CMLS, Cell. Mol. Life Sci. 55 (1999) 3–6
1420-682X/99/010003-04 $ 1.50 + 0.20/0

CMLS Cellular and Molecular Life Sciences

Visions & Reflections

The enigma of intron origins

A. Rzhetsky[a,*] and F. J. Ayala[b]

[a]Columbia Genome Center, Columbia University, 1150 St. Nicholas Avenue, Unit 109, New York (New York 10032, USA), Fax +1 212 304 5515, e-mail: andrey@genome2.cpmc.columbia.edu
[b]275 Leonard St. #2 NW, New York (New York 10013, USA), e-mail: ze@interport.net

Received 5 October 1998; received after revision 3 November 1998; accepted 3 November 1998

Introns-early: an elegant theory in conflict with the data

The introns-early hypothesis for the origin of the genes-in-pieces mosaic architecture of eukaryotic genomes [1] is compelling because it is simultaneously a theory on the origin of extant protein diversity and because it ascribes a biological function to the introns themselves. The theory is built on the following suppositions. First, nuclear introns were present in the common ancestor of all presently living organisms. Accordingly, the machinery for splicing introns from messenger RNA (mRNA) is ancient. Second, natural selection subsequently removed all introns from the bacteria in the interests of metabolic efficiency. Third, the original evolutionary role of introns was to link small pieces of nucleic acid-encoded modules of 15–20 amino acids. Random recombinational shuffling of these modules has resulted in the diverse functional and structural repertoire of current proteins. Finally, the theory postulates that introns may occasionally become displaced by 1–15 nucleotides along the length of the genes in which they reside.

In this model, intron displacement, or sliding, is critically important for explaining the present distribution of introns among orthologous and paralogous genes [2]. The process by which introns slide is not clear, having been posited as some notional combination of transcription, splicing, reverse splicing, reverse transcription and finally recombination [3]. Experimental observation of this process has eluded a vigorous search [4]. Furthermore, the invocation of intron sliding creates a contrariety for the introns-early theory: while the process may account for the wandering distribution of introns, it calls into question one of the theory's central pillars – the validity of reported correlations between the positions of introns within coding regions and the boundaries of protein modules [1, 2, 5, 6]. In fact, whether such correlations exist continues to be a matter of intense debate [7–9].

Exon shuffling is also problematic for the introns-early model. A pronounced disparity exists between the definition of protein modules used by introns-early theorists and that used by structural biologists. The latter define a protein structural module as a conserved compact globular component [10–12], allowing the number of amino acids and the physical dimensions to vary widely. Individual structural modules are usually associated with specific protein functions. By contrast, introns-early studies purporting a correlation between intron boundaries and module boundaries rely on modules defined as contiguous sequences of amino acids that compactly fit into spheres of 28 Å [5, 6, 13]. This definition tends to generate modules that ineptly cut across structural elements such as α-helices and β-sheets, and furthermore tends to generate modules that reside within larger functional structures. As yet, shuffling has only been inferred between modules as defined by the structural biologists [10, 11, 14].

The lack of experimental and statistical support is unfortunate for a theory that would otherwise so elegantly explain the peculiarities of eukaryotic gene structure and evolution.

* Corresponding author.

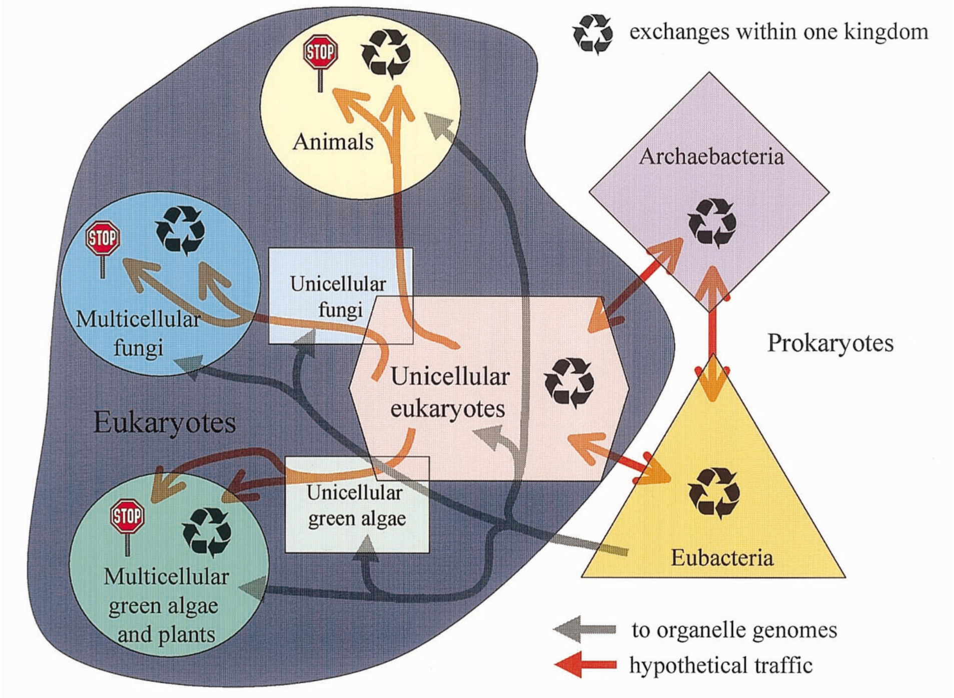

Figure 1. Hypothetical transfers of transposable elements within and between kingdoms. Under the reservoir hypothesis unicellular organisms continuously supply multicellular eukaryotes with new transposons, most of which will eventually die and, in some instances, become introns.

Introns-late: molecular parasites

The alternative theory, introns-late [10, 15], proposes the following. First, extant introns were inserted into their present locations within nuclear genes. Second, the spliceosome is an evolutionary invention exclusive to the eukaryotes; bacteria never possessed spliceosomal introns. Third, introns have not been widely useful in the generation of protein diversity, although exon shuffling has indeed occurred occasionally [11]. Lastly, the notional sliding of introns is not required to explain their distributions.

The introns-late theory portrays introns as molecular parasites, mildly deleterious in that they consume metabolic resources of organisms that must remove them from their transcripts. Introns may have first appeared through the coevolution of early eukaryotic genomes and transposon-like group II introns [16], although the evolutionary emergence of introns and the splicing machinery is obscure under the introns-late as well as introns-early theories.

The pattern of intron boundaries within coding regions is an interesting feature of intron distribution not well explained by the introns-late theory. A significant excess of phase 0 introns (in which the intron boundaries lie between codons) has been observed relative to phase 1 and 2 introns (in which introns break up codons, lying between the first and second codon positions respectively). The asymmetry of intron phase distribution can plausibly be explained by an asymmetric distribution of nucleotide combinations preferential for intron insertion [17], although it is more often cited as the expected outcome under the introns-early theory [18], whereby genes were formed by combinations of ancient functional proto-domains.

Nevertheless, a considerable body of data supports the conjecture that spliceosomal introns have been gained as well as lost along many evolutionary lineages [19–22], intron gain appearing to be the more frequent event. This provides compelling support for a theory that increasingly appears more credible than the competing theory.

The reservoir hypothesis: transposons in circulation

Support for the introns-late insertional theory has recently and unexpectedly come from the immunoglobulin system. The mammalian RAG1 and RAG2 proteins, which normally facilitate VDJ recombination in immunoglobulin minigenes, are now known to have transposase activity [23]. Their mechanism of action is very close to those of Tn7 and Tn10 in bacteria, P-elements in *Drosophila*, and Tc1/mariner elements in eukaryotes. Transposons, like viruses, are molecular parasites that prey on the genomes of eukaryotes while occasionally serving as the proximal cause of genic rearrangements [24]. The fragmentation of immunoglobulin loci into minigenes might be the result of ancient transpositional insertions into the coding regions of ancestral genes. These spacers between minigenes might then be sister structures to nuclear introns.

If so, ancient eukaryotic genomes must have been exposed to transpositional meteor showers resulting in the widespread distribution of elements that eventually either became junk DNA in the case of introns, or acquired functional utility in the case of immunoglobulin spacers. Under the mantles of introns-early and transposons as genomic infiltrators, we present here a new 'reservoir' hypothesis on transposon circulation in nature. We make the following suppositions.

First, transposons are mostly limited to inhabiting unicellular organisms, frequently shuttling between different host species, and are subject to intensive stabilizing selection. Transposons multiply and diversify within this reservoir of unicellular organisms, occasionally becoming transferred to multicellular organisms.

Second, the genomes of multicellular organisms are exposed to transpositional assaults since they regularly come into direct physical contact with unicellular organisms (see fig. 1), often by intracellular infection. A successful inheritable integration of a transposon into the genome of a multicellular organism can ensue only when the transposition occurs within a germline cell.

Third, most transposons that succeed in infecting the genomes of multicellular eukaryotes will not generally be further transmitted from one to other multicellular organism, although certain instances of such horizontal transfer are known (the infection of P-elements from *Drosophila simulans* to *D. melanogaster* [25, 26], for example). Nearly all transposons that enter the genomes of multicellular eukaryotes are doomed to eventual death: successful horizontal transmission is too infrequent to save them from degeneration. Most transposons that succeed in integrating themselves into the germline will have one of two fates: either they lose their functionality due to the accumulation of deleterious substitutions [27–29], and some of these will become introns, or they evolve some function that is useful to the host, as in the case of the immunoglobulin spacers.

The reservoir hypothesis yields a prediction that can be experimentally tested. If correct, all transposons that are found in multicellular organisms will have well-conserved functional progenitors in unicellular organisms. As entire genome sequences for diverse species are elucidated at an increasing rate, we will be able to study the function and evolution of transposable elements and self-splicing introns in the prokaryotes and to look for possible remnants of lost spliceosomal introns. Experimental and observational data may thus contribute to a final resolution of the enigma of intron origins.

1 Gilbert W., de Souza S. J. and Long M. (1997) Origin of genes. Proc. Natl. Acad. Sci. USA **94:** 7698–7703

2 Rzhetsky A., Ayala F. J., Hsu L. C., Chang C. and Yoshida A. (1997) Exon/intron structure of aldehyde dehydrogenase genes supports the 'introns-late' theory. Proc. Natl. Acad. Sci. USA **94:** 6820–6825

3 Cerff R. (1995) The chimeric nature of nuclear genomes and the antiquity of introns as demonstrated by GAPDH gene system. In: Tracing Biological Evolution in Protein and Gene Structures, pp. 205–227, Gö M. and Schimmel P. (eds), Elsevier, New York

4 Stoltzfus A., Logsdon J. M. Jr., Palmer J. D. and Doolittle W. F. (1997) Intron 'sliding' and the diversity of intron positions. Proc. Natl. Acad. Sci. USA **94:** 10739–10744

5 de Souza S. J., Long M., Schoenbach L., Roy S. W. and Gilbert W. (1997) The correlation between introns and the three-dimensional structure of proteins. Gene **205:** 141–144

6 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 zero introns are correlated with the structure of ancient proteins. Proc. Natl. Acad. Sci. USA **95:** 5094–5099

7 Logsdon J. M. Jr., Tyshenko M. G., Dixon C., D-Jafari J., Walker V. K. and 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

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

9 Burke J., Hwang P., Anderson L., Lebo R., Gorin F. and Fletterick R. (1987) Intron/exon structure of the human gene for the muscle isozyme of glycogen phosphorylase. Proteins **2:** 177–187

10 Patthy L. (1996) Exon shuffling and other ways of module exchange. Matrix Biol. **15:** 301–310; discussion 311–312

11 Patthy L. (1991) Exons – original building blocks of proteins? Bioessays **13:** 187–192

12 Doolittle R. F. and Bork P. (1993) Evolutionarily mobile modules in proteins. Sci. Am. **269:** 50–66

13 Gö M. and Nosaka M. (1987) Protein architecture and the origin of introns. Cold Spring Harb. Symp. Quant. Biol. **52:** 915–924

14 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

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

16 Cech T. R. (1986) The generality of self-splicing RNA: relationship to nuclear mRNA splicing. Cell **44:** 207–210

17 Dibb N. J. and Newman A. J. (1989) Evidence that introns arose at proto-splice sites. EMBO J. **8:** 2015–20121

18 Long M., de Souza S. J., Rosenberg C. and Gilbert W. (1998) Relationship between 'proto-splice sites' and intron phases: evidence from dicodon analysis. Proc. Natl. Acad. Sci. USA **95:** 219–223

19 Frugoli J. A., McPeek M. A., Thomas T. L. and McClung C. R. (1998). Intron loss and gain during evolution of the catalase gene family in angiosperms. Genetics **149:** 355–365

20 Logsdon J. M. and Stoltzfus A. (1998) Recent cases of spliceosomal intron gain? Curr. Biol. **8:** R560–R563

21 Kwiatowski J., Skarecky D. and 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

22 Tarrio R., Rodriguez-Trelles F. and Ayala F. J. (1998) New *Drosophila* introns originate by duplication. Proc. Natl. Acad. Sci. USA **95:** 1658–1662

23 Hiom K., Melek M. and Gellert M. (1998) DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell **94:** 463–470

24 King C. C. (1985) A model for transposon-based eucaryote regulatory evolution. J. Theor. Biol. **114:** 447–462

25 Houck M. A., Clark J. B., Peterson K. R. and Kidwell M. G. (1991) Possible horizontal transfer of *Drosophila* genes by the mite *Proctolaelaps regalis*. Science **253:** 1125–1128

26 Clark J. B., Maddison W. P. and Kidwell M. G. (1994) Phylogenetic analysis supports horizontal transfer of P transposable elements. Mol. Biol. Evol. **11:** 40–50

27 Smit A. F. and Riggs A. D. (1996) Tiggers and DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA **93:** 1443–1448

28 Adey N. B., Tollefsbol T. O., Sparks A. B., Edgell M. H., and Hutchison C. A. III (1994) Molecular resurrection of an extinct ancestral promoter for mouse L1. Proc. Natl. Acad. Sci. USA **91:** 1569–1573

29 Capy P., David J. R. and Hartl D. L. (1992) Evolution of the transposable element mariner in the *Drosophila melanogaster* species group. Genetica **86:** 37–46