REVIEW

# Introns and gene evolution

**Sandro J. de Souza, Manyuan Long and Walter Gilbert***

*The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA*

**In one scenario of gene evolution, exon shuffling has a fundamental role in increasing gene diversity. As DNA sequences accumulate in the databases, the picture of the intron/exon structures of genes becomes more and more clear. We discuss in this review some features of this picture that suggest that introns have been present since the early stages of evolution, and that exon shuffling was a fundamental process in the construction of ancient as well as modern genes.**

## Introduction

The first applications of DNA sequencing in the late 1970s soon showed that the genes in eukaryotes were discontinuous, having an intron/exon structure, in distinction to the continuous genes of prokaryotes. However, almost two decades of research have still not answered the questions: What is the role of the introns? and Where do they come from? Even though one of the great achievements of molecular biology during this last 20 years has been the creation of the exponentially expanding dataset of DNA sequences, which today includes almost 500 megabases of sequence, only a small fraction of these sequences represent genomic versions of genes that have an intron/exon structure. Today the databases contain about 2000 independent genes with an intron/exon structure of about 10 000 exons. The next decade, which should carry the databases to 20 gigabases of sequence, will produce an explosion of knowledge about the intron/exon structure of genes and should resolve the current controversies.

In general, introns are widespread in the genes of complex eukaryotes, while simple eukaryotes and prokaryotes lack them. In plants and vertebrates, the introns are about 10-fold longer than the exons, so most of the DNA lies in the introns. If one compares homologous genes in different organisms, the exon sequences are well conserved while the intron sequences, simpler in character, vary rapidly, over evolutionary time, in sequence and in length, but not generally in position. Broadly speaking, organisms with small genomes lack introns or have small introns, while organisms with large genomes have many introns.

The critical issues in the evolution of the intron/exon structures are: What is the origin of the introns? What is the general role of exon shuffling in the evolution of genes? And finally, how is protein diversity created from a limited exon diversity?

The discovery of introns immediately suggested that the introns might have a necessary role in the biology of the eukaryotic cell and thus be preserved because they had some selected function in evolution. However, over the years, no generally necessary role for the introns has emerged. Occasionally a control sequence lies in an intron and so provides that intron with an essential role in the functioning of that gene. However, many genes have 10–50 introns, and these cannot all be essential for control reasons. At one time it was thought that splicing was necessary for a messenger to be transported from the nucleus. However, since some genes in the higher vertebrates are intronless, ranging from the histones, to the olfactory receptors, to pairs of genes for the same enzyme, one of which has a complete intron/exon structure while a second functional copy has no introns, the presence of introns and splicing cannot be a necessary step in message function. But if the introns are not being currently selected for by evolution, why are they there? One attitude, the argument for exon shuffling (Gilbert 1978), argues that they are there simply because they have been used in the past by

* *Corresponding author:* Fax: +1 617 496 4313.

evolution to create novel genes. Introns increase recombination between exons: moving the exons several thousand bases apart along the DNA increases the rate of legitimate recombination between the exons one-hundred- to one-thousand-fold and increases the rate of illegitimate recombination, which creates novel connections between the exons, by factors of $10^6$ or $10^8$. Thus the introns are hot spots for recombination in the formation of new combinations of exons and, hence, new genes. It is not that introns are present because they will be involved in the future in the creation of new genes, although that is a consequence of the intron/exon structure, but, more correctly, that they are present because they have been used in the past history of the organism as the faster pathway to create the genes currently present. The view that introns are used to enhance recombination to produce exon shuffling is independent of the issue of the origin of the introns but provides an evolutionary advantage for them.

An alternative view would be that the introns are entirely functionless, but they appear in certain organisms because some random process introduces them down certain evolutionary lines at a rate more rapidly than counterselection or drift would dispose of them. (The selfish DNA idea was expressed by Doolittle & Sapienza (1980) and Orgel & Crick (1980).) Such selfish DNA conceptions are true for transposable elements. However, selfish DNA generally shows, through sequence comparison, the origin and movement of these families of transposable elements or repetitive sequences. Intron DNA, strikingly, does not have a repetitive character and is generally unique DNA in the organism.

A debate about the origins of the introns arose soon after they were discovered. A conflict between the extreme theories of 'introns-early' (Doolittle 1978; Blake 1978; Darnell 1978; Gilbert 1979; Gilbert & Glynias 1993; Long *et al*. 1995b) or 'introns-late' (Cavalier-Smith 1978, 1985, 1991; Palmer & Logsdon 1991; Stoltzfus *et al*. 1994) lies at the heart of the evolutionary consideration of the intron/exon structure. The introns-early view was expanded in the Exon Theory of Genes (Gilbert 1987). In this review we will sketch a scenario proposed by the Exon Theory of Genes and discuss some features of the present-day genes that can be seen as consequences of the evolutionary process envisaged by this theory.

## The exon theory of genes

The exon theory of genes pursues the idea that the shuffling of a limited number of exons whose products are modular folding elements is an extremely potent way to generate a large diversity of protein structures. This idea is not only attractive for the later stages of evolution, to create the novel genes needed for multicellularity and the Cambrian explosion, but also would make the creation of the first complex proteins easier. The exon theory of genes suggests that the first proteins were essentially aggregates of short polypeptides, perhaps 15–20 amino acid residues in length, and that genes were initially assembled from exons of this size. The processes for evolution were then (i) the shuffling of the small exons to make new proteins; (ii) the sliding and drift of introns to change the exon boundaries and to add new amino acids; and (iii) the loss of introns and, hence, the fusion of small exons to make larger ones. The shuffling of exons requires recombination at the DNA level, both legitimate and illegitimate, for which the introns serve as hotspots. The primary mechanism for the loss of introns is retroposition: a spliced message is copied back into DNA by a reverse transcriptase and then inserted into the genome. Such a cDNA copy can remove introns from the original gene by gene conversion. Alternatively, the reinsertion of the cDNA copy into the genome creates a pseudogene, if the region is not transcribed, or creates a complex C-terminal exon, if the insertion is into an intron in a functional gene. A complete example of this last process was worked out for the gene *Jingwei* in *Drosophila* where the normal four exons of ADH are fused into one and attached to other exons, making a new functional gene (Long & Longley 1993). Several rounds of exon fusions would be required in the course of evolution to produce today's exon spectrum, peaked at length of 35–40 residues, and extending out occasionally to thousands of residues in length.

The critical predictions of the Exon Theory of Genes are that the first exons were small and that the original exons should correspond to modules, compact regions of polypeptide chain, which should in turn correspond to folding elements in protein structure.

## The RNA world

The discovery that RNA molecules could be enzymes (ribozymes) with the potential to catalyse many reactions, including self-splicing, RNA cleavage and RNA polymerization (reviewed by Cech 1990) solved the chicken/egg paradox of the origin of life because now one single molecular type could (i) store information and (ii) catalyse a broad number of reactions and be

the target of natural selection. An RNA world, consisting of RNA genomes and RNA enzymes, became possible (Gilbert 1986).

In this picture, RNA is the carrier of genetic information, but introns, as ribozymes, promote the shuffling of primordial RNA exons that are RNA folding or functional units, and provide the benefits of genetic recombination. With ribozymes developing into the full complement of enzymes for primitive life, the intron/exon structure of RNA genes creates complex ribozymes from simple and reusable parts.

After the development of a translation system that uses ribozymes as ribosomes and activating enzymes (Piccirilli *et al*. 1992), one expects the introduction of amino acids essentially one at a time and, hence, first the appearance of oligopeptide/ribozyme complexes with enhanced enzymatic activities and then the appearance of true protein enzymes. This picture leads to an RNA/protein world in which the RNA genetic information still functions with an intron/exon structure, the exons now related to the protein products of the messenger RNAs. In this period, new and efficient splicing mechanisms could appear using proteins as co-factors along with RNA molecules. The original splicing mechanism, at the ribozyme level, would be the enzymologies responsible for splicing in group I and group II introns.

DNA emerges after the development of a reverse transcriptase and the enzymes, to make the deoxy-ribonucleotide precursors from the ribonucleotide triphosphates. The role of DNA is to be a long, efficient, store of genes, because error correction mechanisms permit long genomes. Such DNA genomes would conserve the intron/exon patterns copied from the RNA genetic information. This produces a DNA/RNA/protein period of evolution, a progenote with an intron/exon structure. This primitive cell lineage ultimately split into three different lines: eukaryotes, eubacteria and archaebacteria. The last two cell lineages reduced their genomes, losing introns due to the pressures involved in specializing for rapid DNA replication. The eukaryotes, however, evolved a more efficient splicing mechanism, the spliceosome, and the spliceosomal-dependent introns presumably evolved from group II introns.

A frequent misunderstanding is the idea that the assumptions of the exon theory of genes are restricted to spliceosomal introns. Spliceosomal introns, in terms of mechanism, probably emerged after the appearance of eukaryotes. However, the exon theory of genes suggests that these introns were present in the progenote in a more primitive form, as group II introns.

## On the role of exon shuffling in evolution

The accumulation of sequence data over the last 15 years has shown that exon shuffling by intron-mediated recombination events has occurred with high frequency during recent evolution. In the middle 1980s the classic example of exon shuffling appeared, that of the LDL receptor (Sudhof *et al*. 1985a) and EGF precursor (Sudhof *et al*. 1985b). To that, one might add the structure of the collagen genes, the structure of the immunoglobulin superfamily and the blood factors (such as IX and X), and the extracellular matrix proteins (Hynes 1990). Furthermore, a specific example of the acquisition of a new function after an exon shuffling event is the gelatinases in the matrix metalloproteinase family of enzymes. In these extracellular matrix degrading proteases, which share at least 50% homology, there is often a clear distinction between catalytic and substrate-binding activities (Matrisian 1992). In the gelatinases, the acquisition of a gelatin-binding domain from fibronectin through an exon shuffling event directed the enzyme to cleave gelatin (Wilhelm *et al*. 1989). Recently, we found a clear example of exon shuffling between the nuclearly encoded plant genes, glyceraldehyde-3-phosphate dehydrogenase (cytosolic GAPDH or GapC) and cytochrome C1 precursor in the potato (Long *et al*. 1996). Three amino-terminal exons of GapC have been donated to cytochrome C1, where, in a new protein environment, they serve as a source of the mitochondrial targeting function. The intron/exon structures in the shuffled region of the two genes are very similar, both shuffled sequences contain the same number of introns with very close or identical positions and phases, showing that the shuffling was a result of shuffling at DNA sequence level rather than a consequence of retroposition.

## Intron phase correlations and exon shuffling

Intron phase, the position of the intron within a codon as defined by Sharp (1981), is a conserved evolutionary character of the intron/exon structure for which the two competing theories generate different predictions. Since the insertion of introns into previously uninterrupted genes, as the 'introns-late' theory advocates, does not bring any change to its protein structure and since they are removed before the message is translated, the position of the introns are unrelated to the reading frame. Thus, the simplest form of this theory predicts a random distribution for intron phases: (i) the frequency

of three phases should be equal; and (ii) there should be no phase correlation within genes. On the contrary, the 'introns-early' theory predicts a nonrandom distribution of the intron phase: (i) the introns should tend to be in the same phase (to preserve reading frame during exon shuffling) and probably phase zero since primordial genes (exons) would be functionally independent units (modules); and (ii) the phases of two or more introns along a gene should be correlated as a consequence of exon shuffling. These predictions formed a foundation for a test for exon shuffling.

Long *et al.* (1995a) constructed an exon database containing 13 042 exons and 11 117 introns from 1925 independent or nearly independent genes. The three intron phases do not appear equally, as a previous survey (Fedorov *et al.* 1992) had shown. Of the 11 117 introns, 48% are phase zero, 30%, phase one, and 22% phase two. These results are consistent with the predictions of the Exon Theory of Genes, however, not a strong argument against the introns-late theory because there might be preferences for intron insertion into special sequences, which might in turn be correlated with reading frame. Such a feature might be a nonrandom distribution of proto-splice sites, a hypothetical insertion site for introns advocated by Dibb & Newman (1989). However, Stephens & Schneider (1992) compared more than 1800 human introns and found that most of the information for splicing is confined within introns and that there is no significant AG|GT bias but possibly a small AG|G conservation in exon sequences. If one uses the AG|G distribution from dicodon frequency tables (Gary Stormo, University of Colorado, personal communication), there is no consistent predicted intron phase distribution. Thus there is no evidence that the intron phase distribution is a result of a nonrandom distribution of proto-splice sites.

The correlations of intron phases are a stronger argument for exon shuffling. There are two types of intron combinations across exons: symmetric, the flanking introns having the same phases, or asymmetric. On an insertional theory, the intron phase associations should equally be any of the nine combinations because insertion at separate sites along the DNA should be uncorrelated. However, in exon shuffling, symmetric exons are an essential requirement, because the addition of an exon into an intron should not change the reading frame. In fact, the frequencies of all three symmetric exons are significantly higher than expectation (based on random association using the observed biased phases) and the frequencies of all asymmetric exons are significantly smaller than expectation. Specifically, there is 30% excess of (1,1) exons over expectation, greater than the 10% excess of (0,0) exons (although the absolute number of (0,0) exons is greater than the number of (1,1) exons). This excess of symmetric exons (and pairs, triples and quadruples of exons) is very significant ($\chi^2$ values of 200). This excess of symmetric exons strongly suggests that there is shuffling in eukaryotes. The actual value of the excess can be turned into an estimate for how many exons have to be involved and, hence, into a minimal estimate for how many exons were involved in exon shuffling. Long *et al.* (1995a) concluded that at least 19% of the database had to be involved in exon shuffling, thus exon shuffling is a very important component of evolution.

## How many exons?

A few years ago, Dorit *et al.* (1990) tried to calculate how many different exons were used in the assembly of current genes, based on the frequency with which exons are reused during evolution. Their estimates suggested that from 1000 to 7000 exons were needed to construct the current pool of genes. Although these calculations were criticized (Doolittle 1991; Patthy 1991), we still accept the general trend of the argument (Dorit *et al.* 1991). This estimate pertains to the roughly 30–40 residue-long exons characteristic of today's spectrum. A 300–400 residue-long gene is made up of 10 such units. There are, in principle, on the order of $10^{40}$ such structures, if there are only $10^4$ exons—a more than adequate diversity. In fact, evolutionary history limits diversity far more profoundly, to about $10^5$ possibilities from $10^4$ exons (Dorit *et al.* 1990). The reason for this is that evolution can search linearly, by adding exons one at a time, rather than combinatorially.

## Introns early or late?

All these arguments for exon shuffling apply to the eukaryotes. What is the connection to early introns? The crucial questions are: were introns present in the progenote? Were exons shuffled to make ancient genes? To address these questions one must study ancient genes, genes conserved between prokaryotes and eukaryotes, whose origin would be the progenote.

In general, there are three types of arguments. The first type are arguments based on phylogeny and phylogenetic classification and direct comparison of genes and gene structures. The second type of argument studies intron phase correlations and tries to show that

one can infer that there was exon shuffling in ancient genes in the progenote. The third type of argument tries to show that introns, as a feature at the DNA level, are associated with aspects of the three-dimensional structure of proteins in such a way as to delineate aspects of tertiary structure. The simplest explanation for this would be that introns shuffled elements of protein structure to make up the first genes.

## Phylogenetic arguments

The first suggestion that the introns might be very ancient, in 1978, proposed a phylogenetic puzzle which has been reiterated strongly recently (Palmer & Logsdon 1991; Kwiatowski *et al*. 1995; Rogers 1985). Broadly speaking, prokaryotes do not have spliceosomal introns (although they do have a truly ancient intron in their tRNA (Xu 1990; Kuhsel 1990)). Simpler (sometimes called primitive) eukaryotes have few or no introns, and only the most complex eukaryotes appear to have a full display of introns. This simple description has led those that take the classical view of evolution, as proceeding from the simple to the complex, from prokaryotes to simple eukaryotic cells to complex eukaryotes, as clear evidence that the introns must be a late feature (Palmer & Logsdon (1991)). In contrast, the theory of ancient introns has always required a post-modern view of evolution: a view that turns the classical attitude on its head and suggests that the first genomes had a complicated intron/exon structure, and similarly, the first cells had a complicated and sloppy ribosome structure, and that the bacteria and many 'simple' organisms are highly evolved, streamlined structures which specialized for rapid DNA replication, small DNA content, rapid protein synthesis, and so forth. This view of evolution holds that the bacteria, as constituted today, are the end-points of a long evolutionary path, involving many more generations than the nuclei of today's eukaryotic cells. One general support of this view is that the spliceosome is an RNA-based enzymology and structure, rather akin to the ribosome, which resembles a product of early evolution.

These two views of evolution involve diametrically opposed ways of looking at a number of phenomenon and are not resolvable on general principles. The classical view looks upon the addition of introns as easy, as a simple derived character, and their loss as difficult. The post-modern view looks upon the addition of introns as a difficult and rare feature and their loss as an easy and common feature. The classical view looks upon the matching of intron positions in different genes as being a statistical accident and the presence of introns at positions separated by one or two bases over in homologous genes as clear evidence for separate acts of addition. The post-modern view looks upon the exact matching of position as evidence of common ancestry and the close matching of position as suggestive of common ancestry followed by movement, such movement created by mutation at the DNA sequence level or excision and reintroduction at the RNA level.

## Intron loss or gain?

There are many example of intron loss, and only a few clear examples of intron gain. The first example of intron loss in the case of introns is the pair of insulin genes in the rat in which one intron was shown to be lost when the gene was duplicated (Perler *et al*. 1980). Another example is the superoxide dismutase genes in the insects (Kwiatowski *et al*. 1992) in which a phylogenetic pattern clearly shows the loss of an intron, a third example is the loss of introns in *Arabidopsis* as opposed to other plants (Chang & Meyerowitz 1986), and we remind our readers of the retroposition example documented in Long & Langley (1993). There has been a recent assertion that some of the introns in triosephosphate isomerase might be examples of gain, but this has not been worked out in a tight phylogenetic fashion (Logsdon *et al*. 1995; Palmer & Logsdon 1991). The clearest examples of intron gain are the insertion of introns into the U6 RNA of some fungi (Tani & Ohshima 1989, 1991; Takahashi *et al*. 1993; Tani *et al*. 1995) in which the pattern of introns supports the argument that the mechanism is actually one of reverse splicing of a newly spliced out spliceosomal intron back into the U6 RNA (involved in the catalysis of that splicing), followed by retroposition and gene conversion events.

The sharpest example of the loss of introns is *Saccharomyces cerevisiae.* This organism was traditionally an example of a simple eukaryote that does not have introns. Fink (1987) developed the argument that the loss of introns in cerevisiae, compared to the other fungi, might be due to a runaway reverse transcriptase which led to the reintroduction of cDNA copies of genes into the genome. This process not only would eliminate introns, but, depending on the enzymology, could eliminate them in a polarized fashion from the 3′ end.

The simplest model for the addition of introns would be a transposable element that bears splicing signals on its ends and, hence, can insert into DNA in an invisible

fashion. Such an element has not yet been seen, and the simplest prediction of such models, that one should see similarity in intron sequences reflective of selfish DNA, has not been observed. Giroux *et al.* (1994) documented the case of a mutagenic plant transposon in which further mutation had led to the development of an intron.

## Arguments from coincidence of position

Consider when introns might have arisen:

(i) The introns could be very recent, currently being added, and in this case we would expect no exon shuffling;

(ii) The introns might arise only since the metazoan radiation (these are theories in which the complicated eukaryotes have introns but there is not any intron structure in single-celled organisms);

(iii) Introns might arise before the Precambrian, in the ancestor of the plants, animals and fungi;

(iv) The introns might arise before any eukaryotic cell, but after the split away from the prokaryotes; and

(v) The introns might be in the progenote, in the last common ancestor of eubacteria, archaebacteria, and eukaryotes.

The issue is basically, are some introns very old? The introns-late view is that no introns are old, that introns fall into one of the classes (i) to (iv).

The most striking test of introns being currently added would be the finding of introns with identical sequences in different genes as a consequence of that addition. Such introns have never been found. The strongest argument that introns arose before the Precambrian and in the ancestor of plants and animals is the correlation of introns in plant and animal genes. One of those correlations is in the triosephosphate isomerase gene. By comparing the gene in chicken and in maize, Marchionni & Gilbert (1986) observed the striking feature that five out of the six introns in chicken had identical positions in maize. Three introns in actin (Shah *et al.* 1983) are shared between plants and animals. In the largest subunit of the RNA polymerase II, there are five introns in identical position between *Arabidopsis* and the vertebrates (Nawrath *et al.* 1990; Accession numbers M12130, M14101 and X52954). Pardo & Serano (1989) reported that in the plasma membrane $H^+$-ATPase gene all 15 introns of the *Arabidopsis* gene are located in positions equivalent to the intron positions of the animal ATPases. In the ubiquitin conjugating enzymes in *Arabidopsis*, four introns in the family corresponding to UBC1 are located in identical positions to the introns in the corresponding mouse gene, the fifth intron in the rat gene is located identically to an intron in a second family of ubiquitin conjugating enzymes in *Arabidopsis* (Sullivan *et al.* 1994; Wing & Bonville 1994). This strong coincidence suggested that the introns were present in the single-cell ancestor that diverged between the plants and the animals about a billion years ago.

For some eukaryotic gene products localized in the mitochondria or the chloroplast whose genes are located in the nucleus, one can compare introns between an organelle and a cytoplasmic version. Amino acid sequence can show that the organelle and cytoplasmic versions represent sequences that trace back to ancient duplications antedating the divergence between the prokaryotes and the eukaryotes. If introns have identical positions in such genes, they themselves are candidates to antedate the prokaryotic/eukaryotic divergence (Cornish-Bowden 1985; Shih *et al.* 1988; Quigley *et al.* 1988). The GAPDH (glyceraldehyde-3-phosphate dehydrogenase) gene provides an excellent example. The chloroplast and cytosolic GAPDH genes are two related nuclear genes in the eukaryotes derived from ancient duplication events. Kersanach *et al.* (1994) pointed out that five introns in these genes are in identical positions, which suggested that these introns were likely to have existed in the common ancestor of eukaryotes and prokaryotes. Further examples are the cytosolic and mitochondrial malate dehydrogenase (MDH) genes where two introns are located in the same positions (Setoyama *et al.* 1988) and the cytosolic and mitochondrial aspartate aminotransferase isozymes (AspAT) genes where five introns are in identical positions (Obaru *et al.* 1988).

However, not all ancient duplication events show duplicated introns. The elongation factors 1 and 2 of protein synthesis EF-1 and EF-2 originated from a duplication event predating the progenote (Iwabe *et al.* 1990), but there is no matching of introns. Similarly between MDH and LDH (malate and lactate dehydrogenase), or the two halves of phosphofructokinase there is as yet no matching. While the reason for this might be the drift of homologous introns or a gene conversion loss of different introns in different lineages, the current state of knowledge about these genes does not yet support ancestral introns. However, carbamoylphosphate synthetase represents an ancient duplication in which homologous positions have been seen (Schofield 1993; van den Hoff 1995).

The general argument about the coincidence of introns in sequences which are very different at the nucleic acid level but show an ancient similarity at the protein level is that this matching is evidence that some

introns in the eukaryotic genes are very old, going back to the progenote. The general attack on this view is the argument that these introns might all have been insertions into previously unbroken genes and that the matching is only statistical coincidence. Logsdon & Palmer (1994) and Stoltzfus *et al*. (1994) criticized the Kersanach paper by suggesting an intron insertion model in which the number of possible sites for introns is limited and, hence, it would not be surprising that coincidences occur, while Kersanach *et al*. (1994) had assumed that the number of sites that introns could insert is equivalent to the number of residues in the gene, which produces a very low probability.

The major difficulties in trying to trace ancestral introns or to attribute introns to an ancestral gene are: (i) that introns may be lost and, hence, that exons may become complex, (ii) that introns may move in position, or (iii) that introns may be added at some later stage of evolution. How clearly can we see a remnant of the original structure under the influence of these forces? Only the collection of more and more data will enable some of these questions to be resolved.

Palmer & Logsdon (1991) argued that the phylogenetic distribution of introns showed that introns are recent acquisitions of the eukaryotic branch. Based on the fact that most of the protist groups lack introns, they discussed a scenario where introns were acquired by what they called AFP lineages (animals, fungi and plants). The phylogeny shown by Palmer & Logsdon, constructed using rRNA sequences, has as a major feature that most of the protist lineages are unrelated, being the results of independent branchings. Thus, Palmer & Logsdon argued that the only way to explain the phylogenetic distribution of introns under an introns-early view was the parallel loss of introns down all the protist branches (they argued that the loss of introns was very difficult; the introns-early view is that the loss of introns is very easy, and that they survive only where there is no selection pressure). Protist phylogeny, however, is controversial (Hashimoto *et al*. 1994; Kumar & Rzhetsky 1996), and the branching order of the groups is not clear, with the exception of an earliest branching for *Giardia*.

In 1991, no introns were known for *Giardia*, *Vairimorpha*, *Naegleria*, *Entamoeba*, *Trypanosoma* and other protists. Today's GENBANK (Version 93), has spliceosomal intron-containing entries for *Entamoeba* (Lohia & Samuelson 1993) and *Naegleria* (Remillard *et al*. 1995). One intron in the cdc2 gene in *Entamoeba* is conserved in position with a yeast gene suggesting a very ancient origin (Lohia & Samuelson 1993). Two introns present in calcineurin B of *Naegleria* seem to separate functional domains of the protein (Remillard *et al*. 1995). Furthermore, there are spliceosomal introns in *Euglena* (Henze *et al*. 1995) and *trans*-splicing in *Trypanosoma* (Agabian 1990). All these introns show that the splicing mechanism exists in protists, which challenges the notion that all early protists are intronless.

Introns have not yet been reported for the early branching protists, *Giardia* and *Vairimorpha*. These species may have suffered a 'streamlining' process to reduce their genome, just as have the bacteria. What might have been the pressure on bacteria and early protists to reduce their genomes? One possibility is competition with mitochondria-containing eukaryotic cells. The symbiotic event that gave rise to mitochondria and chloroplasts gave a selective advantage to the cells containing these organelles. The intronless *Giardia* and *Vairimorpha*, amitochondrial organisms, along with the prokaryotes, may have used smaller genomes and a consequently faster replication rate to compete with the organelle-containing eukaryotes and leave descendants.

## Arguments from phase correlations in ancient genes

Earlier we commented that the excess of symmetric exons shows that exon shuffling is one of the major processes in creating gene diversity. However, exon shuffling could be viewed as a recent phenomenon, as has been claimed by Patthy (1991), Stoltzfus *et al*. (1994), and Palmer & Logsdon (1991). Long *et al*. (1995a) extended the phase correlation analysis to a database of ancient genes. By defining regions of eukaryotic genes that are homologous and colinear to prokaryotic genes, they identified a set of introns that must all be inserted on any introns-late model. However, there are excesses of symmetric exons, pairs, triples and quadruples of exons in the database. This result suggests that shuffling events took place before the divergence of the prokaryotes and eukaryotes. No shuffling in these ancient regions could occur after the divergence because that would break the agreement in sequence between the lineages.

## Arguments from exon/module correlations

In the early 1980s, Gō suggested that the exon products might correspond to modules, by which she meant

a region of continuous polypeptide chain that folds up into a compact unit circumscribable by a sphere 28 Å in diameter (Gō 1981). This feature of compactness could be represented on a two-dimensional map, the Gō plot, in which the distance between the alpha carbons of each pair of residues is plotted, with distances greater or equal to 28 Å coloured black. All the distances within an exon lie within a triangle along the diagonal of such a plot. As a consequence, if an exon is compact, no black spots will appear within the corresponding triangle: while more complex exons will have a black region within their triangle. Gō argued that for globin the central exon encoded two modules, and she predicted the existence of an intron in an ancestral globin that would have been located between these modules (Gō 1981). Soon after her prediction was made, a third intron was discovered in leghaemoglobin that disrupted the central exon exactly between these modules (Jensen *et al*. 1981). More recently, similar introns have been discovered in other globin genes (Moens *et al*. 1992). Nonetheless, Stoltzfus & Doolittle (1993) and Pohajdak & Dixon (1993) have argued that although the two original introns in globin are fixed in position, the central intron seems to be variable in position and, hence, that all these are different insertions. However, the central region of the very distant globins that are compared is such that the sequence alignment is not well defined. This is a deep problem with such evolutionary comparisons. An amino acid alignment is actually a hypothesis about a pattern of mutation and evolution, not a fact. Alternative methods of alignment—for example profile analysis which attempts to take into account the relationship of the amino acids with respect to the three-dimensional structure of the protein (Luthy *et al*. 1991)—give different positions for the introns in the central region. It becomes a question of belief as to which feature is primary: intron position or amino acid matching.

A similar analysis of the relation of exon structure to modules was made for the triosephosphate isomerase (TPI) gene. Like other genes with glycolytic metabolism, TPI enzyme represents an ancient conserved sequence, with more than 40% amino acid similarity between human and bacterial enzymes. On examining the Gō plot for TPI from chicken, Straus & Gilbert (1985) commented that the exons would represent modules in Mitiko Gō's sense if two of the exons (numbers 4 and 5 in chicken) would be found to be broken up in other organisms. When McKnight *et al*. (1986) analysed the gene in *Aspergillus*, two new introns interrupted these two complex exons. Gilbert *et al*. (1986) predicted that there should be a further ancestral intron in the TPI gene that would break up one remaining complex exon. A few years ago Tittiger *et al*. (1993) found this missing intron in the mosquito *Culex tarsalis*. This represents a series of successful theoretical predictions, that if one attributes all of the introns to an ancestral gene those introns break up the ancestral gene into modules—compact segments in Mitiko Gō's sense. However, the comparison of introns in an 'ancestral' TPI gene involves notions of sliding and drift. The comparison between vertebrate, plant and *Aspergillus* introns involves five introns at identical positions in plants and animals and one in an identical position between fungi, plants and animals. But, three introns are at different positions: two introns in fungi that are close to the positions of introns in plants, one one base over and one four bases over, and one intron in plants that is three amino acids (nine bases over) from an intron in chicken. The argument that these three are ancestral introns requires both the assumptions that each of these pairs of introns had a common origin and that there was drift over evolutionary time. The school of belief that introns were added, looks upon each of these as examples of separate addition.

One might wonder about the strength of these arguments about compact modules, because this theory is not self-limiting: a module could be broken up into smaller modules by further introns. The theory seems to offer no guidance as to what such submodules might look like. More and more introns make the exons smaller and smaller, and, in terms of a Gō plot, the exons will ultimately fall into compact regions along the diagonal. Has one said any more than this rather trivial observation? Gilbert & Glynias (1993) tried to answer this question by examining whether the 11 hypothesized ancestral introns of TPI fit the Gō plot in an unusual fashion. They constructed statistical tests by comparing the observed intron structure with the one that would appear had the 11 introns been inserted in the gene in a random fashion. If introns had been purely random, the observed set of positions turns out to fit the Gō plot criterion of modularity better than do 99.98% of random variants. However, exon sizes are clearly nonrandom, and insertion models must make special provision to account for this. If one compares the observed set of 11 introns to alternative sets of 11 introns that have the same distribution of exon sizes, the observed set was better than the constrained random sets only about 95% of the time. Gilbert & Glynias went on to identify four tests of exon structure, one being a module $\leq 28$ Å, a second being the exons containing more of the close contacts, $\leq 12$ Å, a third being the exons containing more of the buried surface along the

backbone, and a fourth being the exons containing more of the backbone hydrogen bonds. By combining all of these tests they could show that the set of 11 introns for TPI were better than 99.4% of the constrained random comparisons. Thus they concluded that one could construct a statistical argument that the intron positions in TPI were unusually related to three-dimensional structure of the protein.

## The attack on an exon–protein structure correlation

Stoltzfus *et al*. (1994) attacked the Exon Theory of Genes and the notion that introns had a correlation with the three-dimensional structure of proteins. They argue that they could not find a correlation between exons and elements of three-dimensional structure, and they thus concluded 'that the Exon Theory of Genes is untenable.' However, much of their work did not test the actual predictions of the Exon Theory but a variety of straw men. Their first analysis was to examine whether or not exons corresponded to alpha helices or to beta sheets: they tested whether introns would lie at the boundaries of such secondary structure elements. They found no correlation. However, Gilbert (1979) had already made the argument that introns often break up alpha helices and that the exons in general would correspond to a tertiary structure elements, such as bends in the protein, rather than to secondary elements. The second major approach in the Stoltzfus *et al*. (1994) paper was to test the notion of centrality: whether or not the introns lie close to a centre of mass of the protein structure. Although such a notion of centrality was one of the ways of looking at Gō's original plot in 1981, by 1985 Straus & Gilbert (1985) had pointed out that centrality was not a valid notion for TPI. Thus it is not surprising that this is not a valid notion in general. Lastly, Stoltzfus *et al*. examined the intron positions in four proteins: TPI, pyruvate kinase, globin and alcohol dehydrogenase. They treated all of the introns observed in any gene as ancestral intron and argued that with the possible exception of TPI, they could not see any correlation between intron positions and a modular structure. They concluded from this failure that there was no correlation.

Recently, Gō & Noguti (1995) tested the same set of four proteins using a different definition of modules and reached the opposite conclusion: they concluded that the distribution of introns did approximate the boundaries of modules at about the $P = 0.05$ level.

## Correlation between introns and modules

Recent work by our group (de Souza, Long, Schoenbach & Gilbert, Harvard University, to be published) has provided a strong statistical argument that introns are related to the tertiary structure of proteins. We defined modules as regions of polypeptide chain that could be circumscribed by a sphere of a given diameter, and then wrote a computer program that will define regions along the protein that are sites at which the boundaries of modules could lie. This defines a set of 'linker regions', such that if an intron were to be placed in each of the linker regions, the protein would be dissected into modules of the specified size. About one-third of the protein is defined to be in such linker regions, but this definition provides a clear statistical test of whether the introns respect the boundaries of modules: is there an excess of introns in these linker regions? or do the introns distribute randomly across the gene? By examining a set of 28 proteins (25 ancient proteins and three others: ras, actin and globin), which contain a total of 579 intron positions, counting every intron position observed as an independent event, we found a statistically significant excess of introns in the linker regions for modules 28 Å in diameter, the standard size used before by Gō and by us. The correlation is significant with a $P = 0.007$. Because the linker regions are predicted by a computer program that will determine their pattern for any diameter of the modules, one can vary that diameter to see how the excess of introns in the linker regions varies. In a plot of the statistical significance of this excess there are four peaks of statistical significance corresponding to modules of about 15, 22, 27 and 33 Å in diameter. The peaks at 15 and 22 Å reach probabilities less than 0.01, the peaks at 27 and 33 Å reach probabilities less than 0.001. This demonstrates that there is a extremely significant correlation of introns with the boundaries of compact size elements in the tertiary structure of proteins. One can understand these elements more simply if one analyses the pattern not in terms of modules fitting a certain diameter but in terms of the average lengths between linker regions, i.e. the average length of the polypeptide chain in the modules or the average length of the hypothetical exons. The four peaks turn out to correspond to polypeptide lengths of 8, 16, 24 and 32 amino acids. Thus there is a significant correlation of introns to the boundaries of tertiary structural elements in the protein corresponding to original exons 8, 16, 24 and 32 amino acids in size. This finding provides strong statistical support for the Exon Theory of Genes. It is exactly the major prediction of that theory, and so

supports the attitude that the original genes were constructed by exon shuffling from short pieces of polypeptide chain.

One can speculate that the shortest pieces of polypeptide chain, those roughly eight amino acids long, probably correspond to beta turns in the protein structure, the longer regions of about 16 residues correspond to a part of an alpha helix plus a beta turn. There is literature that shows that specific amino acid sequences of these lengths, identified from proteins, often have such structures in solution. Mitiko Gō (personal communication) has used a slightly different definition of modules to arrive at a similar conclusion, that introns correlate on the module boundaries. Her definition of modules involves a centripetal scan along the three-dimensional structure of the protein to identify minima in a sliding function that is the sum of a set of interresidue distances squared. Our definition of a module is a compact region of the polypeptide chain circumscribable by a sphere. Nonetheless, both of these approaches identify the correlation of introns with tertiary structure elements of the protein, and this correlation becomes more and more significant as one studies more and more protein genes.

Are there alternative interpretations of these findings, other than the model of early introns? One is the suggestion that there might be some unusual amino acid composition or use at the boundaries of modules so that there might in turn be some unusual compositional bias in the DNA or even special sequences, which could in turn be used as targets for intron insertion. (Such a suggestion might be that module boundaries and introns occur preferentially on the surface of proteins (Craik *et al*. 1982) and that this property might serve as a bases for such a compositional bias. However, introns lie on the surface of proteins just in proportion to the fraction of amino acid residues that lie on the surface of proteins rather than in excess.) One specific counter argument is that de Souza *et al*. (1996; to be published) examined a set of 24 putatively ancient introns, occurring at the same or similar positions in three of the four groups: plants, fungi, invertebrates and vertebrates, found in their collection of 579 introns and could show that these had a 10-fold more significant association with 28 Å modules than did the total population. This is not what one would expect if there had been some DNA bias underlying the result but is consonant with the signal being from a subset of ancient introns.

Other arguments are based on natural selection. One such is that intron insertion might be mutagenic, so that introns could insert only where amino acid changes could be accommodated, and hence introns would be found in regions in which amino acid sequence is not conserved, such as loops or between modules. However, if one examines the distribution of introns, they occur in highly conserved regions as well as unconserved regions in proportion to the extent of these regions in the sequence. In fact, Brändén *et al*. (1987) argued for the alpha-beta-barrel proteins that intron sliding had often been used to add amino acids to form part of the active site and hence that these introns were in extremely conserved regions.

Another argument is that positive selection has preserved added introns in inter-module positions. When this argument is fleshed out, it is the proposal that introns added randomly to an ancient protein and exons that happened to correspond to modules were then shuffled and used to create new proteins on which natural selection could act. However, even though natural selection could fix in the population the novel exon in the new protein, it cannot fix in the population the corresponding structure of the donor molecule, since that gene would be genetically unlinked from the recipient one. Hence these arguments fail.

## Conclusion

Overall, many arguments support the idea that some introns are very old and are remnants of the assembly of the first genes. In each of these arguments, the coincidences of intron positions, the correlation of intron phases, and the correlation of introns with the tertiary structure of proteins, are unified in a simple way by the Exon Theory of Genes. To encompass them in an introns-late view requires a special series of arguments involving special entry sequences and unusual connections between DNA sequence and protein structure. The continuing flow of information about gene structure as complex eukaryotes are sequenced by the genome project will show us more and more about the origin of evolution and about the most ancient genes.

## Acknowledgements

We thank the National Institutes of Health, grant no. GM 37997, for support. S. J. de Souza was supported by Conselho Nacional de Desenvolvimento Cientifico e Technologico (CNPq-Brasil) and the PEW-Latin American Fellows Program.

## References


Agabian, N. (1990) *Trans* splicing in nuclear pre-mRNAs. *Cell* **61**, 1157–1160.

Blake, C.C.F. (1978) Do genes-in-pieces imply proteins-in-pieces? *Nature* **273**, 267.

Brändén, C.-I., Schneider, G., Lindqvist, Y., Andersson, I., Knight, S. & Lorimer, G. (1987) Structural and evolutionary aspects of the key enzymes in photorespiration; RuBisCO and glycolate oxidase. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 491–498.

Cavalier-Smith, T. (1978) Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA *C*-value paradox. *J. Cell Sci.* **34**, 247–278.

Cavalier-Smith, T. (1985) Selfish DNA and the origin of Introns. *Nature* **315**, 283–284.

Cavalier-Smith, T. (1991) Intron phylogeny: A new hypothesis. *TIG* **7**, 145–148.

Cech, T.R. (1990) Self-splicing of group I introns. *Annu. Rev. Biochem.* **59**, 543–568.

Chang, C. & Meyerowitz, E.M. (1986) Molecular cloning and DNA sequence of the *Arabidopsis* thaliana alcohol dehydrogenase gene. *Proc. Natl. Acad. Sci. USA* **83**, 1408–1412.

Cornish-Bowden, A. (1985) Are introns structural elements or evolutionary debris? *Nature* **313**, 434–435.

Craik, C.S., Sprang, S., Fletterick, R. & Rutter, W.J. (1982) Intron-exon splice junctions map at protein surfaces. *Nature* **299**, 180–182.

Darnell, J.E., Jr. (1978) Implications of RNA–RNA splicing in evolution of eukaryotic cells. *Science* **202**, 1257–1260.

Dibb, N.J. & Newman, A.J. (1989) Evidence that introns arose at proto-splice sites. *EMBO J.* **8**, 2015–2021.

Doolittle, R. (1991) Counting and discounting the universe of exons. *Science* **253**, 677–679.

Doolittle, W.F. (1978) Genes in pieces: Were they ever together? *Nature* **272**, 581–582.

Doolittle, W.F. & Sapienza, C. (1980) Selfish genes, the phenotypic paradigm and genomic evolution. *Nature* **284**, 601–603.

Dorit, R.L., Schoenbach, L. & Gilbert, W. (1990) How big is the universe of exons? *Science* **250**, 1377–1382.

Dorit, R.L., Schoenbach, L. & Gilbert, W. (1991) Response to 'Counting and discounting the universe of exons,' by Doolittle, R. [*Science* (1991) **253,** 677–679]. *Science* **253**, 679–680.

Fedorov, A., Suboch, G., Bujakov, M. & Fedorova, L. (1992) Analysis of nonuniformity in intron phase distribution. *Nucl. Acids Res.* **20**(10), 2553–2557.

Fink, G.R. (1987) Pseudogenes in yeast? *Cell* **49**, 5–6.

Gilbert, W. (1978) Why genes in pieces? *Nature* **271**, 501.

Gilbert, W. (1979) Introns and exons: Playgrounds of evolution. In: *Eucaryotic Gene Regulation* (eds R. Axel, T. Maniatis & C. F. Fox). New York: Academic Press.

Gilbert, W., Marchionni, M. & McKnight, G. (1986) On the antiquity of introns. *Cell* **46**, 151–153.

Gilbert, W. (1986) The RNA world. *Nature* **319**, 618.

Gilbert, W. (1987) The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901–905.

Gilbert, W. & Glynias, M. (1993) On the ancient nature of introns. *Gene* **135**, 137–144.

Giroux, M.J., Clancy, M., Baier, J., Ingham, L., McCarty, D. & Hannah, L.C. (1994) *De novo* synthesis of an intron by the maize transposable element *Dissociation. Proc. Natl. Acad. Sci. USA* **91**, 12150–12154.

Gō, M. (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291**, 90–93.

Gō, M. & Noguti, T. (1995) Putative origins of introns deduced from protein anatomy. In: *Tracing Biological Evolution in Protein and Gene Structures* (eds M. Gō & P. Schimmel). Amsterdam: Elsevier.

Hashimoto, T., Nakamura, Y., Nakamura, F., *et al.* (1994) Protein phylogeny gives a robust estimation of early divergence of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia. Mol. Biol. Evol.* **11**, 65–71.

Henze, K., Badr, A., Wettern, M., Cerff, R. & Martin, W. (1995) A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. *Proc. Natl. Acad. Sci. USA* **92**, 9122–9126.

Hynes, R.O., ed. (1990) *Fibronectins*. New York: Springer-Verlag.

Iwabe, N., Kuma, K., Kishino, H., Hasegawa, M. & Miyata, T. (1990) Compartmentalized isozyme genes and the origin of intron. *J. Mol. Evol.* **31**, 205–210.

Jensen, E.O., Paludan, K., Hyldig-Nielsen, J.J., Jorgensen, P. & Marcker, K.A. (1981) The structure of a chromosomal leghaemoglobin gene from soybean. *Nature* **291**, 677–679.

Kersanach, R., Brinkmann, H., Liaud, M.-F., Zhang, D.-X., Martin, W. & Cerff, R. (1994) Five identical intron positions in ancient duplicated genes of eubacterial origin. *Nature* **367**, 387–389.

Kuhsel, M.F., Strickland, R. & Palmer, J.D. (1990) An ancient group I intron shared by eubacteria and chloroplasts. *Science* **250**, 1570–1573.

Kwiatowski, J., Krawczyk, M., Kornacki, M., Bailey, K. & Ayala, F.J. (1995) Evidence against the exon theory of genes derived from the triose-phosphate isomerase gene. *Proc. Natl. Acad. Sci. USA* **92**, 8503–8506.

Kwiatowski, J., Skarecky, D. & Ayala, F.J. (1992) Structure and sequence of the *Cu,Zn Sod* gene in the Mediterranean fruit fly, *Ceratitis capitata*: Intron insertion/deletion and evolution of the gene. *Mol. Phylogenet. Evol.* **1**, 72–82.

Kumar, S. & Rzhetsky, A. (1996) Evolutionary relationships of eukaryotic kingdoms. *J. Mol. Evol.* **42**, 183–193.

Logsdon, J.M.,Jr., Tyshenko, M.G., Dixon, C., Jafari, J.D., Walker, V.K. & Palmer, J.D. (1995) Seven newly discovered intron positions in the triose-phosphate isomerase gene: Evidence for the introns-late theory. *Proc. Natl. Acad. Sci. USA* **92**, 8507–8511.

Logsdon, J.M., Jr. & Palmer, J.D. (1994) Origin of introns-early or late? *Nature* **369**, 526.

Lohia, A. & Samuelson, J. (1993) Cloning of the Eh cdc2 gene from *Entamoeba histolytica* encoding a protein kinase $p34^{cdc2}$ homologue. *Gene* **127**, 203–207.

Long, M. & Langley, C.H. (1993) Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila. Science* **260**, 91–95.

Long, M., Rosenberg, C. & Gilbert, W. (1995a) Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl. Acad. Sci. USA* **92**, 12495–12499.

Long, M., de Souza, S.J. & Gilbert, W. (1995b) Evolution of the

intron/exon structure of eukaryotic genes. *Curr. Opin. Genet. Dev.* **5**, 774–778.

Long, M., de Souza, S.J., Rosenberg, C. & Gilbert, W. (1996) Exon shuffling and the origin of the mitochondrial targetting function in plant cytochrome c1 precursor. *Proc. Natl. Acad. Sci. USA*, in press.

Luthy, R., McLachon, R. & Eisenberg, D. (1991) Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* **10**, 229–239.

Marchionni, M. & Gilbert, W. (1986) The triosephosphate isomerase gene from maize: Introns antedate the plant–animal divergence. *Cell* **46**, 133–141.

Matrisian, L.M. (1992) The matrix degrading metalloproteinases. *BioEssays* **14**, 455–463.

McKnight, G.L., O'Hara, P.J. & Parker, M.L. (1986) Nucleotide sequence of the triosephosphate isomerase gene from *Aspergillus nidulans*: Implication for a differential loss of introns. *Cell* **46**, 143–147.

Moens, L., Vanfleteren, J., De Baere, I., Jellie, A.M., Tate, W. & Trotman, C.N.A. (1992) Unexpected intron location in non-vertebrate globin genes. *FEBS Lett.* **312**, 105–109.

Nawrath, C., Schell, J. & Koncz, C. (1990) Homologous domains of the largest subunit of eucaryotic RNA polymerase II are conserved in plants. *Mol. Gen. Genet.* **223**, 65–75.

Obaru, K., Tsuzuki, T., Setoyama, C. & Shimada, K. (1988) Structural organization of the mouse aspertate aminotransferase isozyme genes: Introns antedate the divergence of cytosolic and mitochondrial isozyme genes. *J. Mol. Biol.* **200**, 13–22.

Orgel, L.E. & Crick, F.H.C. (1980) Selfish DNA: The ultimate parasite. *Nature* **284**, 604–606.

Palmer, J.D. & Logsdon, J.M., Jr (1991) The recent origins of introns. *Curr. Opin. Genet. Dev.* **1**, 470–477.

Pardo, J.M. & Serrano, R. (1989) Structure of a plasma membrane $H^+$-ATPase gene from the plant *Arabidopsis thaliana*. *J. Biol. Chem.* **264**, 8557–8562.

Patthy, L. (1991) Exons—original building blocks of proteins? *BioEssays* **13**, 187–192.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) The evolution of genes: The chicken preproinsulin gene. *Cell* **20**, 555–566.

Piccirilli, J.A., McConnell, T.S., Zaug, A.J., Noller, H.F. & Cech, T.R. (1992) Aminoacyl esterase activity of the *Tetrahymena* ribozyme. *Science* **256**, 1420–1424.

Pohajdak, B. & Dixon, B. (1993) A commentary on: 'Unexpected intron location in non-vertebrate globin genes,' by Moens *et al.* [*FEBS Lett.* (1992) **312**, 105–109]. *FEBS Lett.* **320**, 281–283.

Quigley, F., Martin, W.F. & Cerff, R. (1988) Intron conservation across the prokaryote–eukaryote boundary: Structure of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. *Proc. Natl. Acad. Sci USA* **85**, 2672–2676.

Remillard, S.P., Lai, E.Y., Levy, Y.Y. & Fulton, C. (1995) A calcineurin-B-encoding gene expressed during differentiation of the amoeboflagellate *Naegleria gruberi* contains two introns. *Gene* **154**, 39–45.

Rogers, J. (1985) Exon shuffling and intron insertion in serine protease genes. *Nature* **315**, 458–459.

Schofield, J.P. (1993) Molecular studies on an ancient gene encoding carbamoyl-phosphate synthetase. *Clin. Sci. Colch.* **84**, 119–128.

Setoyama, C., Joh, T., Tsuzuki, T. & Shimada, K. (1988) Structural organization of the mouse cytosolic malate dehydrogenase gene: Comparison with that of the mouse mitochondrial malate dehydrogenase gene. *J. Mol. Biol.* **202**, 355–364.

Shah, D.M., Hightower, R.C. & Meagher, R.B. (1983) Genes encoding actin in higher plants: Intron positions are highly conserved but the coding sequences are not. *J. Mol. Appl. Genet.* **2**, 111–126.

Sharp, P.A. (1981) Speculations on RNA splicing. *Cell* **23**, 643–646.

Shih, M.C., Heinrich, P. & Goodman, H.M. (1988) Intron existence predated the divergence of eukaryotes and prokaryotes. *Science* **242**, 1164–1166.

Stephens, R.M. & Scheider, T.D. (1992) Features of spliceosomal evolution and function. *J. Mol. Biol.* **228**, 1124–1136.

Stoltzfus, A. & Doolittle, W.F. (1993) Slippery introns and globin gene evolution. *Curr. Biol.* **3**, 215–217.

Stoltzfus, A. (1994) Origin of introns—early or late? *Nature* **369**, 526–527.

Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, J.M., Jr. & Doolittle, W.F. (1994) Testing the exon theory of genes: The evidence from protein structure. *Science* **265**, 202–207.

Straus, D. & Gilbert, W. (1985) Genetic Engineering in the precambrian: Structure of the chicken triosephosphate isomerase gene. *Mol. Cell. Biol.* **5**, 3497–3506.

Sudhof, T.C., Goldstein, J.L., Brown, M.S. & Russell, D.W. (1985a) The LDL receptor gene: A mosaic of exons shared with different proteins. *Science* **228**, 815–822.

Sudhof, T.C., Russell, D.W., Goldstein, J.L., Brown, M.S., Sanchez-Pescador, R. & Bell, G.I. (1985b) Cassette of eight exons shared by genes for LDL receptor and EGF precursor. *Science* **228**, 893–895.

Sullivan, M.L., Carpenter, T.B. & Viestna, R.D. (1994) Homologues of wheat ubiquitin-conjugating enzymes, TAUBC1 and TAUBC4, are encoded by small multigene families in *Arabidopsis* tholioma. *Plant Mol. Biol.* **24**, 651–661.

Takahashi, Y., Urushiyama, S., Tani, T. & Ohshima, Y. (1993) An mRNA-type intron is present in the Rhodotorula hasegawae U2 small nuclear RNA gene. *Mol. Cell. Biol.* **13**, 5613–5619.

Tani, T. & Ohshima, Y. (1989) The gene for U6 small nuclear RNA in fission yeast has an intron. *Nature* **337**, 87–90.

Tani, T. & Ohshima, Y. (1991) mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing. *Genes Dev.* **5**, 1022–1031.

Tani, T., Takahashi, Y., Urushiyama, S. & Ohshima, Y. (1995) Spliceosomal introns in the spliceosomal small nuclear RNA genes. In: *Tracing Biological Evolution in Protein and Gene Structures* (eds M. Gō & P. Schimmel). Amsterdam: Elsevier.

Tittiger, C., Whyard, S. & Walker, V.K. (1993) A novel intron site in the triosephosphate isomerase gene from the mosquito *Culex tarsalis*. *Nature* **361**, 470–472.

Wilhelm, S.M., Collier, I.E., Marmer, B.L., Eisen, A.Z., Grant, G.A. & Goldberg, G.I. (1989) SV40-transformed human lung fibroblasts secrete a 92-kDa type IV collagenase which is identical to that secreted by normal human macrophages. *J. Biol. Chem.* **264**, 17213–17221.

van den Hoff, M.J.B., Jonker, A., Beintema, J.J. & Lamers, W.H. (1995) Evolutionary relationship of the carbamoylphosphate synthetase genes. *J. Mol. Evol.* **41**, 813–832.

Wing, S.S. & Bonville, D. (1994) The 14-KDa ubiquitin conjugating enzyme: Structure of the rat gene and regulation of mRNA levels upon fasting and by insulin. *Am. J. Physiol. 267*, E39-E48.

Xu, M.-Q., Kathe, S.D., Goodrich-Blair, H., Nierzwicki-Bauer, S.A. & Shub, D.A. (1990) Bacterial origin of a chloroplast intron: Conserved self-splicing group I introns in cyanobacteria. *Science* **250**, 1566–1570.