Insights on the Origin and Evolution of Introns

Maria Dulcetti Vibranovski
Ludwig Institute for Cancer Research,
Sao Paulo Branch, Sao Paulo, Brazil.
Ph.D Program, Departamento de Bioquímica,
IQ-USP, Sao Paulo, Brazil.
maria@compbio.ludwig.org.br

Sandro José de Souza
Ludwig Institute for Cancer Research,
Sao Paulo Branch, Sao Paulo, Brazil.

Abstract

Since the discovery of introns (intervening non-coding sequences in a gene), many questions about their origin have been raised. Currently, the main questions are: why introns exist in eukaryotic organisms and not in prokaryotes and when and how did they originate? Mainly, there are two opposing hypotheses explaining the origin of introns: “introns-early” and “introns-late”. The first suggests that introns and exons already existed in the first genes and were lost later in the bacteria lineage. The opposing hypothesis, introns-late, assumes that introns were inserted late in evolution, in eukaryotic organisms only. Here, we review this issue considering characteristics related to the phenomenon of exon-shuffling. As a consensus between the two hypotheses has emerged in the form of a synthetic theory of intron evolution, we also discuss the perspectives in the field.

Key words
intron phase, exon-shuffling, symmetry of exons, modules.

Origin and evolution of introns

Introns are non-coding sequences of DNA located between exons (coding sequences) in a gene. After transcription, they are removed from mature messenger RNA prior to translation. Thus, the protein formed is the product of

fusion of translated exons (1). Since the discovery of introns, much has been discussed about their origin (2,3). These studies aimed at answering questions such as why introns are present in eukaryotes and absent in prokaryotes and how they originated (4). There are two main hypotheses about their origin, namely: “introns-early” and “introns-late”. The first one, “introns-early”, assumes that introns were present in the genes of the progenote (the common ancestor of prokaryotes and eukaryotes) and were lost in prokaryotes by a process of “streamlining” – reducing genome size due to a selective pressure for fast DNA replication (2,4,5). In contrast, the “introns-late” hypothesis argues that introns were inserted late in evolution, after the eukaryote-prokaryote divergence (3,6,7,8).

Sequences from different genes may suffer recombination during cell division to form new sequences, either functional or not. Recombination within introns between different genes allows and increases the frequency of exchanging *complete* exons and, consequently, increases the probability of forming new functional proteins (1). This phenomenon has been named exon-shuffling by Gilbert (1). As exon-shuffling has only been possible since the appearance of introns, the two hypotheses on the origin of introns disagree about the period in evolution in which this phenomenon arose. “Introns-early” assumes that exon-shuffling has occurred since the progenote and played an important role in the creation of new proteins early in evolution (2,4,9). In contrast, “introns-late” argues that exon-shuffling is a recent phenomenon important in protein variability in complex eukaryotes only (7,10,11).

Analysis of intronic and exonic features

The two hypotheses are based on properties of introns and exons observed in nature. However, what are the characteristics that can be analyzed in order to test the two hypothesis? One may consider comparison of intron sequences among evolutionarily distant species as a feature to be studied. However, as intron sequences have a higher mutation rate, the identity between sequences of distant species is undetectable (12). On the other hand, positions of introns, *i.e.*, the corresponding localization of the intron

projected in the amino acid sequence of the protein encoded by exons, could be conserved among species. If introns were present in the progenote, their position in a gene should be conserved among different species. On the other hand, if they were (randomly) inserted in eukaryotic sequences during evolution, one would not expect to find matching positions in the same genes between different organisms. Using triosephosphate isomerase, it has been first shown that certain intron positions are conserved between organisms (13). Later, it was demonstrated that other introns were inserted in different positions during eukaryotic evolution (14,15). These last data seemed to be an argument in favor of the "introns-late" hypothesis. However, "introns-early" argued that introns are able to shift their position during evolution (a process called "intron sliding"). Therefore, they questioned if correspondence of introns' positions among species could be a reliable feature, even after Stolzfus and colleagues (16) showed that intron sliding could not explain the difference in 205 positions among species in five proteins studied.

Other features, such as intron phase and symmetry of exons, have been intensively studied throughout this debate. These features are deeply related to the phenomenon of exon-shuffling. Intron phase is defined as the position where an intron lies in a codon (set of three nucleotides that encode an amino acid). Phase zero introns lie between codons; phase one, between the first and the second nucleotides of the codon; phase two, between the second and the third nucleotides. Another trait, exon symmetry, is the correlation between introns that flank the same exon. Symmetric exons are those flanked by introns of the same phase, *e.g.*, 0-0, 1-1 or 2-2. The relation between these characteristics and exon-shuffling is based on the "success" of the phenomenon. Exons can be exchanged among genes, but this does not mean that it will yield a functional protein. Shuffling of exons flanked by introns of different phases has less probability to generate functional genes due to changes in the reading frame as demonstrated in Figure 1. The insertion of a symmetric exon (0-0) from gene A in a sequence B by recombination of introns of the same phase (0) produces a different protein containing all exons from gene B, plus the exact amino acid sequence from

the inserted exon of the gene A. Now, imagine that the inserted exon was still symmetric although flanked by a different type of intron (phase one, for example). The new protein formed would contain the exon with the same nucleotide composition from gene A, except that it would be in a different reading frame. Moreover, if the exon inserted was not symmetric, the protein resulting from shuffling would not only have a modified reading frame of the inserted exon, but the reading frame of all subsequent exons would be changed too. Thus, the probability of forming a functional protein by exon-shuffling decreases from exons flanked by introns of the same phase to non-symmetric exons. As symmetric exons always contain nucleotides in multiples of three, their insertion in a sequence never changes the reading frame of the subsequent exons.

According to "introns-early", symmetric exons should be present in excess even in ancient proteins (the ones found in both prokaryotes and eukaryotes) because exon-shuffling is assumed to be an ancient phenomenon (17). Contrarily, "introns-late" does not predict such an excess as it believes that introns were inserted randomly in exon sequences that do not correspond to independent units (8). In addition, as shuffling of symmetric exons of the same phase has a higher probability of producing functional proteins, "introns-early" expects that introns of a certain phase will have a higher frequency throughout the population of introns.

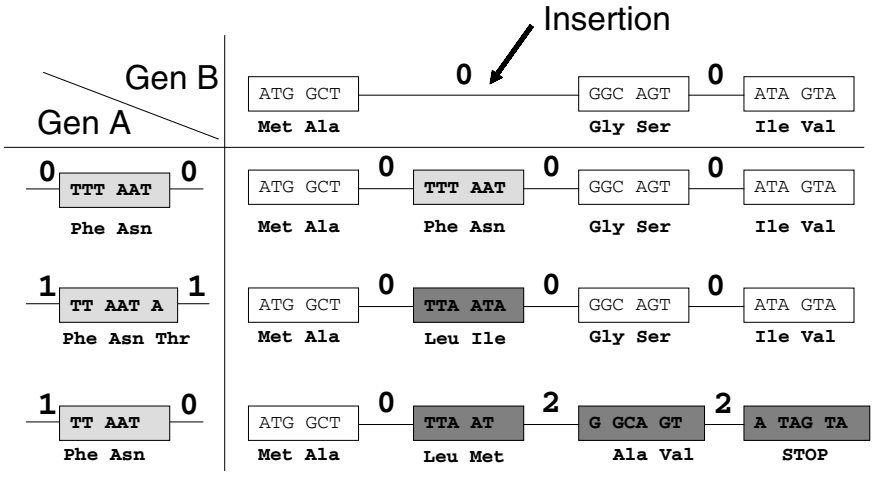

Figure 1 – Scheme of Exon-shuffling – An exon from gene A is inserted into gene B and formed a new protein. Boxes represent exons and lines represent introns. Numbers above introns correspond to intron phases.

Inside each box are the exon nucleotide sequences. Amino acid sequences encoded by each exon are represented below. Light gray boxes correspond to exons containing the same amino acid sequence as the original one (from gene A). Dark gray boxes correspond to exons that had their reading frame changed after insertion. STOP means a premature stop codon generated by change of the reading frame.

Indeed, Fedorov and colleagues (18) found a non-uniform distribution of intron phases. Their data, confirmed later by Long *et al.* (17) with a larger dataset, showed a distribution biased to phase zero introns. The important contribution of Long and colleagues was that they made the whole analysis of distribution of intron phases and symmetric exons within ancient coding regions (ACRs), *i.e.*, linear regions homologous between eukaryotes and prokaryotes. They found phase zero as the most abundant intron phase (~50%), followed by phase one (~30%) and phase two (~20%) (17,19). The analysis of ACRs enabled them to observe the signal of exon-shuffling in ancient periods. In relation to the distribution of symmetric exons, they found a depletion of non-symmetric exons and an excess of symmetric exons in all genes analyzed and in the ACRs separately (17,19). These findings were important landmarks in favor of the "introns-early" hypothesis.

One of the points raised by "introns-late" regarding the biased distribution of intron phases is a supposed non-random insertion pattern of introns (20). Their argument is based on the existence of proto-splice sites, which are consensus sequences in coding regions recognized for intron insertion. This argument was kept alive until Long and colleagues (21,22) demonstrated in two different ways that the distribution of proto-splice sites observed today does not explain the biased intron phase distribution, neither the correlation of intron phases found in nature. Their first approach consisted of dicodon analysis in six model species looking for proto-splice sites and calculating what would be the distribution of the intron phases if proto-splice sites were used to insert introns (21). Their second contribution utilized all proto-splice sites in all coding sequences to identify pseudo-exons, which are virtual exons delimitated by two consecutives proto-splice sites. That way, the authors computed the correlation between intron phases

that flank these pseudo-exons. Using this information, they calculated the excess and depletion of symmetric pseudo-exons and compared them to the distribution of excess of real symmetric exons (22). Both approaches could not relate virtual distributions based on proto-splice sites to real distributions found in nature, which rules out the “introns-late” argument of a biased pattern of intron insertion due to the existence of proto-splice sites.

Exons as independent units

Another issue discussed in the debate is the correlation of exons with independent protein units. For “introns-late”, exons do not necessarily correspond to protein units as introns were inserted in former continuous sequences of exons. Blake (5) was the first to mention that if new combinations of exons are able to produce new proteins, exons must correspond to independent units – domains or other structural units, able to properly fold independently. Modules are units defined by a protein backbone that fits in a sphere of a given diameter. Go (23) observed the correspondence of exons from a hemoglobin gene to modules of 28 Å of diameter. This paper was followed by a series of other works that found correspondence between compact structural units (modules) and exons using other proteins, such as: lysozyme (24), triosephosphate isomerase (13), glyceraldehyde phosphate dehydrogenase (25) and others.

The central argument of “introns-late”, showing that a general correspondence between structural units of proteins and exons could not be shown, was based on results obtained by another group testing this correspondence (8). They analyzed 62 intron positions in four different proteins (alcohol dehydrogenase, globins, pyruvate kinase and triosephosphate isomerase) and did not find evidence of correlation among intron positions and boundaries of secondary structures, modules or globular domains. Hence, there was no place in the debate anymore for other individual examples of correlation among exons and structurally independent protein units. The discussion rather required a large-scale analysis supported by statistical significance.

In 1996, the number of sequences stored in public databases (especially those including protein structures) was large enough to support this analysis. De Souza and colleagues (26) demonstrated a correlation between the positions of introns and module boundaries in a set of 32 ancient proteins using the method of Go (23) to construct the modules. The analysis from de Souza *et al.* (26) was also possible because of the development of an algorithm that allowed the automated study of several proteins at once. They first found significant correlation using modules of 28 Å of diameter. Later, varying the modules' size, they showed two other peaks of significant correlation of intron positions and module boundaries at 21 Å and 32 Å (26). A few years later, the same group using a sample of 43 proteins was able to detect that this statistical signal was due to phase zero introns (19). Furthermore, a recent work with a larger sample verified that the three peaks of module sizes correlated to phase zero introns' positions merged in just one (28 Å) for ancient proteins (27). Moreover, no correlation was found for this phase of introns and module boundaries in non-ancient proteins, which is of great significance.

The module correlation was not the only example of exons as independent protein units. Domains were also studied in respect to their shuffling in the human genome (28). First, the authors observed excess of symmetrical intron phase combinations in the boundaries of domains whereas exons flanked by non-boundary introns showed no excess of symmetry. Concerning exon-shuffling in general, this result indicates that the phenomenon involves domains as units, where exons and sets of exons could correspond to the units shuffled. However, their most important contribution to the "introns-early" X "introns-late" debate was the distribution of symmetric exons or set of exons investigated taking into account the "age" of domains. When the authors considered only new domains (those present only in metazoans), 1-1 domains were found in significant excess. However, this excess disappeared when only old domains were analyzed and 0-0 symmetrical domains tended to be over-represented.

These two forms of correlation of exons and independent units composed an important landmark in favor of the "introns-early" hypothesis. However, a consensus between

the two hypotheses in the debate was about to be constructed.

Consensus in the debate

"The synthetic theory of intron evolution", cited by de Souza (29), has arisen from joining ideas and arguments from both hypotheses. Actually, the first paper toward a resolution of the discussion claimed that introns found nowadays are the result of ancient introns (present in the progenote) and new inserted ones (19). This argument is based on the observation of excess of symmetric exons, excess of phase zero in the distribution of introns and the tendency of phase zero introns to be located in module boundaries. The authors suggested that equal numbers of introns of all three phases had been added during evolution (19). They claimed that within the distribution of introns (56% are phase zero, 23% are phase one and 21% are phase two) 35% of all introns (only taken from the set of phase zero introns) should be considered ancient because this percentage represents the excess of phase zero introns in relation to the other phases.

In the mean time, discoveries have been made in molecular biology that, like everything in science, somehow contributed to the discussion. As more genomes have been sequenced and the amount of sequences deposited has increased, a large number of introns has been identified. Less complex eukaryotic organisms, as *Giardia lamblia* and *Carpediemonas membranifera*, were discovered to contain spliceossomal introns (30,31). Moreover, group II introns known as a possible precursor of spliceossomal introns were found in an archaea genome (32) which is evidence against Cavalier-Smith's argument (3) ("introns-late"). As archaea is considered the host genome in the symbiosis between a eubacteria and an archaea cells that formed the primitive eukaryotic cells, the latter finding means that nuclear eukaryotic introns have other possible origins besides insertion from eubacteria's group II introns (the symbiotic cell). They could have evolved from archaea's group II introns.

The synthetic theory of intron evolution combines the concepts of "introns-early" and "introns-late" hypotheses by the acceptance that introns known today have two origins

(ancient and recent by insertion) (23). The arguments and the data supporting the view that the eukaryotic genome consists, at least partially, of ancient introns cannot be ruled out.

Perspectives

The perspectives for the field of origin and evolution of introns are encouraging. Although the debate seems to have achieved a consensus, it is quite clear that there is still a great amount of work to be done. Everything discovered in the last two decades and the reviews about the issue in the last few years (29,33,34) are a source for further progress in the field. Exon-shuffling patterns in ancient proteins would be one of the first phenomena to be explored. The synthetic theory of intron evolution predicts reuse of ancestral modules by exon-shuffling to form new genes. As the evolutionary distance might be too large to show sequence similarity between shuffled exons, we may look for structural similarity at the protein level. Moreover, studying the structures encoded by these exons and the pattern of their flanking introns would be interesting to consolidate the theory of intron evolution.

Another perspective, still concerning exon-shuffling, is to define the exons in a protein that might have been shuffled recently and in ancient times. This could be done by analysis of the patterns and distribution of symmetric exons in proteins. Ancient proteins found today should be the product of modern and ancient exon-shuffling and there might exist a pattern of distribution of both such events within these proteins.

In addition, exon-shuffling is a phenomenon that can be compared with alternative use of exons (a type of alternative splicing). The latter is the process by which non-consecutive exons are spliced together to form different mature mRNAs. Hence, multiple protein isoforms are generated from a single gene (35). These two processes are comparable, as both are capable to produce variability in terms of proteins. Exon-shuffling increases variability during evolution and alternative splicing does so during the cell's life. Hence, they may present common exon and intron features or even common origin. For this reason, the alternative use of exons

can be analyzed in the light of the proteins module structure and intron phase correlation and be compared to possible patterns related to exon-shuffling.

For all this, intron origin and evolution is still an exciting theme in biology. The research possibilities in the field are wide open. The advance in the studies of genomics and proteomics of the last decade increased the amount of data available in the public databases. Hence, bioinformatics, one more time, can be used as a tool to solve questions concerning biology and evolution.

Acknowledgments

The authors would like to thank Natanja Kirschbaum-Slager and Noboru Jo Sakabe for careful reading of the manuscript. MDV is supported by FAPESP fellowship.

References

[1] Gilbert, W., 1978. Why genes in pieces? Nature 271: 501.

[2] Gilbert, W., 1987. The Exon Theory of Genes. Cold Spring Harb Symp. Quant. Biol. 52: 901-905.

[3] Cavalier-Smith, T., 1991. Intron phylogeny: a new hypothesis. Trends Genet. 7: 145-148.

[4] Doolittle, W. F., 1978. Genes in pieces: where they ever together? Nature 272: 581-582.

[5] Blake, C.C.F., 1978. Do genes-in-pieces imply proteins-in-pieces? Nature 273: 267.

[6] Roger, J., 1985. Exon-shuffling and intron insertion in serine protease genes. Nature 315: 458-459.

[7] Palmer, J.D. and Logsdon, J.M., Jr., 1991. The recent origins of introns. Curr. Opin. Genet. Dev. 1: 470-477.

[8] Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, J.M., Jr. and Doolittle, W.F., 1994. Testing the Exon Theory of Genes: The Evidence from Protein Structure. Science 265: 202-207.

[9] Blake, C.C.F., 1979. Exons encode protein functional units. Nature 277: 598.

[10] Doolittle, W.F. and Stoltzfus, A., 1993. Genes-in-pieces revisited. Nature 361: 403.

[11] Patthy, L., 1999. Genome evolution and the evolution of exon-shuffling – a review. Gene, 238: 103-114.

[12] Graur, D., and Li, W., 2000. Fundamentals of Molecular Evolution. Chapter four. Sinauer Associates Inc., Massachusetts.

[13] Gilbert, W., Marchionni, M and McKnight, G., 1986. On the Antiquity of Introns. Cell, 46:151-152.

[14] Kwiatowski, J., Krawczyk, M., Kornacki, M., Bailey, K. and 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.

[15] Logsdon, J.M.,Jr, Tyshenko, M.G., Dixon, C., Jafari, J.D., Walker, V.K. and Palmer, J.D., 1995. Seven newly discovered intron position in the triose-phosphate isomerase gene: Evidence for the introns-late theory. Proc. Natl. Acad. Sci. USA 92: 8507-8511.

[16] Stoltzfus, A., Logsdon, J.M.,Jr., Palmer, J.D., Doolittle, W.F., 1997. Intron “sliding” and the diversity of introns positions. Proc. Natl. Acad. Sci. USA, 94: 10739-10744.

[17] Long, M., Rosenberg, C., and Gilbert, W., 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. Proc. Natl. Acad. Sci. USA 92: 12495-12499.

[18] Fedorov, A., Suboch, G., Bujakov, M. and Fedorova L., 1992. Analysis of nonuniformity in intron phase distribution. Nucl. Acids Res. 20: 2553-2557.

[19] De Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S. and Gilbert, W., 1998. Toward a resolution of the introns early/late debate: Only phase 0 introns are

correlated with the structure of ancient proteins. Proc. Natl. Acad. Sci. USA 95: 5094-5099.

[20] Dibb, N.J. and Newman, A.,J., 1989. Evidence that introns arose at proto-splice sites. Embo J. 8: 2015-2021.

[21] Long, M., de Souza, S.J., Rosenberg, C. and Gilbert, W., 1998. Relation between “proto-splice sites” and intron phases: Evidence from dicodon analysis. Proc. Natl. Acad. Sci. USA 95: 219-223.

[22] Long, M. and Rosenberg, C., 2000. Testing the “Proto-splice Sites” Model of Intron Origin: Evidence from Analysis of Intron Phase Correlations. Mol. Biol. Evol. 17: 1789-1796.

[23] Go, M., 1981. Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature 291: 90-92.

[24] Go, M., 1983. Modular structural units, exons, and function in chicken lysozyme. Proc. Natl. Acad. Sci. USA 80: 1964-1968.

[25] Long, M., de Souza, S.J., Rosenberg, C. and Gilbert, W., 1996. Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. Proc. Natl. Acad. Sci. USA 93: 7727-7731.

[26] De Souza, S.J., Long, M., Schoenbach L., Roy, S.W. and Gilbert, W., 1996. Intron positions correlate with modules boundaries in ancient proteins. Proc. Natl. Acad. Sci. USA 93: 14632-14636.

[27] Fedorov, A., Cao, X., Saxonov, S., de Souza, S.J., Roy, S.W. and Gilbert, W., 2001. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. Proc. Natl. Acad. Sci. USA 98: 13177-13182.

[28] Kaessmann, H., Zöllner, S., Nekrutenko, A. and Li, W., 2002. Signatures of Domain Shuffling in the Human Genome. Genome Res. 12: 1642-1650.

[29] De Souza, S.J., 2003. The emergence of a synthetic theory of intron evolution. Genetica 118: 117-121.

[30] Nixon, J.E.J., Wang, A., Morrison, H.G., McArthur, A.G., Sogin, M.L., Loftus, B.J. and Samuelson, J., 2002. A spliceosomal intron in *Giardia lamblia*. Proc Natl Acad Sci USA. 99: 3701-3705.

[31] Simpson, A.G., MacQuarrie, E.K. and Roger, A.J., 2002. Eukaryotic evolution: early origin of canonical introns. Nature 419: 270.

[32] Dai, L. and Zimmerly, S., 2003. ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. RNA 9 :14-19.

[33] Roy, S.W., 2003. Recent evidence for Exon Theory of Genes. Genetica 118: 251-266.

[34] Saxonov, S. and Gilbert, W., 2003. The universe of exons revisited. Genetica 118: 267-278.

[35] McKeown, M., 1992. Alternative mRNA splicing. Annu. Rev. Cell. Biol. 8, 133-155.