Volume 268, number 2, 339-343 FEBS 08634 August 1990

Minireview

The role of introns in evolution

John H. Rogers

Department of Physiology, University of Cambridge, Cambridge CB2 3EG, UK

Received 15 May 1990

What are the roles of 'classical' introns in the evolution of nuclear genes, and what was the origin of these introns? Exon shuffling has been important in the evolution of cell surface and extracellular proteins, but the evidence for it in respect of intracellular proteins is weak. Intron distributions imply that some introns have been removed while others have been inserted in the course of evolution; ancestral patterns of introns may thus have been obscured. Recent evidence on the self-splicing and reverse-splicing abilities of Group II introns supports the hypothesis that these could have been the ancestors of classical introns.

Intron; Exon shuffling; Domain; Chloroplast

1. INTRODUCTION

When introns were discovered in 1977, violating the prevailing notions of how genes ought to be organised, they immediately posed the questions of their present-day function (if any), their role in gene evolution, and their ultimate origin.

These are 3 logically distinct questions. First, what functions may introns perform now in individual genes? Many examples of alternative splicing are now known [1], and in some cases functional consequences can be identified, ranging from regulation of efficiency of *Ras* gene expression [2] to tissue-specific variation of the adhesive interactions of fibronectins [3-5]. But these individual functions do not answer the second question: what role have introns played in the general evolution of genes? This minireview will particularly discuss whether the possibility of exon shuffling has increased the potential for evolution. But since nature selects not for potential, but for achievement, such a role would not explain why RNA splicing existed in the first place. Thirdly, therefore, what was the origin of introns?

In the space of this minireview it is impossible to cite all the authors who have contributed to our present picture of introns in evolution. Therefore, most of the references will be to the more recent reviews and examples, from which earlier references can be obtained. This minireview will deal mainly with 'classical' introns in nuclear genes, since they are the ones that most affect the general evolution of the cell and of the organism, but first it is relevant to compare the various types of introns which may have different origins.

Correspondence address: J.H. Rogers, Department of Physiology, University of Cambridge, Cambridge CB2 3EG, UK

2. MULTIPLE TYPES OF INTRONS

Classical introns: in nuclear genes of eukaryotes, almost always beginning with the dinucleotide GT and ending with AG, and spliced by small nuclear ribonucleoprotein particles (snRNPs), via a mechanism that involves a 'lariat' topology.

Group I: in many genomes including mitochondria, chloroplasts, at least one nuclear gene, and even bacteriophage T4 [6]. Some group I introns in different genes are clearly homologous and can apparently propagate as autonomous elements. Many of them encode proteins that enable them to transpose at the DNA level into homologous sites [7], and many of them are self-splicing as pure RNAs [8], although some require ancillary proteins in vivo. The self-splicing capability may enable them to avoid damaging genes into which they insert themselves.

Group II: in mitochondria and chloroplasts [9]. The splicing mechanism of group II introns is different from that of group I, but similar to that of classical introns. These introns may also be autonomous elements, since they too are self-splicing. They have not actually been shown to transpose, nor to encode a transposase. Instead, many of them encode a large protein which includes homology to reverse transcriptase [9], and there is evidence for reverse transcriptase activity, in that group II introns are required for loss of other introns by

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/90/$3.50

apparent reverse transcription events in mitochondria [10]. (A reverse transcriptase-like gene product may also have transposase activity [11].) A possible first step in self-insertion of the intron is the reversal of the self-splicing reaction, which has recently been demonstrated in vitro [12,13].

Group III: short, very (A,T)-rich introns in *Euglena* chloroplasts, whose splicing mechanism and sequence requirements are unknown [14]. Similar unusual introns have been found in *Drosophila* [15] and chicken [16].

Transfer RNA introns: short introns which are spliced by a different mechanism again, and may best be viewed as a special case of tRNA processing, evolutionarily unrelated to other types of splicing.

3. EXON SHUFFLING?

The most widely discussed role of introns in the evolution of genomes has been 'exon shuffling' - the promotion of non-homologous rearrangements between genes [17]. The large size of introns means that random rearrangements within them can bring exons into new combinations with much higher frequency than would be possible for rearrangements in continuous coding sequences. (However, this is also responsible for the high frequencies of deleterious rearrangements in large genes that lead to common genetic diseases, such as muscular dystrophy and familial hypercholesterolaemia.)

Individual exons encode many extracellular protein domains, in both membrane and secreted proteins, and these exons have been extensively shuffled in evolution [18–21]. The immunoglobulin (Ig) genes were the first to show homologous domains encoded by separate exons, and this arrangement has been shown by all members of the Ig superfamily since, including cell adhesion molecules and growth factor receptors as well as almost all the surface molecules involved in the immune response. The introns separating Ig-type domains are always between the first and second nucleotides of a codon (phase I). The great proliferation of this superfamily may have been owed to the fortuitous existence of a proto-domain exon flanked on each side by a phase I intron (which may be called a phase I-I exon), so that duplications of this exon automatically created tandemly arranged domains which could assemble properly in the protein.

The Ig-type domain is only one of at least 7 types of domains that are characteristically encoded by phase I-I exons and that appear in different genes in a variety of combinations [19]. Indeed, several new candidates for shuffled phase I-I domains have been identified [20] by looking for further homologies between proteins that were already known to contain such domains. The Ig-type domain itself has been found in many adhesion molecules (such as NCAM [22]) in tandem with domains distantly related to the type III repeats of fibronectin, and these 'FnIII-type' domains form one of the most widespread of these families, also occurring in cytokine receptors [21] and *sevenless* [23]. Other widespread phase I-I domain families are related to epidermal growth factor, to fibronectin type I and II repeats, and to lectin domains [20,24]. They are well represented in the non-enzymatic domains of serine proteases [18,19,25].

It is not known why all these domain families are made up of phase I-I exons, but this could be merely the result of a random selection. As argued by Patthy [19], there could have been phase 0–0 and phase II–II domain families as well, but the success of several phase I domain families in creating new proteins by exon shuffling meant that there were more phase I–I exons available for future shuffling, and that new phase I–I domains could fit into the expanding set of phase I–I exons, whereas exons with boundaries in other phases would have remained isolated.

All this concerns domains expressed outside the cell [19]. However, Gilbert and others (e.g. [26,27]) have argued for a much more general rule that exons originally encoded protein domains, or even smaller 'modules' of protein structure that were convenient for assembling stable domains. The advocates of this strong theory of exon shuffling have claimed in particular that many genes for ancient intracellular enzymes have introns at divisions between domains or modules. This view was encouraged by Gō's analysis of the haemoglobin structure, which mapped the two introns to divisions between modules in the protein, and identified a third such division which was subsequently found to coincide with an intron in the homologous leghaemoglobin gene. A similar analysis of triose phosphate isomerase [26,28] also showed significant correlations, and an early survey [29] found that introns tended to map to protein surfaces. Otherwise, objective evidence for the supposed 'modules' seems to be lacking, and it is puzzling that Gō's model of triose phosphate isomerase [26,28] assigns divisions within secondary structure elements like β strands, whereas other authors are happy to locate divisions between such elements. In most published claims for exon-module correlations, the supposed structural modules are only identified with hindsight, apparently on aesthetic grounds [30]. (Other possible explanations for the apparent regularities will be considered below.)

If one looks at the most likely candidate for primitive exon shuffling, the mononucleotide binding fold of dehydrogenases and other metabolic enzymes, it has been claimed that it is more-or-less bounded by introns in several genes. But the introns do not coincide in the different genes, and they are not even in the same phase of the reading frame [30]. Thus proponents of the strong exon shuffling theory have to invoke intron

movement, regardless of the implausibility of this occurring without inactivating the gene (see below), and of the logical difficulty of reconstructing ancestral patterns if the introns are supposed to have moved.

If the strong exon shuffling theory were true, it would imply that introns were present in prokaryotes (see below), and an interesting corollary would be that the splicing machinery would have been in the same compartment as the ribosomes and thus could have been sensitive to the coding phase of potential introns. Thus the predominance of phase I–I exons in known exon shuffling events might reflect an ancestral mechanism geared to phase I splicing, which would certainly have facilitated evolution. However, there is no need of any such hypothesis, and phase I introns are not preferred among genes in general, nor at supposed module boundaries in metabolic enzymes [30].

4. ORIGIN OF CLASSICAL INTRONS

How far back can introns be traced? They are present in all eukaryotes, and some individual introns are in the same positions in mammals and plants, implying that they date from the very earliest eukaryotes [26]. There are very few known cases of introns which have been moved or inserted within the vertebrate lineage.

However, there are many differences in intron positions if one compares homologous sequences that were separated before the vertebrate radiation: between orthologous genes in different phyla, or between members of dispersed gene families, or between internal repeats within a single gene [32,33]. Therefore there were large-scale rearrangements of introns in the earlier phases of evolution. Either many introns have moved, or many ancestral introns have deen deleted, or many new introns have been inserted. Or all three may have occurred.

Intron movement cannot be a general explanation of the distributions, as many of the discordant introns would have to have moved across conserved coding sequences or across a non-integral number of codons. While schemes for achieving this can be devised, and there even appears to be one example in a carbonic anhydrase gene [31], the calculated probability of the required double frameshift events (or of the genes surviving a transitional state) seems much too low to invoke such events as a general phenomenon [19,32,33].

Intron removal certainly can occur. The most abundant examples are the processed pseudogenes of mammals, and reverse transcriptase is likely to be available (thanks to retrotransposons) throughout the eukaryotic kingdom. There are several examples of functional genes is vertebrates where individual introns have been removed (e.g. [34]), and intron removal seems to have been more prevalent in small-genome organisms like yeast [35] or *Drosophila*. For example, these organisms lack several introns shared by plants and mammals in the genes for actin [36], triose phosphate isomerase [26], and glyceraldehyde-3-phosphate dehydrogenase (GAPDH). However, intron removal alone, from an ancestral gene with many introns of which we now see only subsets, cannot account for the present haphazard distributions of introns; some ancestral genes, encoding for example the serine protease catalytic domain or the 'EF hand' calcium-binding domain [25,37], would have to have started off with a vast number of introns, many separated by only one or a few nucleotides.

Many authors have tried to interpret intron distributions in terms of a mixture of removal and movement of introns, but the difficulty of phase-shifting movement must cast serious doubt on these interpretations. In contrast, some gene families for which phylogenies can be traced show patterns that clearly indicate intron insertions.

The first clear evidence for intron insertions was seen in the catalytic domains of the serine protease family [25]. Discordant introns most easily explained by insertion have also been noted in other genes or gene families, including genes for collagens, a zinc finger protein, actins, and tubulins [32,36].

The superfamily of calcium-binding proteins with their repeated homologous domains provides a large set of examples [37,38]. Most strikingly, the family of four genes that includes calmodulin and myosin alkali light chain shares common introns in domains I, II, and IV, but each has a different intron in domain III – apparently inserted after the separation of the four genes, close to the middle of what would then have been the longest exon. These events predated the origin of vertebrates, as each of the shared introns is present in at least one gene of the family in *Drosophila*, as is the gene-specific intron in domain III of myosin alkali light chain. (The insect genes also demonstrate intron removal as they each lack at least one of the shared introns.) But *Chlamydomonas* calmodulin [39] and other subfamilies of calcium-binding proteins have different sets of introns. In this superfamily it appears that introns have been inserted preferentially into certain regions, and have tended to divide the genes into exons of uniform size, thus producing apparently regular distributions which nevertheless owe little or nothing to any ancestral arrangement [37,38].

Most graphically, discordant introns are found within some Ig-like domains, for example in NCAM [22], and within some FnIII domains in fibronectin [3,5]. In these genes, some of the phase I–I domains are not only flanked by the usual conserved introns, but are split by an extra intron that has no fixed position nor phase. Again, these seem to be cases of intron insertion near the middle of a pre-existing exon.

If most introns have been inserted, an explanation is needed for the apparent semi-regularities in their distribution: the fairly uniform size of exons [30] and

the tendency in some genes to map near protein structural divisions. These patterns could be caused by sequence specificity in the insertion process, or by selection for efficiency of splicing after insertion, which might depend on local exon sequences, on the proximity of other splice sites, or on the maintenance of secondary structure in the pre-mRNA.

How might introns have been inserted? Several types of mechanism can be envisaged. First, they might be a type of transposable element, which had splice sites at its ends so that it would not damage genes into which it inserted. Unfortunately the 'GT-AG' splice sites are not compatible with the characteristic terminal sequences of any known type of transposable element. However, an approximation to this scenario is achieved by the McClintock transposons of maize [40], which contain a cryptic internal 5′ splice site such that they can be imprecisely spliced out and some gene function maintained. The *Drosophila* retroposon *suffix* also carries a splicing site [15], although apparently not suitable for excising the retroposon.

A second possible mechanism for creating introns would be to duplicate exonic sequences which happen to contain a cryptic splice site, but the evidence from conserved sequences flanking introns is against such an origin [33]. A third possible mechanism which has been advocated [33] is to mutate introns of non-classical types which probably do behave as self-inserting elements (see below).

Whatever the mechanism of intron insertion, the evidence that it has occurred – in parallel with intron removal – means that reconstructing ancestral intron distributions is likely to be a hazardous affair. In large-genome organisms such as mammals and plants, many of the ancestral and inserted introns seem to have been retained. In small-genome organisms such as *Drosophila* and yeast [35], most of the ancestral introns have been removed, and the introns that are present (often in different positions from those of mammals) are probably comparatively recent insertions. How complex and uncertain these histories can be is shown by the genes for myosin heavy chain, which show patterns suggestive of intron removal in the 'head' region but of intron insertion in the 'tail' region [41].

How old are the oldest introns? It has been reported that 'intron existence predated the divergence of eukaryotes and prokaryotes' [42,43], on the grounds that some intron positions are shared between the genes for cytosol GAPDH and chloroplast GAPDH. Although both genes are now nuclear, the gene for the chloroplast enzyme seems to have been transferred from the prokaryotic endosymbiont that evolved into the chloroplast. And yet genes for chloroplast GAPDH include two introns in positions, respectively identical to an intron of chicken GAPDH [42] and an intron of nematode GAPDH [43], while a third 'chloroplast' intron, in a non-conserved region, falls only one codon away from an intron in the 'cytosol' genes of both plant and chicken [42]. Even to proponents of primordial introns, it should seem surprising that such a distinctly eubacterial organism as the ancestor of chloroplasts should have retained classical introns even though all present-day eubacteria apparently lack them. But it would be at least as surprising if intron insertion into the 'chloroplast' gene had produced two or three introns in identical positions. And recombination between the 'chloroplast' and 'cytosol' genes in the nucleus seems to be ruled out because the affinities of the 'chloroplast' gene with prokaryotes (including *B. subtilis* [44] as well as thermophilic bacteria) are unmistakable in the immediate vicinity of the two shared introns. The evolution of the GAPDH introns therefore remains a puzzle.

Going back still earlier, Archaebacteria have introns of the tRNA type and possibly related types [45], but classical introns have not yet been found in them.

Were introns present in the very earliest genomes, or were they inserted later? The idea that they were present from the beginning was offered by Doolittle [46], who argued that the first organisms would not have been able to replicate full-length genes with sufficient accuracy, and that RNA splicing would have offered a way of piecing together genes from a large number of fragments. This scenario seemed attractive, partly because it fitted in with the emerging arguments for exon shuffling, and partly because it avoided having to find a reason and a mechanism for insertion of introns later. Both these considerations are now much weaker, as there is strong evidence that some classical introns were inserted into existing genes, and that group I and perhaps group II introns are fairly simple entities capable of self-inserting as autonomous elements. It is quite possible that introns arose as 'selfish DNA' (or 'selfish RNA') like group I or group II introns, which parasitised the genome to such an extent that they became an ineradicable part of it [47]. The autonomous precursor of classical introns may have been lost aeons ago, but it may have been very like the group II introns, which have the same splicing mechanism as classical introns. It has been proposed [48] that the snRNP machinery evolved from self-splicing introns like group II which were able to splice other introns in *trans*. It also seems possible that individual group II insertions could evolve into classical introns, since only a single base change is needed to convert some group II introns into sequences conforming to the classical splice site consensus [33].

Thus there is no need to believe that introns were primitive, but the occurrence of intron deletions and intron insertions (probably selective) may have obscured original patterns. In view of the lack of evidence about the earliest genomes, and the manifest capabilities of autocatalytic RNA, we still cannot be certain where introns came from.

REFERENCES

[1] Breitbart, R.E., Andreadis, A. and Nadal-Ginard, B. (1987) Annu. Rev. Biochem. 56, 467–495.
[2] Cohen, J.B., Broz, S.B. and Levinson, A.D. (1989) Cell 58, 461–472.
[3] Schwarzbauer, J.E., Patel, R.S., Fonda, D. and Hynes, R.O. (1987) EMBO J. 6, 2573–2580.
[4] Ffrench-Constant, C. and Hynes, R.O. (1989) Development 106, 375–388.
[5] Hynes, R.O. (1989) Fibronectins, Springer, New York.
[6] Cech, T.R. (1988) Gene 73, 259–271.
[7] Lambowitz, A.M. (1989) Cell 56, 323–326.
[8] Cech, T.R. (1987) Science 236, 1532–1539.
[9] Michel, F. and Jacquier, A. (1987) Cold Spring Harbor Symp. Quant. Biol. 52, 201–212.
[10] Levra-Juillet, E., Boulet, A., Séraphin, B., Simon, M. and Faye, G. (1979) Mol. Gen. Genet. 217, 168–171.
[11] Xiong, Y. and Eickbush, T.H. (1988) Cell 55, 235–246.
[12] Augustin, S., Müller, M.W. and Schweyen, R.J. (1990) Nature 343, 383–386.
[13] Mörl, M. and Schmelzer, C. (1990) Cell 60, 629–636.
[14] Christopher, D.A. and Hallick, R.B. (1989) Nucleic Acids Res. 17, 7591–7608.
[15] Tchurikov, N.A., Ebrlidze, A.K. and Georgiev, G.P. (1986) EMBO J. 5, 2341–2347.
[16] Kiss, I., Deák and Lukácsovich, T. (1988) Nucleic Acids Res. 16, 1211.
[17] Gilbert, W. (1978) Nature 271, 501.
[18] Patthy, L. (1985) Cell 41, 657–663.
[19] Patthy, L. (1987) FEBS Lett. 214, 1–7.
[20] Patthy, L. (1988) J. Mol. Biol. 202, 689–696.
[21] Patthy, L. (1990) Cell 61, 15–16.
[22] Owens, G.C., Edelman, G.M. and Cunningham, B.A. (1987) Proc. Natl. Acad. Sci. USA 84, 294–298.
[23] Norton, P.A., Hynes, R.O. and Reese, D.J.G. (1990) Cell 61, 15–16.
[24] Suter, U., Bastos, R. and Hofstetter H. (1987) Nucleic Acids Res. 15, 7295–7307.
[25] Rogers, J. (1985) Nature 315, 458–459.
[26] Gilbert, W., Marchionni, M. and McKnight, G. (1986) Cell 46, 151–154.
[27] Blake, C.C.F. (1985) Int. Rev. Cytol. 93, 149–185.
[27a] Gō, M. (1981) Nature 291, 90–92.
[28] Gō, M. and Nosaka, M. (1987) Cold Spring Harbor Symp. Quant. Biol. 52, 915–924.
[29] Craik, C.S., Sprang, S. Fletterick, R. and Rutter, W.J. (1982) Nature 299, 180–182.
[30] Traut, T.W. (1988) Proc. Natl. Acad. Sci. USA 85, 2944–2948.
[31] Yoshihara, C.M., Lee, J.-D. and Dodgson, J.B. (1987) Nucleic Acids Res. 15, 753–770.
[32] Rogers, J. (1986) Trends Genet. 2, 223.
[33] Rogers, J.H. (1989) Trends Genet. 5, 213–216.
[34] Srikantha, T., Landsman, D. and Bustin, M. (1990) J. Mol. Biol. 211, 49–61.
[35] Fink, G.R. (1987) Cell 49, 5–6.
[36] Dibb, N.J. and Newman, A.J. (1989) EMBO J. 8, 2015–2021.
[37] Wilson, P.W., Rogers, J., Harding, M., Pohl, V., Pattyn, G. and Lawson, D.E.M. (1988) J. Mol. Biol. 200, 615–625.
[38] Perret, C., Lomri, N. and Thomasset, M. (1988) J. Mol. Evol. 27, 351–364.
[39] Zimmer, W.E., Schloss, J.A., Silflow, C.D., Youngblom, J. and Watterson, D.M. (1988) J. Biol. Chem. 263, 19370–19383.
[40] Wessler, S.R. (1989) Gene 82, 127–133.
[41] Dibb N.J., Maruyama, I.N., Krause, M. and Karn, J. (1989) J. Mol. Biol. 205, 603–613.
[42] Shih, M.-C., Heinrich, P. and Goodman, H.M. (1988) Science 242, 1164–1166.
[43] Quigley, F., Martin, W.F. and Cerff, R. (1988) Proc. Natl. Acad. Sci. USA 85, 2672–2676.
[44] Viaene, A. and Dhaese, P. (1989) Nucleic Acids Res. 17, 1251.
[45] Kjems, J. and Garrett, R.A. (1988) Cell 54, 693–703.
[46] Doolittle, W.F. (1978) Nature 272, 581.
[47] Cavalier-Smith, T. (1985) Nature 314, 283–284.
[48] Sharp, P. (1985) Cell 42, 397–400.