Back to Basics: Structure, Function, Evolution and Application of Homing Endonucleases and Inteins

MARLENE BELFORT

It is a profound and necessary truth that the deep things in science are not found because they are useful; they are found because it was possible to find them.

Robert Oppenheimer 1904–1967

1 Introduction: Back to Basics

Oppenheimer's words resonate with this book's theme, which is how both applied and theoretical science can emanate from answers to basic questions – whether they are being asked in model organisms or in test tubes – about molecular structure and mechanism. Thus, fundamental work on DNA, RNA and proteins, which encode or constitute homing endonucleases and inteins, is leading to refined theories of evolution in prokaryotes and eukaryotes on the one hand, and the development of laboratory tools and health-care reagents on the other.

Homing endonucleases and inteins, sometimes referred to as "protein introns", are linked at many levels. First, homing endonucleases are frequently encoded by introns that self-splice at the RNA level, in analogy to inteins that self-splice at the protein level. Second, homing endonucleases similar to those encoded by introns are often found embedded within and co-translated with inteins. Third, both types of intervening sequence are mobile elements, capable of movement from genome to genome. Fourth, the endonuclease component of both introns and inteins imparts their mobility. Fifth, each of these mobile intervening sequences is thought to have originated from invasion of the gene encoding the self-splicing element, the intron or intein, by

M. Belfort (e-mail: belfort@wadsworth.org)
Wadsworth Center, New York State Department of Health, Center for Medical Sciences, 150 New Scotland Avenue, Albany, New York 12208, USA

Nucleic Acids and Molecular Biology, Vol. 16
Marlene Belfort et al. (Eds.)
Homing Endonucleases and Inteins

an endonuclease gene, the primordial mobile element (Fig. 1). The final unifying theme is the exploitation of introns, inteins and homing endonucleases by chemists, geneticists, structural biologists and engineers, to generate reagents and tools that are useful in basic research, biotechnology and medicine. This chapter serves as an introduction to the volume entitled *Homing Endonucleases and Inteins*, which provides a wonderful illustration of the point that fundamental studies of structure and mechanism fuel evolutionary theory and technology development alike.

2 What Is a Homing Endonuclease?

Homing endonucleases are rare-cutting enzymes that are most often encoded by introns or inteins, but they can also be free-standing, occurring between genes. The genesis of the homing endonuclease field dates back to 1970, with the observation, in genetic crosses between yeast mitochondria, of a significant polarity of recombination for markers of an rRNA gene (Dujon, this Vol.). In 1985, this phenomenon became attributable to an intron-encoded homing endonuclease that initiated recombination within the rRNA gene. Minimally, homing endonucleases are protein enzymes that make a site-specific double-strand break (DSB) at the "homing" site in intron-less or intein-less alleles, thereby initiating a gene conversion event through which the intron or intein is copied into the break site (Fig. 2A, B; reviewed by Chevalier and Stoddard 2001; Belfort et al. 2002; Dujon, this Vol.). For the group I and archaeal intron endonucleases and inteins, the recombinogenic ends created at the DSB engage in a strictly DNA-dependent recombination process that duplicates

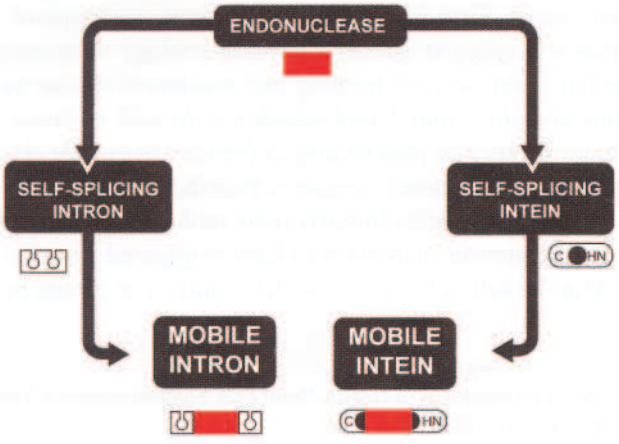

Fig. 1. Evolution of mobile introns and inteins. Endonuclease genes (*red*) are proposed to have invaded DNA encoding self-splicing introns or inteins, to generate mobile genetic elements. (Dassa and Pietrokovski, this Vol.)

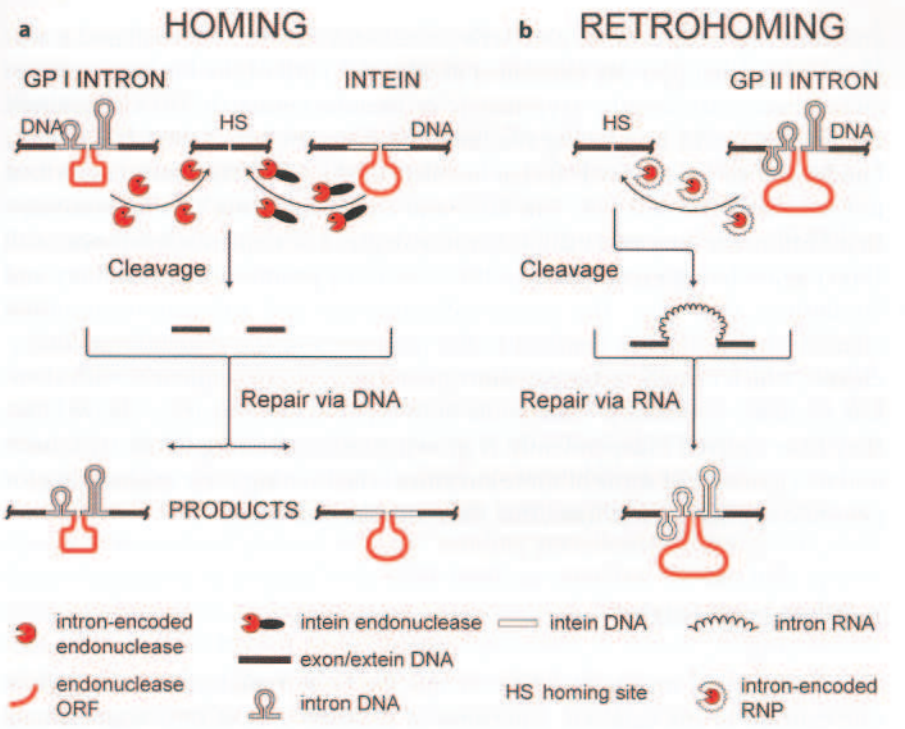

Fig. 2. Mobility of introns and inteins. **a** DNA-based homing of group I introns and inteins. The intron or intein endonuclease cleaves the homing site of a cognate intron- or intein-less allele. Gene conversion repairs the break to generate intron- or intein-containing products (Dujon, this Vol.). **b** Retrohoming of a group II intron. In this case, the homing site DNA is invaded by intron RNA, and the opposite strand is cleaved by an intron-encoded protein, which is part of an RNP complex. The intron is copied into cDNA to generate the intron-containing product (Lambowitz et al., this Vol.).

the intron or intein (Fig. 2A). The group II intron-encoded proteins are more complex, forming a ribonucleoprotein (RNP) particle with the intron RNA (Fig. 2B). The intron invades the DNA sense strand (mRNA-like strand) by reverse-splicing, whereas the endonuclease domain of the protein nicks the antisense strand. The intron acquisition event is completed with a cDNA copy of the intron, in a process termed retrohoming (Lambowitz et al., this Vol.).

The homing endonucleases fall within four families, characterized by the sequence motifs LAGLIDADG (Caprara and Waring, this Vol.; Chevalier et al., this Vol.; Dujon, this Vol.; Haber and Wolfe, this Vol), GIY-YIG (Edgell, this Vol.; Van Roey and Derbyshire, this Vol.), His-Cys box (Galburt and Jurica, this Vol.; Keeble et al., this Vol.) and HNH (Keeble et al., this Vol.). However, recent structural data support the hypothesis that the His-Cys box and HNH

enzymes share features at their active sites, and should be considered a single family, called ββα-Me (Keeble et al., this Vol.). All of the homing endonucleases recognize lengthy asymmetric or pseudosymmetric DNA sequences, ranging from a 14-bp homing site for I-DmoI, a member of the LAGLIDADG family, to a 40-bp site for I-TevI, a member of the GIY-YIG family (described in Chevalier et al., this Vol.; Van Roey and Derbyshire, this Vol., respectively). In addition, the enzymes exhibit varying degrees of sequence tolerance, with I-TevI again being exceptional, in this case in its promiscuity (Van Roey and Derbyshire, this Vol.). The conserved sequences and substrate-recognition characteristics stand in contrast to the properties of the restriction endonucleases, which usually recognize short palindromic DNA sequences with absolute sequence specificity. Thus, while both types of endonuclease cleave DNA, they have evolved independently. A growing understanding of the structures and mechanisms of some of these enzymes is facilitating their engineering for genomic applications (Dujon, this Vol.; Gimble, this Vol.).

3 What Is an Intein?

The discovery of inteins and protein splicing represented a breakthrough in our concept of the catalytic repertoire of proteins and of post-translational modification (Perler, this Vol.). Conserved residues at the intein–extein junctions facilitate splicing (Mills and Paulus, this Vol.; Perler, this Vol.). Several inteins are bifunctional proteins that not only catalyze protein splicing, but also function as endonucleases, to initiate homing of the intein gene (Figs. 1 and 2A). Additionally, several inteins have motifs suggesting an evolutionary relationship to intron-encoded homing endonucleases. The endonuclease and the protein-splicing component are genetically, structurally, and functionally separable (Dassa and Pietrokovski, this Vol.; Moure and Quiocho, this Vol.; Perler, this Vol.), supporting the hypothesis that the endonuclease genes invaded the genes of these self-splicing elements, which provided safe havens, while themselves acquiring mobile properties (Fig. 1).

4 Inteins and Homing Endonucleases as Molecular Mosaics

The invasion of self-splicing introns and inteins by endonuclease genes appears to have occurred multiple times, given that these elements encode endonucleases of different families and that some endonuclease genes of the same family emanate from different positions of group I introns. It appears that some of these endonucleases then adapted to function in other process-

es, e.g., repression of transcription (Van Roey and Derbyshire, this Vol.), and promotion of splicing through acquisition of maturase activity (Caprara and Waring, this Vol.).

Inteins share an ancestry with metazoan hedgehog proteins, which undergo a self-cleavage reaction. The common Hint (*hedgehog/intein*) domain is a structural unit with a mechanistic identity (Dassa and Pietrokovski, this Vol.). It is apparent that composite elements like mobile introns, inteins and hedgehog proteins have interchanged functional domains in the course of evolution.

Endonucleases themselves have evolved specificity by fusion of a catalytic domain, containing the conserved motif, with variant DNA-binding domains, e.g., the GIY-YIG and HNH endonucleases (Van Roey and Derbyshire, this Vol.). The modular nature of these enzymes is further illustrated for I-TevI, in which the DNA-binding domain is itself an assembly of small DNA-binding units, some of which are present in other homing endonucleases. These enzymes have evolved a broad range of binding specificities, through the shuffling of catalytic cartridges with DNA-binding cassettes. We can only speculate as to how such molecular mosaics are formed. The most popular view is that proposed for hybrid bacteriophage genomes, in which "illegitimate recombination takes place quasi-randomly along the recombining genomes, generating an unholy mélange of recombinant types" (Pedulla et al. 2003). This sloppy, non-homologous recombination would generate a mound of genetic junk, with only a miniscule number of recombinants being selected, on the basis of their function and/or viability.

A lingering question is whether homing endonucleases and their genes are maintained specifically to promote their own selfish lifestyles and that of their host elements (introns and inteins), or whether they additionally serve some useful function for the organism. While their invasiveness and success as selfish intruders are undisputed, a potential advantage to the host organism has been observed, in experiments with phage T4 and its relative T2 (Edgell, this Vol.). Here, GIY-YIG homing endonucleases act to promote the spread of genes from their host organism to its relatives. This is a satisfying observation, considering that 8% of the phage T4 genome comprises endonuclease genes. Another "useful" homing enzyme is HO endonuclease, the first member of the LAGLIDADG family to be discovered, and the first shown to make a DSB. This intriguing enzyme catalyzes mating-type switching in yeast (Haber and Wolfe, this Vol.).

5 Applications: "Turning Junk into Gold"

The title of this subsection has been borrowed from that of an essay on the practical application of introns and inteins (Wickelgren 2003). Basically, the topic breaks down into the utility of homing endonucleases, group II intron RNPs, and inteins.

5.1 Site-Specific Group I Intron and Intein Endonucleases

The engineering of DNA often demands cleavage at rare sites and, to some extent, the highly site-specific endonucleases of group I introns and inteins fulfill the requirement. One need simply open the catalog of a molecular cloning company to identify those enzymes that cut the bacterial genome of ~5000 kb seven times, or a yeast genome of ~13,000 kb only once. They sound like restriction enzymes, bearing odd names like I-CreI and PI-PspI, with the prefixes I and PI, respectively, designating intron-encoded and protein intron-encoded (Perler, this Vol.). The engineering of single sites of the historic ω intron-endonuclease, I-SceI, into everything from yeast to mammals has facilitated not only genome sequencing projects, but also studies of DSB repair and non-homologous end joining (Dujon, this Vol.). However, the dream is not only for rare cleavage, but also for customized recognition sequence specificity. LAGLIDADG endonucleases are indeed starting to be engineered to alter their recognition-site specificity, through domain shuffling, and selection for new amino acid-DNA contacts, as described by Gimble later in this volume (Fig. 3A).

5.2 Gene Targeting by a Group II Intron RNP Complex

The novel DNA insertion mechanism of group II introns, via intron RNA-target DNA base pairing, has enabled the development of group II introns as site-specific gene targeting agents (Lambowitz et al., this Vol.). Through genetic manipulation, the intron RNA sequences that base pair with DNA, the so-called exon binding sequences, are changed (Fig. 3B). Thereby the group II Ll.LtrB intron has been re-targeted to specific genes in several bacteria. Additionally, this intron has been directed to HIV provirus and a plasmid-borne HIV coreceptor, and remained functional in transfection assays. This group II intron system is poised for different kinds of targeted gene manipulation (Fig. 3B), including gene disruption by integration into the antisense of an expressed gene, and conditional disruption via insertion into the sense strand and regulation of splicing. Such approaches will greatly facilitate functional

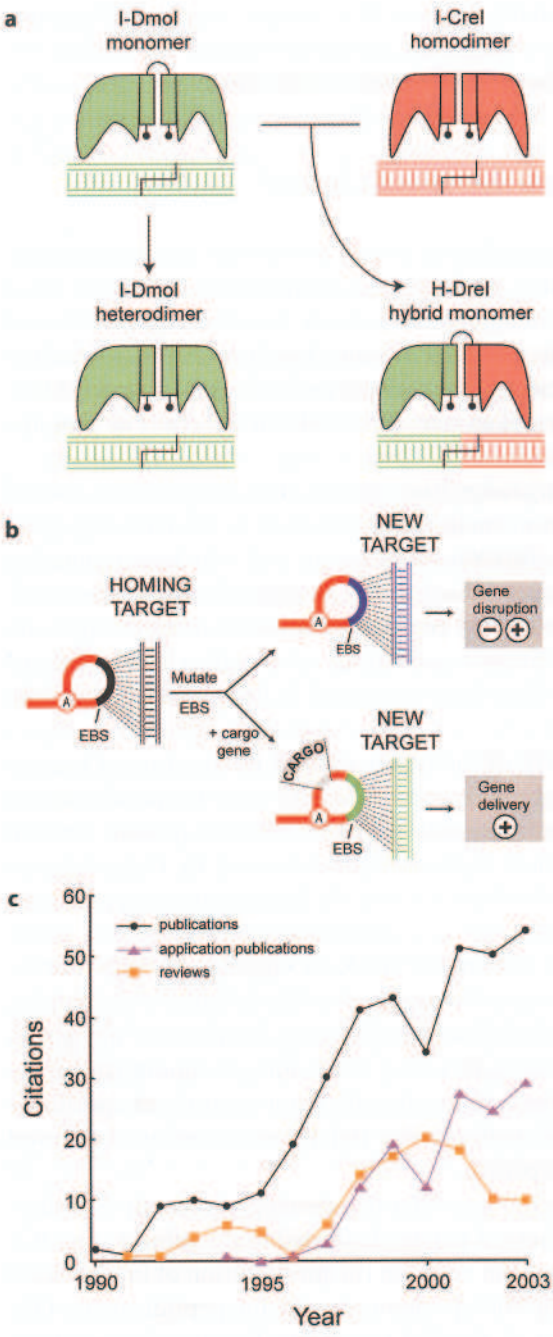

Fig. 3. Practical application of introns and inteins. **a** Domain manipulation of group I intron endonucleases. The splitting of monomers into heterodimers and the generation of hybrid nonomers pave the way for increasing the substrate repetoire of LAGLIDADG homing endonucleases (Gimble, this Vol.). **b** Group II introns as gene-targeting agents. The wild-type intron lariat targeting its natural homing site through its exon binding sequence (EBS; both in *black*) is shown on the *left*. Mutation of the EBS to target new sites is reflected on the *right* by a color change of the EBS that matches that of the new site (*blue* or *green*). Gene disruption, which can be either absolute (–) or conditional (+), is shown at the *top*, and delivery of a foreign cargo gene, with maintenance of gene function (+), is shown at the *bottom* (Lambowitz et al., this Vol.). **c** Burgeoning intein technology since the discovery of inteins in 1990 (Perler, this Vol.).

genomics. Gene therapy, involving delivery of a foreign sequence, is a major goal, pending delivery to mammalian cells, transport through the nuclear envelope, and integration into the desired chromosomal sites.

5.3 "Inteins ... Nature's Gift to the Protein Chemist"

The metaphor in this subheading derives from a recent comment made by intein researcher Tom Muir (2004, pers. comm.). In the short time span since their discovery barely 15 years ago, intein activity has been harnessed and controlled to build an entirely new area of biotechnology (Fig. 3C), which is predicted to explode in the coming years. Intein technology has already facilitated protein purification, as well as providing tools for the study of proteins both in vitro and in vivo.

Switches that control splicing range from temperature and pH shifts (Wood et al., this Vol.) to the addition of small molecules, such as the reducing agent dithiothreitol (Chong and Xu, this Vol.), rapamycin and 4-hydroxytamoxifen (reviewed by Ozawa and Umezawa, this Vol.). In all cases, the intein was modified in order to impart controllability. For facilitation of such intein engineering, and for the monitoring of intein activity in vitro and in vivo, a battery of different reporter systems have been developed (Chong and Xu, this Vol.; Wood and Skretas, this Vol.).

By using the basic chemistry of the intein, mutant intein–protein fusions can be chemically ligated to a protein or peptide with an N-terminal cysteine residue. This process has been variously termed expressed protein ligation (EPL) or intein-mediated protein ligation (IPL) (reviewed by Perler later in this volume). This EPL/IPL technology has already had enormous application, as for example in protein stabilization and potentiation by cyclization (Tavassoli et al., this Vol.), segmental labeling for NMR, or tagging with different reporters (Muir 2003). Equally impressive are in vivo intein-based technologies, many of them based on split inteins, and many using fluorescence or light as the reporter (Ozawa and Umezawa, this Vol.). These include monitoring of intracellular protein–protein interactions; identification of proteins specifically imported into mitochondria, endoplasmic reticulum, or nucleus; and even generation of safer transgenic plants.

This staggering array of technology that has developed recently is underscored by the increasing numbers of papers dedicated to the subject (Fig. 3C). Looking into the crystal ball, we can envision the proliferation of intein-based biosensors (Chong and Xu, this Vol.), proteomics utilizing peptide arrays (Tavassoli et al., this Vol.; Wood et al., this Vol.), and even functional proteomics in whole animals (Ozawa and Umezawa, this Vol.). Nature's gift, indeed!

6 The Message

What, then, are the lessons to be learned from these mobile elements that roam diverse genomes? To ponder their origins, we must first know their substance; to harness their ingenuity, we must first understand their action. The investment in discerning their substance and their action is yielding high payoffs. We are already gaining insight into how inteins, introns and homing endonucleases may have evolved, and how they may influence the evolution of genomes. Also, we have exploited the exquisite specificity of the endonucleases to manipulate DNA in vitro and in vivo. Finally, we have manipulated the activity of inteins to purify proteins, study them in test tubes, cells and multicellular organisms, and to build safer transgenic plants, as inteins provide the prospect of performing functional proteomics in whole animals.

Among the ultimate satisfactions for a scientist is to achieve mechanistic and evolutionary insights. To then apply these insights to practical problem-solving is frosting on the cake. Where, then, do we start? What is the message to our students, who, in their youthful enthusiasm, are likely to want both the thrill of discovery and the satisfaction of application? Don't start with the frosting! Study the fundamentals, and be vigilant! In that way, the potential for doing it all well will be maximized, and further compelling examples of the unpredicatable value of basic research will continue to emerge. We must keep bearing Oppenheimer's words in mind, that "the deep things in science are not found because they are useful". A terrific case in point is the study of a bizarre genetic phenomenon in 1970, which led to the discovery of the first homing endonuclease in 1985 that in turn resulted in a reagent, I-SceI, that is widely used in megasequencing projects of today (Dujon, this Vol.).

Acknowledgments. I dedicate this chapter to Bernard Dujon, who gave birth to this field, and who beautifully illustrated the message, that is, to study the fundamentals, to be vigilant, and then to seize the opportunities for application. I am grateful to all of the authors who contributed to this book and particularly to my coeditors Vicky Derbyshire, Barry Stoddard and Dave Wood, who have, through their hard work and shared insights, helped make this volume come to life. I thank Maryellen Carl for preparing this manuscript and handling the 19 others, and John Dansereau for providing the figures. Work in our laboratory is supported by NIH grants GM39422 and GM44844 and NSF grant NIRT0210419.