Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 1356–1357, February 1998

Commentary

Lightning strikes twice: Intron–intein coincidence

*Victoria Derbyshire and Marlene Belfort**

Molecular Genetics Program, Wadsworth Center, New York State Department of Health, P.O. Box 22002, Albany, NY 12201-2002

"Multiple introns, and the prospect that these occur within several genes in the same metabolic pathway, suggest a regulatory role for splicing . . . " (1). That statement was made more than a decade ago by our colleagues and ourselves in reference to the phage T4 group I introns, all three of which reside in genes involved in nucleotide metabolism. The intervening years were distinguished by a striking absence of any demonstration of a regulatory role for these introns. Furthermore, they seem not to confer any selective advantage to T4 phage, and, despite the discovery of more introns in genes of DNA metabolism, the question of "why exclusively in *those* genes?" has been all but forgotten. Now we are faced with a report in this issue of the *Proceedings* of an unrelated phage and a different type of splicing element, an intein; not only does this element coexist with a group I intron in the same gene but also the gene is one of nucleotide metabolism (2). With lightning apparently striking twice in the same place, we are again forced to confront our doubts about pure chance. What might account for the colocalization of two different types of intervening sequence in the same gene on the same metabolic pathway?

The now familiar group I introns self-splice at the RNA level (3), whereas inteins, which are in-frame protein fusions, self-splice at the protein level (ref. 4; Fig. 1). The first inteins were described just a few years ago, and immediately examples emerged in all three biological kingdoms, the archaea, bacteria, and eukarya (5–7). The recent identification of more than 50 inteins, mostly by sequence comparisons (refs. 4, 8, and 9; New England Biolabs Intein Database Intein Registry at http://www.neb.com; S. Pietrokovski at http://www.blocks.fhcrc.org/~pietro/inteins) has shown them to be present in a wide range of organisms. However, until now, none had been found in bacteriophage. Until, that is, the above-mentioned report, which identifies an intein and a group I intron in the gene for a putative ribonucleotide reductase subunit (the *bnrdE* gene) in the *Bacillus subtilis* bacteriophage SPβ (2).

Lazarevic *et al.* (2) also identified a group I intron in the neighboring *bnrdF* gene, encoding a second putative subunit of SPβ ribonucleotide reductase. The *bnrdE* and *bnrdF* introns are in the same general family as the three group I introns of phage T4 (1, 10) and the introns in *Bacillus* phages β22 (11) and SPO1 and three of its close relatives (12). Remarkably, all but one of the introns and the single intein so far identified in bacteriophage are located in genes involved in DNA metabolism (10–13). The SPO1 intron is in the DNA polymerase gene, whereas β22 and T4 have introns in their thymidylate synthase genes. The remaining T4 introns are in *nrdB*, which encodes a small subunit of ribonucleotide reductase, and *sunY*, renamed *nrdD*, encoding an anaerobic ribonucleotide reductase. Given the small number of group I introns so far discovered in bacteriophage genes and the apparently very low occurrence of inteins in these organisms, these findings are provocative indeed.

Although self-splicing introns in bacteria and lower eukaryotes are not confined to genes of DNA metabolism, but rather are situated in a different spectrum of loci, including tRNA, rRNA, and energy metabolism genes (14), the genes playing host to inteins are once again heavily biased toward DNA metabolism (refs. 4, 8, and 9; New England Biolabs Intein Database Intein Registry at http://www.neb.com; S. Pietrokovski at http://www.blocks.fhcrc.org/~pietro/inteins). Approximately 70% of known inteins are confined to such functions. They include DNA polymerases, helicases, gyrases, RecA recombinase, and, once again, ribonucleotide reductases, both aerobic and anaerobic.

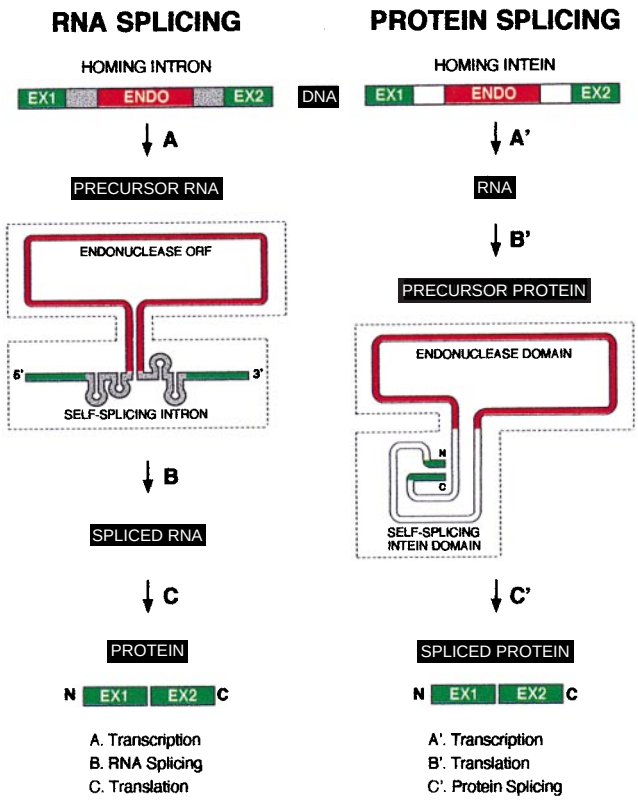

FIG. 1. Comparison of RNA and protein splicing. EX1 and EX2 (green) represent exons (*Left*) or exteins (*Right*), which flank the intron or intein, respectively. Mobile group I introns and inteins are bipartite elements. The intron consists of the catalytic core (shaded) and the endonuclease (ENDO) open reading frame (ORF, red). The intein consists of the protein splicing domain (white) and the endonuclease domain (red) (8, 19, 20).

In addition to splicing at the RNA or protein levels, many group I introns and inteins are mobile genetic elements at the DNA level, by virtue of endonucleases encoded within them (Figs. 1 and 2). These endonucleases cleave intron-minus or intein-minus alleles of their cognate genes, initiating a unidirectional gene conversion event that results in copying of the intein or intron DNA (Fig. 2). There has been much discussion of the evolution of mobile group I introns and inteins. It has been argued that endonuclease genes are the ancestral mobile elements that

*To whom reprint requests should be addressed. e-mail: marlene.belfort@wadsworth.org.

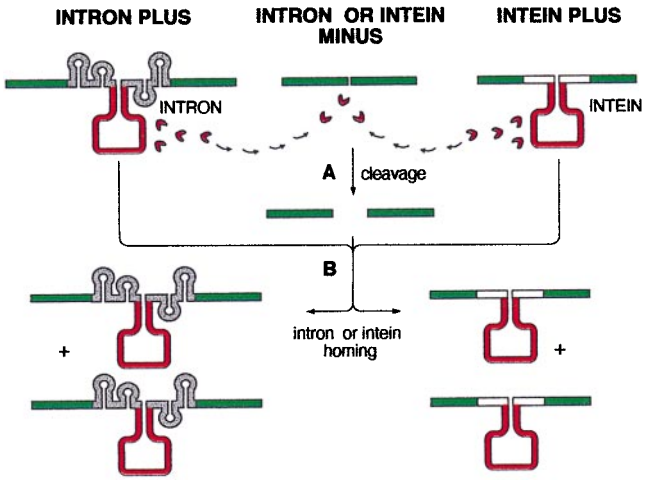

FIG. 2. Intron and intein homing. (*A*) The intein or intron-encoded endonuclease (small red symbols) cleaves the intein-minus or intron-minus allele. (*B*) Double-strand break repair of the recipient results in duplication of the intein or intron DNA. Conventions as in Fig. 1.

colonized ancient introns and inteins (15–20). Endonuclease invasion of essential genes would naturally be lethal because it would abolish gene expression. However, insertion into DNA encoding introns or inteins would afford endonuclease genes safe haven because of the ability of the intron or intein to splice and therefore to preserve expression of the host gene. The intron or intein in turn acquires mobility. While the presence of nonmobile introns in the same genes of widely differing organisms is consistent with their being of ancient lineage, the mobility of the endonuclease-containing elements allows for lateral transfer between genes and organisms in more recent times.

Why then is the distribution of particular introns and the inteins so narrow? We offer four different possibilities. First, back to querying the original suggestion: do these elements regulate the genes in which they reside, and thereby give a selective advantage to the organism, too subtle to be measured under standard laboratory conditions, but manifest under the nutrient-deprived, oxygen-limited, temperature-stressed environments prevalent in their natural habitats? Although no evidence has yet been forthcoming for a regulatory role for introns or inteins, they may confer a selective advantage by other means. It has been shown, for example, that in the archaeon *Sulfolobus acidocaldarius* an rDNA intron provides cells with an advantage over intronless cells (21). Additionally, it would appear that the intron endonucleases can confer a selective advantage. The endonuclease of the subtilis phage SP82 preferentially cleaves the DNA of its close relative SPO1 in the vicinity of its intron, thereby effecting exclusion of SPO1 markers and promoting the propagation of its own in mixed infections (22). Furthermore, very different inteins reside at different positions in the *recA* genes of *Mycobacterium tuberculosis* and *Mycobacterium leprae* but are not found in the nonpathogenic mycobacteria. This has been interpreted to suggest that inteins may confer some advantage to their pathogenic hosts (23).

Second, might facilitated entry of introns and inteins into DNA account for their prevalence at particular loci? Possibilities here are that distinctive DNA or nucleoid structures could provide easy access to invasive elements. Such an explanation has been suggested for the high incidence of some mobile elements, such as pathogenicity islands and *Ty* retrotransposons, in highly transcribed regions of DNA (24, 25).

Third, might the sites of intron and intein occupancy be in functions that cannot readily tolerate their loss? Because of the powerful invasive potential of these elements, they would tend to simply reenter sites of perfect excision by endonuclease-mediated mobility. Imperfect excision, on the other hand, might abolish gene function, resulting in a lethal event, thereby favoring retention of the element. Indeed, it has been noted that archaeal and group I introns are often positioned in functionally important regions of genes (26), and a recent systematic alignment of the sites of intein insertion shows them to be at or close to residues involved in catalysis or substrate or cofactor binding (8).

Finally, might some of the preferred target genes in phage act as sinks for intervening sequences because they encode duplicated functions already present in the bacterial host? Thymidylate synthase, and both aerobic and anaerobic ribonucleotide reductases in the coliphage- and subtilis phage-host systems represent such duplicated functions. The "short-term" disadvantage afforded by insertion of a novel intron or intein, which may splice poorly until the element and host environment can adapt to one another, might be tolerated under these circumstances if the bacterial enzymes can be utilized to maintain phage survival.

Determining which, if any, of these possibilities apply, must await more experimental data and genome analyses. However, it is not unreasonable to suspect that one or another of these alternatives may apply to different inteins and introns in different organisms and at different genetic addresses.

We thank Maryellen Carl for expert manuscript preparation, Maureen Belisle for the figures, and David Shub and members of our laboratory for their comments on the manuscript. Work in our laboratory is supported by National Institutes of Health Grants GM39422 and GM44844.

1. Gott, J. M., Shub, D. A. & Belfort, M. (1986) *Cell* **47,** 81–87.
2. Lazarevic, V., Soldo, B., Düsterhöft, A., Hilbert, H., Mauël, C. & Karamata, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1692–1697.
3. Cech, T. R. (1990) *Annu. Rev. Biochem.* **59,** 543–568.
4. Perler, F. B., Olsen, G. J. & Adam, E. (1997) *Nucleic Acids Res.* **25,** 1087–1093.
5. Kane, P. M., Yamashiro, C. T., Wolczyk, D. F., Neff, N., Goebl, M. & Stevens, T. H. (1990) *Science* **250,** 651–657.
6. Perler, F. B., Comb, D. G., Jack, W. E., Moran, L. S., Qiang, B., Kucera, R. B., Benner, J., Slatko, B. E., Nwankwo, D. O., Hempstead, S. K., Carlow, C. K. S. & Jannasch, H. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 5577–5581.
7. Davis, E. O., Jenner, P. J., Brooks, P. C., Colston, M. J. & Sedgwick, S. G. (1992) *Cell* **71,** 201–210.
8. Dalgaard, J. Z., Moser, M. J., Hughey, R. & Mian, I. S. (1997) *J. Comput. Biol.* **4,** 193–214.
9. Pietrokovski, S. (1998) *Protein Sci.* **7,** 64–71.
10. Shub, D. A., Gott, J. M., Xu, M.-Q., Lang, B. F., Michel, F., Tomaschewski, J., Pedersen-Lane, J. & Belfort, M. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 1151–1155.
11. Bechhofer, D. H., Hue, K. K. & Shub, D. A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 11669–11673.
12. Goodrich-Blair, H. & Shub, D. A. (1994) *Nucleic Acids Res.* **22,** 3715–3721.
13. Mikkonen, M. & Alatossava, T. (1995) *Microbiology* **141,** 2183–2190.
14. Damberger, S. H. & Gutell, R. R. (1994) *Nucleic Acids Res.* **22,** 3508–3510.
15. Perlman, P. S. & Butow, R. A. (1989) *Science* **246,** 1106–1109.
16. Bell-Pedersen, D., Quirk, S., Clyman, J. & Belfort, M. (1990) *Nucleic Acids Res.* **18,** 3763–3770.
17. Dalgaard, J. Z., Garrett, R. A. & Belfort, M. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5414–5417.
18. Loizos, N., Tillier, E. R. M. & Belfort, M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 11983–11987.
19. Duan, X., Gimble, F. S. & Quiocho, F. A. (1997) *Cell* **89,** 555–564.
20. Derbyshire, V., Wood, D. W., Wu, W., Dansereau, J. T., Dalgaard, J. Z. & Belfort, M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 11466–11471.
21. Aagaard, C., Dalgaard, J. Z. & Garrett, R. A. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 12285–12289.
22. Goodrich-Blair, H. & Shub, D. A. (1996) *Cell* **84,** 211–221.
23. Davis, E. O., Thangaraj, J. S., Brooks, P. C. & Colston, M. J. (1994) *EMBO J.* **13,** 699–703.
24. Curcio, M. J. & Morse, R. H. (1996) *Trends Genet.* **2,** 436–438.
25. Hacker, J., Blum-Oehler, G., Muhldorfer, I. & Tschape, H. (1997) *Mol. Microbiol.* **23,** 1089–1097.
26. Garrett, R. A., Dalgaard, J., Larsen, N., Kjems, J. & Mankins, A. S. (1991) *Trends Biochem. Sci.* **16,** 22–26.