Volume 325, number 1,2, 135–139 FEBS 12449 June 1993

Minireview

Why do genes have introns?

N.J. Dibb

Department of Haematology, Royal Postgraduate Medical School, Hammersmith Hospital, Du Cane Road, London W12 ONN, UK

Received 30 March 1993

This review outlines some of the models that have been proposed to explain why genes have introns and then explores the possibility that alternative splicing might be the cause rather than a consequence of split genes.

Intron origin; Alternative splicing; Evolution

1. INTRODUCTION

1.1. *The discovery of introns*

In 1977 several independent groups published their remarkable discovery that the coding sequences of eukaryotic genes are interrupted by non-coding DNA. These groups are listed by Gilbert [1] who also coined the terms exon and intron for the coding and non-coding regions of split genes. Fig. 1 illustrates that split genes can have a variety of different structures. In some ways, genes such as those coding for actin, globin and tubulin have the most puzzling structures. These genes, which are represented by Fig. 1a, encode single proteins that cannot be made until all introns are precisely removed from their pre-mRNA by splicing. The discovery of such introns was entirely unexpected and there is still much debate as to why they exist (see below). Experiments in which split genes are converted into contiguous genes (by DNA replacement), have contributed to but have not resolved this issue; the expression of some genes is inhibited by this treatment [2], whereas others are not noticeably affected [3].

1.2. *Alternative splicing*

The split structures of the remaining genes illustrated in Fig. 1 have more obvious functional consequences. These genes contain alternatively spliced exons in addition to constitutively spliced exons and so are able to encode more than one protein isoform through alternative splicing. This mechanism is used by a wide range of eukaryotes and it allows a single troponin-T gene, for example, to encode 6 known protein isoforms and possibly many more [4,5]. Protein isoforms produced from the same gene through alternative splicing have, in many cases, been shown to have functional differences [5]. The isoforms of certain transcription factors, for example, differ to the extent of either activating or inhibiting transcription [6,7]. Alternative splicing is also subject to regulation, which means that the type and/or proportions of protein isoforms that are expressed by a single gene may change during development and can also differ between mature, cell types [4,5,8].

Correspondence address: N.J. Dibb, Department of Haematology, Royal Postgraduate Medical School, Hammersmith Hospital, Du Cane Road, London W12 ONN, UK.

1.3. *The spliceosome*

Both intron removal and the generation of alternative transcripts occur in a structure called a spliceosome which has only been found in eukaryotes and is at least as complex as a ribosome. The spliceosome or splicing machinery consists of a set of five small nuclear ribonucleoprotein (snRNP) particles and a number of accessory proteins [8,9]. The spliceosome is a dynamic structure that is assembled upon pre-mRNA and dissociates from mRNA following splicing. During this process splice sites are recognised, cleaved and spliced in a two-step transesterification reaction (Fig. 2).

1.4. *Self-splicing introns*

There are three known groups of self-splicing introns (I, II and III) and of these group II introns have the most similar splice sites and mechanism of removal to classical or spliceosomal introns [10]. Group II introns have been found in mitochondria and chloroplasts but not in nuclear genes.

1.5. *Splice sites*

It is often difficult to identify splice sites with certainty because many splice sites do not flank introns (Fig. 1) and although splice sites conform to the consensus sequences shown in Fig. 2, only the GT and AG

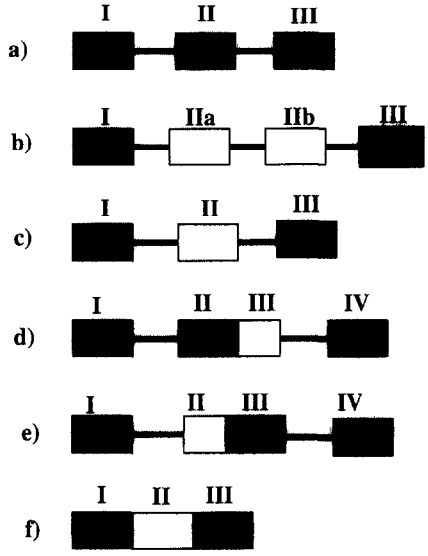

Fig. 1. Some examples of the different structures that eukaryote genes can have. (a) A gene with three exons and two introns both of which must be constitutively removed from pre-mRNA in order to generate a single type of protein encoded by exons I, II, and III. (b) A gene that contains two exons IIa and IIb that undergo mutually exclusive, alternative splicing to generate two protein isoforms I, IIa, III or I, IIb, III. (c) A gene with an alternatively spliced exon that is either kept or removed during pre-mRNA splicing and so can generate two protein isoforms I, II, III or I, III. (d) A gene in which the 5′ splice site of an alternatively spliced exon III is located entirely within the coding region, two possible protein isoforms I, II, III, IV or I, II, IV. (e) A gene in which the 3′splice site of an alternatively spliced exon II is located entirely within the coding region, two possible protein isoforms I, II, III, IV or I, III, IV. (f) A gene in which an alternatively spliced exon is not flanked by introns, two possible isoforms I, II, III or I, III. Actual genes may well contain a mixture of the different types of alternatively spliced exons. Examples of such genes are described in a recent review [5]. Black box, constitutively spliced exon; white box, alternatively spliced exon; line, intron (non-coding segment of the gene). Modified from ref. [5].

bases of 5′ and 3′ splice sites respectively are highly conserved. Consequently, there are numerous sequences within pre-mRNAs that have good matches with splice sites yet are not used [11].

Many genetic diseases are caused by mutations that alter the pattern of splicing of individual genes [12]. Such mutations may either disrupt normal splice sites or create new ones. Often the mutation of an authentic splice site leads to the activation of nearby cryptic splice sites [12–14] or causes exon skipping [15]. Numerous cryptic splice sites have now been identified that lie dormant within either introns or exons. It is generally held that 'authentic' and 'cryptic' splice sites in a pre-mRNA compete for the splicing machinery and that authentic splice sites dominate because they have the highest affinity [8,9].

2. HOW OLD ARE INTRONS?

Ever since their discovery there has been much debate about whether introns are recent inserts into eukaryote genes [16] or ancient relics that have been lost by the prokaryotes [17]. There is also a related controversy as to whether exon shuffling has occurred early or late in gene evolution [18,19]. Table I outlines the differences and similarities between the early and late models of intron origin and two other models.

2.1. *Exon shuffling*

Patthy [19,20] has catalogued a large number of eukaryotic genes whose exons are very likely to have been assembled by intergenic recombination between introns (exon shuffling), as predicted by Gilbert [1]. These genes are all unique to eukaryotes indicating that exon shuffling has occurred relatively late in gene evolution.

In support of early shuffling, Gilbert et al. [21] reported that the exons of the gene for triosephosphate isomerase (TIM) encode single or multiple protein modules consisting of tightly clustered amino acids, as originally defined by Go for the globin gene [22]. The significance of this correlation has been strengthened by the

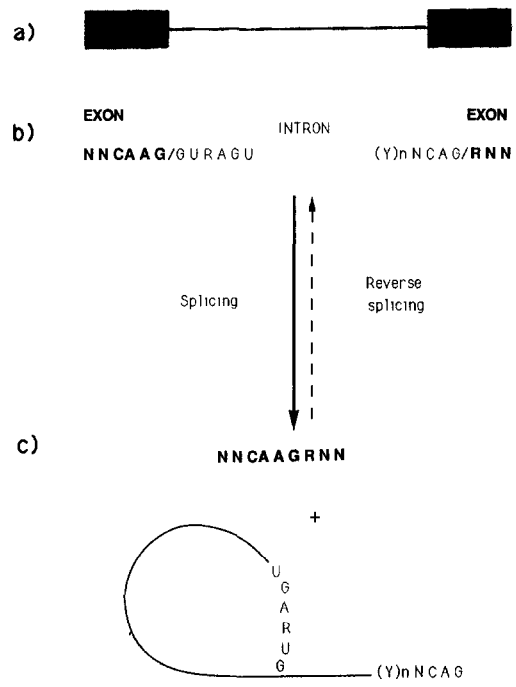

Fig. 2. Intron removal by the splicing machinery. (a) A gene with an intron that is flanked by two exons. (b) pre-mRNA showing the sequence of the 5′ and 3′ splice sites, the diagonal lines indicate the sites of cleavage by the splicing machinery. The sequence between the sites of cleavage is usually non-coding and is of variable length. In general, only the 5′ and 3′ splice sites of introns are highly conserved, in particular, the GU at the 5′ splice site and the AG at the 3′ splice site are normally invariant. Introns are also flanked by a conserved coding sequence shown in bold type. (c) The intron is removed in the form of a lariat and the flanking coding sequences are joined by the splicing machinery. The dotted arrow indicates that intron removal is reversible in theory. As indicated, reverse splicing would be expected to insert introns between the G and R of the coding sequence CA, A, G, R. Modified from ref. [35].

discovery of a TIM intron [23] that was predicted to exist by Gilbert et al. [21]. They question whether this correlation could have resulted from intron insertions and instead suggest that the TIM gene was assembled by exon shuffling [21]. Because this gene has a prokaryote counterpart, it follows that exon shuffling must have occurred early in gene evolution before the separation of the eukaryotes and prokaryotes.

Against this interpretation is the fact that TIM introns are restricted to just some of the eukaryotes suggesting that they were inserted [16,24]. In accordance with this, Patthy's analyses [19,20] show that successful exon shuffling needs to be mediated by introns of the same phase, whereas TIM introns are a mixture of all three.

It should be noted that all proteins have probably been assembled from smaller motifs [25]. This outcome was anticipated before the discovery of introns; when it was assumed that the necessary recombination events occurred between contiguous genes. This view may still be largely correct. It is by no means established that the smallest protein motifs, in general, are or were encoded by exons [19,26–28].

2.2. *Intron insertion*

Hundreds of thousands of introns appear to have been inserted into genes during the eukaryotic radiation [24], so supporting late rather than early models of intron origin (Table I). This means that the split structure of many genes, such as actin and tubulin genes [29], cannot be of very great evolutionary significance simply because their introns were inserted after they had evolved.

It has been suggested that introns are degenerate derivatives of group II intron insertions [30], which if true would support the selfish DNA model of intron origin proposed by Cavalier-Smith [16]. However, it is clear that most introns have been inserted between the nucleotides G and R (where R is a G or A) of the consensus sequence C or A, A, G, R (termed a proto-splice site), which is virtually identical to the coding sequence which flanks introns in general [29,31] (Fig. 2). This is not an expected target site for group II intron insertions [32,33].

Proto-splice sites are the expected target sites for the insertion of spliceosomal introns by reverse splicing [34] (Fig. 2). Spliceosomal introns that are inserted by this mechanism will have all of the necessary signals for their efficient removal from pre-mRNA. Consequently, these intron insertions could spread within a population as neutral mutations and later acquire useful functions. Alternatively, such insertions might well confer a small but immediate selective advantage [35], so causing their fixation by selection.

The insertion of introns by reverse splicing is similar to the mechanism proposed to explain the loss of introns by organisms such as *Saccharomyces cerevisiae* [36], it seems likely that both mechanisms have operated during evolution [35].

3. HOW OLD IS ALTERNATIVE SPLICING?

As discussed above, it is possible that most introns originated from pre-existing introns that were inserted into genes by the mediation of the very machinery which now removes them from pre-mRNA. This supports the view that alternative splicing is more ancient and of greater evolutionary significance than intron removal [4,5]. It should be noted that alternative splicing can occur in the absence of introns (Fig. 1d,e,f).

Alternative splicing would also be expected, on occasion, to generate new introns from the coding sequences of genes [4,5,30,35]. Consequently, it is formally possible that alternative splicing is responsible for the origin of the introns that appear to have been spread by reverse splicing during evolution.

4. THE ADVANTAGES OF ALTERNATIVE SPLICING

4.1. *Many proteins from a single gene*

Complex eukaryotes would appear to benefit from the functional diversity afforded by multiple genes or by single genes that have alternative splice sites [4,5]. Because eukaryotes often have excessively large genomes it seems implausible to regard the use of single genes that encode many proteins as a parsimonious adaptation. Instead it simply suggested that this strategy allows such genes to generate the equivalent advantages of heterosis without incurring a segregation load [37].

4.2. *Increased variation*

Splicing acts to delete regions of pre-mRNA (Fig. 1), and so has an equivalent effect to an intragenic deletion. However, only alternative splicing allows a single gene to encode both a normal and a truncated protein [1]. A consequence of this is that for the same mutation load, splicing mutations can reach, on average, far higher frequencies in a population than can equivalent intragenic deletions.

Often, splicing acts in concert with intragenic duplications so allowing a region of a protein to be replaced with a region of similar structure but different sequence (Fig. 1b). This happens because intragenic duplications often duplicate splice sites [5]. The net result is that alternative splicing allows intragenic duplications to reach much higher frequencies in a population than they otherwise could [5].

The combination of alternative splicing and intragenic duplications greatly increases the amount of radical variation that can be maintained in a population. Therefore, it is hardly surprising that this variation has been exploited by the eukaryotes during evolution [5].

Table I
Models that have been proposed to explain introns.

Class of model	I	I	II	II
Subclass	Early	Late	Non-coding	Coding
Origin of classical introns	Generated by the splicing machinery from self-splicing introns that split the most ancient of genes	Generated by the splicing machinery from self-splicing introns that were inserted into genes	By-products that were generated by the splicing machinery from non-coding sequences that split ancestral genes	By-products that were generated within genes during the evolution of alternative splicing.
Origin of splice sites	Splice sites of self-splicing introns	Splice sites of self-splicing introns	Stop codons	Proto-splice sites
Driving force for the evolution of the splicing machinery	To facilitate intron removal	To facilitate intron removal	To enlarge the coding sequence of genes	To enhance the production of advantageously spliced mRNA.
Advent of exon shuffling	Used to assemble the earliest of genes	After intron insertion	Could have occurred early in gene evolution	After intron evolution and insertion
Advent of alternative splicing	An opportunistic development of intron removal that could have occurred early in gene evolution	An opportunistic development of intron removal that could only have occurred late in gene evolution	An opportunistic development of intron removal that could have occurred early in gene evolution	Prior to intron removal. Most constitutively spliced introns proposed to have originated by reverse splicing.
Reasons why introns might correlate with protein structure	Selection of advantageous exon shuffling events	Selection of advantageous exon shuffling events	Selection of advantageous splicing and exon shuffling events	Selection of advantageous splicing and exon shuffling events
References	17, 28	16	40	29,35

Class I models propose that classical or spliceosomal introns originated from self-splicing introns, these have been subclassified into early and late models. Class II models propose that classical introns did not originate from self-splicing introns but from either non-coding or coding sequences. From [35].

5. A MODEL FOR THE EVOLUTION OF ALTERNATIVE SPLICING AND THE ORIGIN OF INTRONS

The marked similarity between group II introns and the spliceosome has led to the generally held view that the spliceosome originated from an ancestor of a group II self-splicing intron that acquired the ability to splice *in trans* [10,34,37,38]. Although it has been argued that their similarity could have arisen through convergent evolution [39].

Consider the consequences of an ancestral self-splicing intron that, for whatever reason, started to splice *in trans*; albeit at a fraction of the efficiency of the present spliceosome. Its 'services' would presumably have become available to any RNA sequence that conformed to an ancestral group II splice site. Numerous such sequences would be expected to have occurred by chance within the contiguous coding sequences of ancestral mRNAs. These may well have competed for recognition by such an ancestral intron, just as present splice sites compete for recognition by the spliceosome.

It has been suggested that there are a number of expected outcomes of these starting conditions [35,37]. One, the tiny minority of 'proto-splice sites' whose use resulted in the production of advantageous protein isoforms would be under selective pressure to evolve towards a sequence better able to compete for recognition. Two, from this it follows that any increase in the ability of the ancestral spliceosome to act *in trans* would be selected. Three, introns would be expected to be generated within contiguous genes as by-products of this process.

This model (the fourth of Table I) can also explain the similarity of group II and spliceosomal introns. It differs from the other models listed in Table I in that it predicts that the spliceosome evolved for the purpose of alternative splicing rather than intron removal.

Further elucidation of both the mechanism of splicing of all intron types and of the gene structures of the lesser known eukaryotes promises to yield answers, and no doubt more surprises, about one of the most fascinating of evolutionary problems.

Acknowledgements. I thank Nick Cross, James Crow, Lucio Luzzatto, Andy Newman, Sara Pappworth, Irene Roberts, John Maynard Smith and Tom Vulliamy for their help. This work was supported by the Hunting Gate Group and the MRC.

REFERENCES

[1] Gilbert, W. (1978) Nature 271, 501.
[2] Buchman, A.R. and Berg, P. (1988) Mol. Cell Biol. 8, 4395–4405.
[3] Gross, M.K., Kainz, M.S. and Merrill, G.F. (1987) Mol. Cell Biol. 7, 4576–4581.
[4] Andreadis, A., Gallego, M. and Nadal-Ginard, B. (1987) Annu. Rev. Cell Biol. 3, 207–242.
[5] Smith, C.W.J., Patton, J.G. and Nadal-Ginard, B. (1989) Annu. Rev. Genet. 23, 527–577.
[6] Foulkes, N. S., Mellstrom, B., Benusiglio, E. and Sassone-Corsi, P. (1992) Nature, 355, 80–84.
[7] Roman, C., Cohn, L. and Calame, K. (1991) Science 254, 94–97.
[8] Green, M.R. (1991) Annu. Rev. Cell Biol. 7, 559–599.
[9] Padgett, R.A., Grabowski, P.J., Konarska, M.M., Seiler, S. and Sharp, P.A. (1986) Annu. Rev. Biochem. 55, 1119–1150.
[10] Jaquier, M. (1990) Trends Biochem. 15, 351–354.
[11] Ohshima, Y. and Gotoh, Y. (1987) J. Mol. Biol. 195, 247–259.
[12] Kuivaniemi, H., Kontusaari, S., Tromp, Zhao, M., Sabol, C and Prockop, D.J. (1990) J. Biol. Chem. 265, 12067–12074.
[13] Treisman, R., Orkin, S.H. and Maniatis, T. (1983) Nature, 302, 591–596.
[14] Wieringa, B., Meyer, F., Reiser, J. and Weissmann, C. (1983) Nature 301, 38–43.

[15] Hayashi, S.-I., Kunisada, T., Ogawa, M., Yamaguchi, K. and Nishikawa, S.-I. (1991) Nucleic Acids Res. 19, 1267–1271.
[16] Cavalier-Smith, T. (1991) Trends Genet. 7, 145–148.
[17] Darnell, J.E. and Doolittle, W.F. (1986) Proc. Natl. Acad. Sci. USA 83, 1271–1275.
[18] Dorit, R.L., Schoenbacher, L and Gilbert, W. (1991) Science 250, 1377–1382.
[19] Patthy, L. (1991) Bioessays 13, 187–192.
[20] Patthy, L. (1991) Curr. Opinion Struct. Biol. 1, 351–361.
[21] Gilbert, W., Marchionni, M. and McKnight, G. (1986) Cell 46, 151–154.
[22] Go, M. (1981) Nature 291, 90–92.
[23] Tittiger, C., Whyard, S. and Walder, V.K. (1993) Nature 361, 470–472.
[24] Palmer, J.D. and Logsdon, J.M. (1991) Current Opinion Genet. Dev. 1, 470–477.
[25] Bairoch, A. (1992) Nucleic Acids Res. 20, 2013–2018.
[26] Doolittle, R.F. (1985) Trends Biochem. 10, 233–237.
[27] Blake, C. (1983) Nature 306, 535–537.
[28] Gilbert, W. (1987) Cold Spring Harbor Symp. Quant. Biol. 52, 901–904.
[29] Dibb, N.J. and Newman, A.J. (1989) EMBO J. 8, 2015–2021.
[30] Rogers, J.H. (1989) Trends Genet. 5, 213–216.
[31] Lee, V.D., Stapleton, M. and Huang B. (1991) J. Mol. Biol. 221, 175–191.
[32] Augustin, S., Müller, M.W. and Schweyen, R.J. (1990) Nature 343, 383–386.
[33] Morl, M. and Schmelzer, C. (1990) Cell 60, 629–636.
[34] Sharp, P.A. (1985) Cell 42, 397–400.
[35] Dibb, N.J. (1991) J. Theor. Biol. 151, 405–416.
[36] Fink, G.R. (1987) Cell 49, 5–6.
[37] Cech, T.R. (1986) Cell 44, 207–210.
[38] Newman, A.J. and Norman, C. (1992) Cell 68, 743–754.
[39] Weiner, A.M. (1993) Cell 72, 161–164.
[40] Senapathy, P. (1988) Proc. Natl. Acad. Sci. USA 85, 1129–1133.