and group II introns indicates that the former might have evolved from the latter. In this picture, the snRNPs in spliceosomes could have originated as fragments of group II introns.

Possible Relationship Between Splicing and Stop Codons

In a computer modelling conducted by Senanpathy, the open reading frames in random sequences have an upper limit of 600 nucleotides (or 200 codons have an upper limit of 600 nucleotides (or 200 codons) before being interrupted by stop codons. Comparison of length distributions of open reading frames in random sequences and those in existing sequences in DNA databases leads to the conclusion that current protein-coding genes look like random sequences; the longest exons, with rare exceptions, are shorter than 200 codons.

Thus, it was suggested that split genes with spliceable introns were a means to skip clustered stop codons to obtain structurally complete proteins. According to this split gene hypothesis, the earliest protein-coding genes possessed a structure disrupted by introns harbouring stop codons. This hypothesis also suggested that prokaryotes evolved from hypothetical primitive eukaryotic organisms. The prokaryotic genes lost introns and the prokaryotic cells lost nuclear membranes that had existed in the primitive eukaryotic cell, because it was observed that the codon distribution mimics the distribution of spliced eukaryotic genes. One argument for this hypothesis is the nonsense codon-like trinucleotides in the consensus sequences for splicing in introns.

However, two observations may cloud this hypothesis. One is that the distribution of exon length in DNA sequence databases have both too many short and too many long exons; this is inconsistent with the modelled random distribution. The second is that the nonsense codon-like trinucleotides in the intron consensus are not in positions where they would serve as stop codons.

Intron phases: statistical analysis

The intron phase refers to the position of an intron within or between codons. An intron located between two codons is defined as phase 0, after the first nucleotide as phase 1, and after the second nucleotide as phase 2. Intron phase is a highly conserved evolutionary characteristic, because any change requires a simultaneous insertion and deletion to avoid an often detrimental reading frameshift or other more complex mechanisms. A statistical analysis of intron phases revealed some unexpected patterns in the evolution of exon–intron structures.

Since with minor exception, introns are functionless, a random distribution of introns in eukaryotic genes was expected both with respect to the proportions of the three phases of introns and their distribution within genes. Two simple forms of quantitative expectation for this prediction are (1) the proportion of a single intron phase (f_i) given by:

$$f_i = 1/3$$

where the subscript $i = 0$, 1 and 2 stands for the intron phase, and (2) the association frequency of introns that flank a single exon (f_{ij}) given by:

$$f_{ij} = f_i \times f_j$$

These predictions were found to be inconsistent with the observations, however, when a large number of introns from all independent gene families with sequence data available in sequence databases were investigated.

For example, in the GenBank database, release 101, there are 25 666 intron-containing genes. After redundant sequences were discarded, 1997 independent genes whose exon–intron structures had been experimentally determined were recovered. It was found that of the 13 290 introns contained in these genes, 48% were phase 0 introns, while only 30% and 22% were phase 1 and 2, respectively. These significantly different proportions of the three intron phases indicate a nonrandom distribution. Furthermore, when the distribution of intron phases within genes was examined, it was found that not only are the introns not distributed randomly, but the introns with the same phase tend to cluster together (**Table 1**). That is, there is an excess of exons that are flanked by introns of the same phase. Such exons are called symmetric exons, and exons flanked by introns of different phases are called asymmetric exons.

There is a connection between these statistical phenomena and one molecular mechanism of creating new genes: exon shuffling. One critical condition for exon shuffling is the maintenance of the original reading frame in acceptor genes. This is achieved only if the length of the inserted (shuffled) exons is a multiple of three; if it is not, the original reading frame will be broken. The outcome of successful exon shuffling is the creation of a symmetric disposition of introns. Thus the excess of symmetric exons as shown in **Table 1** indicates that a large amount of exon shuffling has occurred in eukaryotic genes. A conservative statistical estimate based on intron phase analysis is that at least 20–30% of exons in eukaryotic genes have been involved in exon shuffling, suggesting that more than 50% of modern eukaryotic genes originated from exon shuffling. No better alternative explanation than exon shuffling could be provided from our present-day knowledge of molecular biology.

The large proportion of phase 0 introns is also in accord with the exon theory of genes, which proposes that early genes encoded short peptides and modern proteins evolved from combinations of these original short peptides. However, stronger evidence came from the analysis of intron phases in ancient conserved regions (ACRs). ACRs are the regions (partial or complete) of eukaryotic genes that have counterparts in prokaryotic genes. After constructing an ACR database by using pairwise comparison of eukaryotic independent intron-containing gene databases with prokaryotic gene databases in GenBank 101, more than 910 genes with 4151 introns that flank at least one exon were isolated. The statistical distribution in this ACR database is similar to that of the overall database. The proportions of the three intron phases are: 54% were phase 0; 25% were phase 1; and 21% were phase 2. The distribution of introns within genes also revealed a significant excess of symmetric exons (**Table 1**).

As is the case in the overall database, these excess symmetric exons in the ACR database are also the products of exon shuffling. It might be thought that these shufflings would have happened after the divergence of eukaryotes from prokaryotes. However, the late shuffling in the single lineage of eukaryotes would have broken the colinearity between eukaroytic ACRs and their prokaryotic counterparts. Thus the inference is that exon shuffling in the ACR database must have happened before eukaryotes diverged from prokaryotes. The observation of ancient exon shuffling in the ACRs of eukaryotic genes

Table 1 Observed and expected intron phase associations

	Symmetric exons			Asymmetric exons						
	(0,0)	(1,1)	(2,2)	(0,1)	(0,2)	(1,2)	(1,0)	(2,0)	(2,1)	Exon no.
Overall database										
Observed number	3051	1303	620	1321	1184	749	1408	1219	704	11 559
Expected number	2709	1013	558	1657	1230	752	1657	1229	752	
ACR database										
Observed number	1046	237	165	378	346	170	386	351	162	3241
Expected number	934	210	140	444	362	172	444	362	172	

The observed number is the actual count in the exon databases. The expected number is calculated as a product of $P_i \times P_j \times N$ based on the assumption of random distribution of introns, where N is the total number of exons and P_i is the proportion of the intron of phase i. (i, j) indicates an internal exon that is flanked by a 5′ intron of phase i and a 3′ intron of phase j.

provides the exon theory of genes with a strong line of independent evidence.

Nonsplicing Information in Introns

No obvious biological functions have been observed in the spliceosomal introns of protein-coding genes. However, in some rare cases introns may contain nonsplicing information, including transcription elements, small nuclear RNAs (snRNAs) and even independent genes.

Transcription elements

Intron-encoded promoters are often located at sites near the starting point of RNA. For example, the first intron of the *Adh* gene in *Drosophila melanogaster* harbours a promoter for larval mRNA transcription. A promoter in the human γ-glutamyltransferase gene resides in intron 7. Enhancers were found in the first intron of the human tyrosine phosphatase gene PRL-1 and in the second intron of the human apolipoprotein B gene. An attenuator was found to be encoded in the intron of the c-*fms* proto-oncogene in mammals. An unusual regulatory element is found in human globin genes where the intron transcripts can bind to promoters to down- and upregulate expression. These examples suggest the important roles of some spliceosomal introns in the regulation of gene expression.

Genes within introns

A gene encoded in an intron creates a nested coding structure. A classic example is the *Gart* locus in *Drosophila*. The opposite strand of the first intron in this gene encodes a gene for pupal cuticle protein. A recent example is that the third intron of the yellow emperor gene in *Drosophila* harbours an unrelated gene, musashi, which encodes RNA-binding protein. A more extreme case was found in the gene *F10F2.2* encoding FGAM (phosphoribosylformylglycinamidine) synthase of *Caenorhabditis elegans*. Four open reading frames are in one large intron of *F10F2.2* and the fifth is in a second intron.

Small nucleolar RNAs in introns and degenerate exons

It is generally accepted that exons encode functionally useful mature RNA and introns are just junk. However, one remarkable case displaying the opposite situation is the mammalian gene *UHG* (U22 snoRNA host gene). This gene encodes eight snoRNAs U22 and U25–U31 in its introns, but its mature RNA does not encode any protein and degrades very quickly in the cell.

Further Reading

DeSouza SJ, Long M, Scheonbach L, Roy SW and Gilbert W (1996) Intron positions correlate with module boundaries in ancient proteins. *Proceedings of the National Academy of Sciences of the USA* **93**: 14632–14636.

Logsdon JM Jr (1998) The recent origins of spliceosomal introns revisited. *Current Opinion in Genetics and Development* **8**: 501–508.

Long M and DeSouza SJ (1998) Intron–exon structures: From molecular to population biology. *Advances in Genome Biology* **5A**: 143–178.

Long M, Rosenberg C and Gilbert W (1995) Intron phase correlations and the evolution of intron/exon structure of genes. *Proceedings of the National Academy of Sciences of the USA* **92**: 12495–12499.

Sharp PA (1994) Split genes and RNA splicing. *Cell* **77**: 805–815.