CMLS, Cell. Mol. Life Sci. 55 (1999) 1304–1326
1420-682X/99/101304–23 $ 1.50 + 0.20/0

CMLS Cellular and Molecular Life Sciences

Review

Homing endonucleases: structure, function and evolution

M. S. Jurica and B. L. Stoddard*

Division of Basic Sciences, Fred Hutchinson Cancer Research Center and the Graduate Program in Molecular and Cell Biology, University of Washington, 1100 Fairview Ave. N. A3-023, Seattle (Washington 98109, USA), Fax + 1 206 667 6877, e-mail: bstoddar@fred.fhcrc.org

Received 6 January 1999; received after revision 24 February 1999; accepted 24 February 1999

Abstract. 'Homing' is the lateral transfer of an intervening genetic sequence, either an intron or an intein, to a cognate allele that lacks that element. The end result of homing is the duplication of the intervening sequence. The process is initiated by site-specific endonucleases that are encoded by open reading frames within the mobile elements. Several features of these proteins make them attractive subjects for structural and functional studies. First, these endonucleases, while unique, may be contrasted with a variety of enzymes involved in nucleic acid strand breakage and rearrangement, particularly restriction endonucleases. Second, because they are encoded within the intervening sequence, there are interesting limitations on the position and length of their open reading frames, and therefore on their structures. Third, these enzymes display a unique strategy of flexible recognition of very long DNA target sites. This strategy allows these sequences to minimize nonspecific cleavage within the host genome, while maximizing the ability of the endonuclease to cleave closely related variants of the homing site. Recent studies explain a great deal about the biochemical and genetic mechanisms of homing, and also about the structure and function of several representative members of the homing endonuclease families.

Key words. Homing endonuclease; mobile intron; intein; crystallography.

Genetic and biochemical studies

The discovery and basic molecular biology of homing endonucleases and mobile introns

'Homing' is a high-frequency, site-specific gene conversion event, where a mobile intervening sequence, either a group I or group II intron or an intein, is copied and transferred to a specific insertion site within a cognate allele of the host gene missing the intervening sequence (fig. 1) [1–7]. The end result is the duplication of the mobile intron or intein within a diploid genome and a depletion of I^- alleles from the population, conferring an advantage for these elements. The mobile elements avoid disrupting their host gene function by splicing themselves out at either the RNA level (group I and group II introns) or at the protein level (inteins). Intron mobility is initiated by double-strand breaks (DSBs) in the recipient allele. The DSB is catalyzed by an endonuclease encoded by an open reading frame (ORF) located within the mobile intron or as a domain within an intein. Although a homing event may be loosely thought of as a unique form of genetic transposition, one should make a clear distinction between a mobile intron and its homing endonuclease as opposed to a transposon and its transposase factor. Both processes involve transfer of a mobile element initiated by a protein encoded within that sequence. However, whereas a transposase recognizes and interacts with the ends of the transposon (and acts in a manner similar to a recombinase or integrase), homing endonucleases do not recognize their corresponding mobile DNA but act instead by simply cleaving their target site.

* Corresponding author.

The first mobile intron discovered was a genetic marker in *Saccharomyces cerevisiae* termed 'ω', which was observed in the early 1970s to be transferred to ω^- strains when they were crossed with strains carrying the marker [8]. This marker was shown to correspond to a 1.1-kb group I intron found in the large (21S) ribosomal RNA (rRNA) gene of the mitochondrial genome in ω^+ yeast strains [9, 10]. After this locus was mapped, several studies showed that the gene conversion event was accompanied by a transient DSB at the insertion site for the intron [11, 12]. Further work revealed that the expression of an ORF within the mobile intron is essential for both DNA cleavage and intron mobility [13, 14]. The ORF product was shown to be a site-specific DNA endonuclease that acts to initiate the homing event [12]. This protein, now called I-*Sce*I, was the first known representative of the intron-encoded proteins now known collectively as homing endonucleases. These proteins are widespread, found throughout many biological phyla and encoded within many different types of intervening sequences. They comprise several different endonuclease families that appear to have arisen independently of one another.

Several features of these proteins make them particularly attractive subjects for genetic, biochemical and structural studies. First, their structure and mechanism may be directly compared and contrasted with traditional restriction endonucleases, as well as other proteins involved in nucleic acid cleavage and rearrangements. As discussed below, homing endonucleases share many similarities with such proteins, while also exhibiting unique folds and catalytic mechanisms. Second, homing endonucleases are encoded by reading frames within mobile introns or inteins, which influences their size. For example, the group I intron-encoded catalysts are generally quite small compared with traditionally encoded proteins of similar function. Finally, homing endonucleases act as protein catalysts that initiate the specific transfer of what may be thought of as very simple invasive elements. These proteins display sufficient DNA recognition specificity to avoid catalyzing nonspecific, lethal DSBs throughout the host genome, while at the same time maintaining cleavage activity against mutated variants of the homing site [15–20]. This unique pattern of DNA recognition specificity may have evolved to maximize the potential for

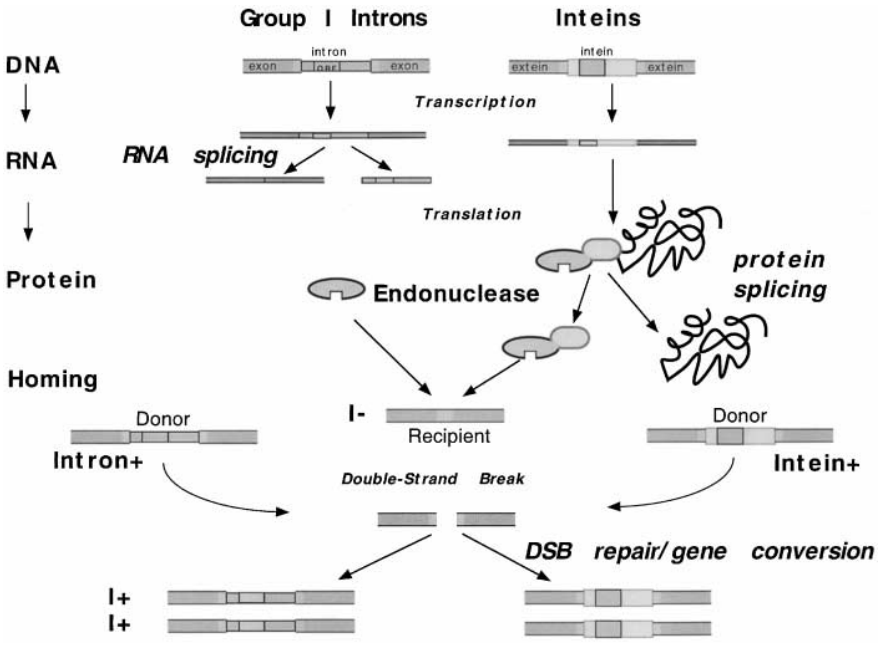

Figure 1. Schematic of homing. Homing is the unidirectional lateral transfer of an intervening sequence to a specific insertion site within a homologous allele of the intron's normal host gene lacking the sequence. Homing is initiated by a DSB in the recipient allele (I−) catalyzed by an endonuclease encoded within the intervening sequence and results in the duplication of the mobile intron or intein to two alleles (I+) within a diploid genome. The intervening sequence avoids disrupting the host gene in two ways. In the case of mobile introns (left), the intron is removed at the RNA level by self-splicing. In contrast, mobile inteins and their nested endonuclease (right) are translated as a fusion protein in frame with the surrounding host gene product. The intein and endonuclease are spliced out through a complex protein-based transesterification mechanism.

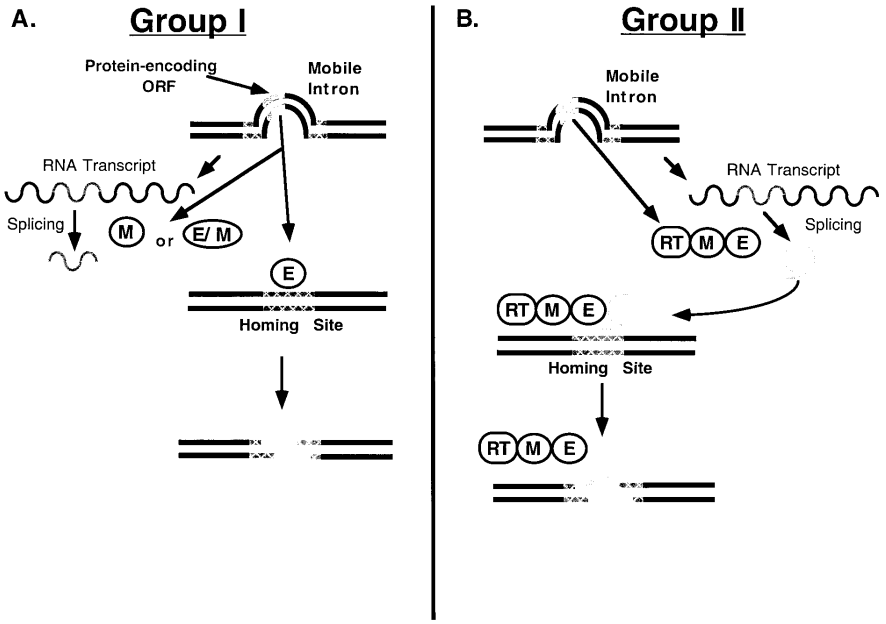

Figure 2. Group I intron- versus group II intron-encoded proteins. (*A*) Group I-encoded proteins often display the single activity of site-specific cleavage of double stranded DNA homing sites ('E'). Transfer of the group I intron is then completed by a gene conversion event. Some group I intron-encoded proteins display a maturase activity ('M') and act to facilitate the splicing of the RNA intron by binding the intron and stabilizing it in an active conformation. A subset of these proteins are bifunctional ('E/M') and carry both homing endonuclease and maturase activity. (*B*) Group II intron-encoded proteins are generally larger, trifunctional proteins that possess endonuclease, maturase and reverse transcriptase (RT) activity. These proteins facilitate a more complex homing mechanism than observed for the mobile group I introns. Homing requires a splicing-proficient intron and maturase, and proceeds through the consecutive activities of endonuclease and reverse transcriptase domains. The protein binds the excised group II intron (as an RNA lariat) to form a stable RNP; the endonuclease uses the 3′ OH of the intron as a nucleophile to initiate the DSB, forming a DNA/RNA hybrid.

transfer between closely related host species. Homing endonucleases achieve this combination of a high degree of overall sequence specificity, while tolerating individual mutations within their target sites, by reading out very long DNA sequences and thereby minimizing the relative energetic penalty of any one particular 'mismatch' throughout the protein/DNA interface. The second half of this review discusses recent advances in addressing these issues of homing endonuclease structure and function.

Mobile introns and homing endonucleases are widespread

A subset of group I and group II introns demonstrate genetic mobility. Group I introns display the most widespread phylogenetic diversity of any class of intervening sequences. The majority have been found in the mitochondrial DNA of fungi, but these elements have also been characterized in mitochondrial and chloroplast genomes of plants and algae, as well as in the nuclei of ciliates, slime molds, algae and fungi [5, 6]. They have also been found in eubacterial and bacteriophage genomes. Of the known group I introns, approximately 30% harbor internal reading frames. Most, if not all, of these ORFs encode proteins involved either in mobility of the intron (endonucleases) or in splicing of the intron after transcription (maturases; fig. 2A). The first reported mobile group I introns (with associated intron-encoded homing endonucleases) were identified in the *Saccharomyces* mitochondrial genome. Other mobile group I introns have been found in the *td* and *sunY* genes of T4 phage [21, 22], in a chloroplast rRNA gene of *Chlamydomonas* [23], in a nuclear rRNA gene of *Physarum polycephalum* [24] and in the cytochrome oxidase subunit I (*COX1*) gene of yeast [25]. Since these initial observations, the list of mobile group I introns has grown rapidly. When one considers the wide range of host organisms and cellular compartments containing mobile group I introns, it becomes evident that a significant fraction of these sequences are either actively mobile or have descended from mobile ancestors.

Group II introns have been found in the genomes of fungal and plant mitochondria, plant and algae chloro-

plasts and eubacteria [4–6, 26]. They comprise the majority of intervening sequences in choroplast DNAs and higher plant mitochondrial DNAs. Most are found in protein-encoding genes, with a small fraction found in transfer RNA (tRNA) and rRNA genes. A smaller percentage of group II introns possess internal ORFs or exhibit mobility than is observed for group I introns. The wide phylogenetic distribution of mobile group II introns again implies a long history of mobilization of these sequences [27]. As discussed below, the proteins encoded by group II intron ORFs tend to be more complex than those associated with group I introns, containing multiple domains of differing functions (fig. 2B) [3]. The participation of each of these protein domains in both splicing and intron mobility has been demonstrated conclusively.

Homing mechanisms of mobile group I and group II introns vary dramatically

Homing of group I introns (fig. 3A) requires homology between exon sequences and is initiated by endonucleases encoded within the mobile introns as described above. The endonuclease catalyzes a DSB in the recipient allele, which stimulates the recombination events of the DSB repair pathway, leading to intron transfer and gene conversion. Initial studies of group I intron homing with the *td* intron of phage T4 were carried out in a λ-phage system [28, 29]. These studies indicated the requirement of a 5′,3′ λ exonuclease, the *Escherichia coli* 3′–5′ exonuclease III, and the *E. coli* recombinase RecA. These functions allow formation of single-stranded DNA tails that undergo homologous strand invasion of an intron-containing allele, facilitating repair of the DSB and precise intron insertion.

Subsequent studies using a T4 phage infection system, however, have indicated that group I intron homing also uses additional gene conversion pathways after the initial DSB (fig. 3A) [30]. In vivo assays using hosts deficient in enzymes responsible for crossover resolution (resolvases) still show transfer of these mobile elements, although at reduced efficiency. Furthermore, an underrepresentation of crossover products indicates that group I intron homing can also proceed via a recombination-independent pathway. Possible mechanisms might include synthesis-dependent strand annealing (SDSA) or topoisomerase-mediated (TM) pathways, in which the migration of branchpoints, strand passages and/or templated strand synthesis result in resolution of the initial intermediate with noncrossover products [30].

In contrast, homing of mobile group II introns has recently been shown to proceed via a novel pathway completely dissimilar to the mobile group I elements (fig. 3B) [27]. The pathway makes use of three distinct protein activities, all present on a single intron-encoded protein. Group II intron mobility is also initiated by an endonuclease-catalyzed DSB, with the added feature that the excised RNA intron participates in this step and is reverse-spliced into the cleaved target site, forming a covalently linked RNA-DNA intermediate. Homing of group II introns has a strict requirement for a splicing-proficient intron [31, 32]. The homing process can be interrupted either by mutations in the intron that cause misfolding or inactivation of the RNA molecule, or alternatively by specific mutations in the maturase domain of the intron-encoded protein. This requirement implies the direct involvement of the spliced RNA intron in homing. Genetic and biochemical studies of the mobile aI2 intron in the mitochondrial yeast *COX1* gene suggest an intricate mechanism mediated by the spliced group II intron in complex with the intron-encoded protein, a process termed 'retrohoming' [31–38]. The spliced intron associates with the intron-encoded protein to form a stable ribonucleoprotein (RNP). The terminal 3′ OH group of the intron serves as a nucleophile for the cleavage of one strand of the DNA homing site, forming a reverse-spliced, covalently linked DNA-RNA species at the insertion site. The endonuclease domain of the protein is responsible for completing the DSB. The inserted RNA intron then serves as the template for complementary (cDNA) synthesis. Group II intron homing displays many features of retrotransposon mobility, leading to conjecture as to the evolutionary and/or mechanistic relationship between these elements.

Distribution and classification of group I homing endonucleases

The group I homing endonucleases have been more extensively characterized structurally and mechanistically and are the focus of the remainder of this review. These proteins are grouped into four separate families (fig. 4) on the basis of conserved sequence motifs [2, 7]. These families appear to have arisen independently; however, all have optimized their ability to recognize long DNA targets, to tolerate small variations in the homing site sequence and to efficiently catalyze DSBs.

The LAGLIDADG family. This large protein family, with more than 130 known members, has variously been termed the 'decapeptide', 'dodecapeptide', 'dodecamer', 'DOD' or 'LAGLIDADG' family and contains the largest known collection of intron-encoded proteins [2, 39]. Members of this family are grouped together on the basis of their most recognizable conserved sequence motif: one or two copies of a 10-residue sequence known as a dodecapeptide or LAGLIDADG motif. Endonucleases that contain a single copy of this motif, such as I-*Cre*I and I-*Ceu*I, tend to be $\sim$18–20 kDa in mass and act as homodimers. Proteins that have two

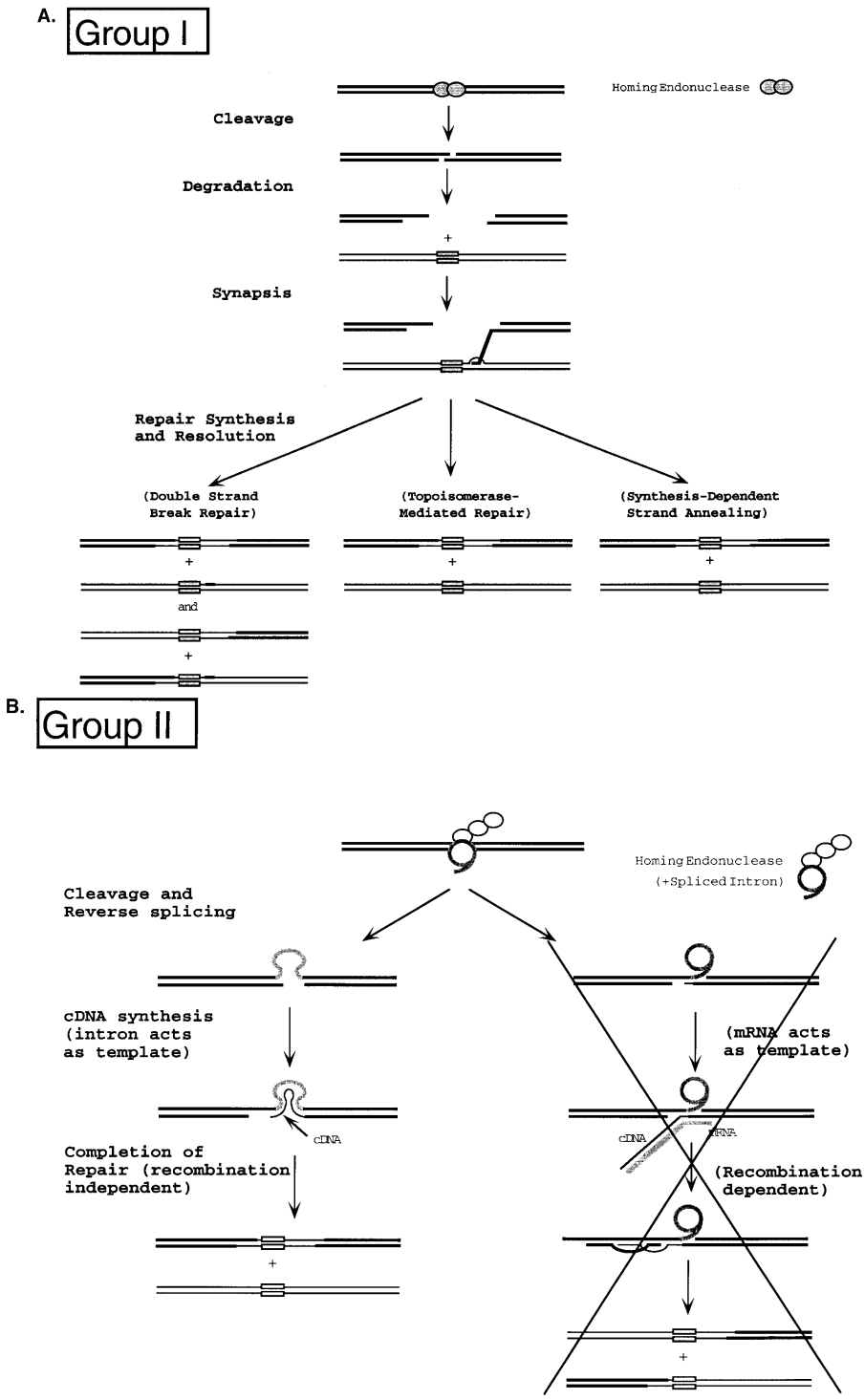

Figure 3. General features of homing pathways for mobile group I introns and group II introns. (*A*) Group I intron mobility is initiated by a DSB made by the homing endonuclease, followed by exonucleolytic degradation, homologous sequence alignment and 3′ end invasion of the donor allele. The donor allele serves as a template for repair synthesis. A variety of pathways (recombination dependent and independent) are then used to repair the DSB and resolve the homologous alleles. (Based on a similar figure in Mueller et al. [30]). (*B*) Like the group I mobile introns, the DSB can potentially be repaired through recombination-independent (left) or -dependent (right) pathways. However, the lack of crossover products and the fact that the intron itself, rather than the mRNA, is used as a template for cDNA synthesis implies that homing proceeds entirely by complete reverse splicing, independent of homologous DNA recombination (left), as described in Cousineau et al. [38].

copies of this motif, such as I-*Dmo*I (encoded within an archaeal intron) and the endonuclease domain of the PI-*Sce*I intein, generally are about twice that size. These proteins are active as monomers and have 80–150 residues separating the two boxes. Aside from intron-encoded homing endonucleases, proteins containing the LAGLIDADG motif include RNA maturases, free-standing nuclear-encoded proteins such as yeast HO mating-type switch endonuclease, and the endonuclease domains of some self-splicing inteins as described below. All the endonucleases containing the LAGLIDADG motif recognize long, pseudopalindromic homing sites of 14–30 bp in length and cleave their homing site DNA to generate 4nt, 3′ extensions. They are dependent on divalent cations for activity, similar to most nuclease catalysts. In the past 2 years, the structure of a dimeric LAGLIDADG endonuclease (I-*Cre*I) has been reported both as apo-enzyme [40] and bound to a synthetic DNA homing site [41]. The structures of two monomeric endonucleases (I-*Dmo*I and PI-*Sce*I) have been reported in the absence of DNA [42, 43]. These structures are summarized in the second half of this review.

The His-Cys box family. The His-Cys box endonuclease family is a much smaller collection of enzymes and contains the only eukaryotic homing endonucleases known to be encoded within nuclear genomes. The enzymes are homodimers of apparent molecular mass 18–20 kDa. The proteins in this family are remarkably rich in His and Cys residues, particularly over a central region of approximately 100 residues [44, 45]. As an example, the I-*Ppo*I endonuclease from *Physarum polycephalum*, the first endonuclease discovered in this family (and the only one to be extensively characterized), contains the His-Cys-rich sequence H-X_4-H-X_{15}-C-X_3-H-X-C-H-X_3-C-H-X_3-H-X-C-X_{12}-C-X-C-X_6-H-X_3-C. Biochemical studies of I-*Ppo*I indicate there are one or more bound zinc atoms per protein subunit [46]. Although the His-Cys-rich sequence clearly has the potential to bind zinc, it does not align well with the consensus sequence of any previously identified zinc binding domain. The DNA-binding properties of I-*Ppo*I have been well characterized [15, 47–50]. This endonuclease binds and cleaves an asymmetric pseudopalindromic homing site and induces a significant DNA bend. The specific homing site for I-*Ppo*I is long (14 bp) compared with restriction endonucleases, but shorter than the sites recognized by most LAGLIDADG endonucleases (fig. 4). I-*Ppo*I can be activated in vitro by several divalent metal ions, including Mg^{2+}, Mn^{2+},

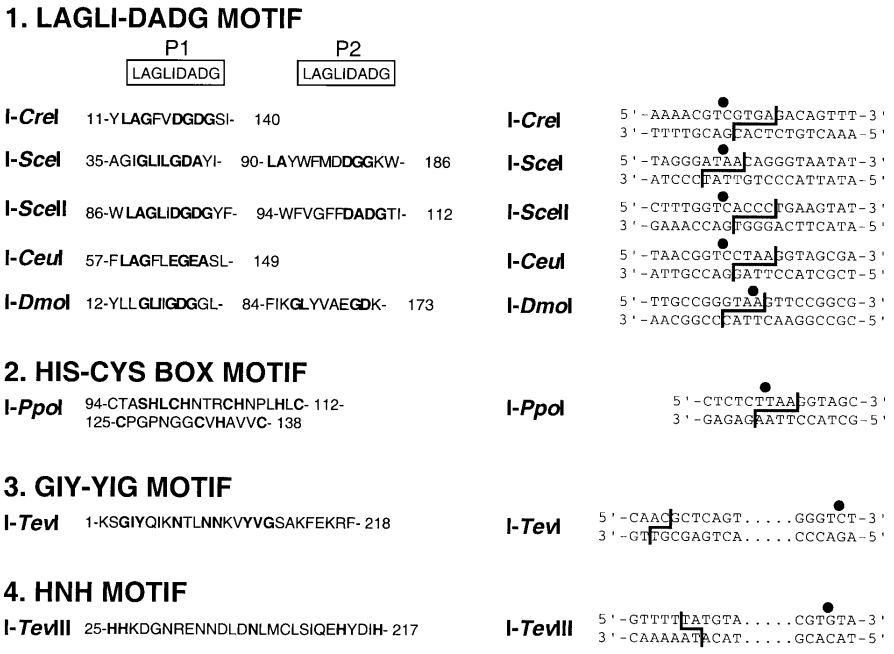

Figure 4. Representative proteins encoded by ORFs found in group I introns. Four families have been characterized. Of these, crystal structures have been reported for members of two families (LAGLIDADG and His-Cys box endonucleases). Shown on the left are examples of the conserved namesake motifs for these proteins, and on the right the length, sequence, cleavage patterns and intron insertion site (closed circles) for their DNA target sites.

Ca^{2+}, Co^{2+} and Zn^{2+}. Homing site binding and cleavage are salt-dependent, with K_d's ranging from 1 to 100 nM in the presence of 10–275 mM NaCl. Cleavage is particularly efficient, with a k_{cat}/K_m of 10^8 M^{-1} s^{-1} for I-*Ppo*I. Cleavage occurs at the center of the homing site to generate 4 nt, 3′ overhangs. The structure of I-*Ppo*I in complex with a synthetic DNA homing site has been determined at high resolution within the past year [51], and is also summarized in the second half of this review.

The GIY-YIG family. The GIY-YIG family includes intron-encoded and traditionally encoded endonucleases from phage and fungi. For these proteins, the cleavage and intron insertion sites are well separated, resulting in homing sites of up to 40 bp long [52]. In contrast, LAGLIDADG and His-Cys box endonuclease homing sites are coincident or nearly so. The two best-studied enzymes in the GIY-YIG family are the T4 phage endonucleases I-*Tev*I and I-*Tev*II, both of which are encoded within mobile group I introns. I-*Tev*I possesses a bipartite structure with separable DNA binding and nuclease domains separated by a flexible linker, similar to type IIs restriction endonucleases such as *Fok*I [55–57]. This hinged protein is a monomer with the catalytic domain containing the GIY-YIG motif at the amino terminus and a carboxy-terminal DNA binding domain [53]. The protein binds with significant sequence tolerance across the homing site [20, 52]. The enzyme significantly distorts its bound DNA homing site and cleaves to generate 2 nt, 3′ extensions. Unlike the LAGLIDADG and His-Cys box endonucleases, I-*Tev*I interacts with its DNA substrate primarily via the minor groove and the phosphate backbone [52]. The enzyme is dependent on divalent cations for DSB formation, although nicks are catalyzed in the absence of added metal in vitro. Although high-resolution structures of I-*Tev*I or any other member of this family are not yet available, a significant amount of structure/function studies have been carried out on this catalyst as described later in this review.

The HNH family. The HNH proteins are the only known family of homing endonucleases that have not been characterized structurally. These proteins contain a consensus sequence spanning 30–33 residues, with two pairs of conserved His residues flanking a conserved Asn [54, 55]. Several properties of these enzymes are distinctive from other homing endonucleases. The T-even phage enzyme I-*Tev*III is reported to make a DSB with 5′ extensions unique among all the known homing enzymes [56]. Similarly atypical are the *Bacillus subtilis* phage enzymes I-*Hmu*I and I-*Hmu*II, which cleave only one strand of the substrate DNA [57]. The endonuclease domains found in some of the multifunctional proteins encoded within mobile group II introns (such as I-*Sce*V, I-*Sce*VI and I-*Lla*I) contain the HNH motif. These proteins form an RNP complex with their corresponding spliced intron RNA [35, 58]. Both the protein and RNA components of this complex participate directly in DNA cleavage as described previously. The colicins, antibacterial nucleases from *E. coli*, also have an HNH motif [54, 55], as do a very small number of the known intein-associated homing endonucleases [59, 60]. The presence of the HNH motif in intron-encoded, intein-encoded and traditionally encoded endonucleases likely reflects the successful exploitation and divergence of this motif from a common ancestral HNH protein.

Homing site recognition: length, asymmetry and flexible specificity

Homing endonucleases face two somewhat contradictory selective pressures. First, these proteins must display sufficient sequence specificity to avoid the cleavage of essential genes within the host genome. However, it is beneficial that the mobile element be capable of insertion into sequence variants of the wild-type homing site in order to promote lateral transfer between closely related host species. Homing endonucleases generally are prevented from nonspecific cleavage of the host genome by three properties. First, these enzymes read out homing site sequences of anywhere from 14 to 30 bp in length (up to 40 for the GIY-YIG proteins), with 20 bp being the average (fig. 4) [2–4, 7]. Even though many of the base pairs within these sites may be mutated individually while still allowing recognition and cleavage, the remaining specificity caused by the long site is sufficient to substantially decrease nonspecific cleavage frequencies across the host genome. Second, most homing endonucleases are localized to specific compartments within the cell (mitochondrial, chloroplast and nuclear) that keep them effectively sequestered from the entire complement of genomic DNA. Finally, very low expression levels of these endonucleases in cells may also suppress nonspecific binding and cleavage. For most of these proteins, overexpression in common bacterial and yeast strains induces substantial cell lysis and death.

Below we summarize sequence specificity data for the enzymes for which extensive structural and biochemical information is available, for the purpose of illustrating the general principles of homing-site recognition. This includes the LAGLIDADG endonucleases I-*Cre*I, PI-*Sce*I and I-*Dmo*I, the His-Cys box endonuclease I-*Ppo*I and the GIY-YIG enzyme I-*Tev*I.

The LAGLIDADG family: I-SceI, I-DmoI, I-CreI and PI-SceI. The first homing site characterized was for I-*Sce*I, formerly known as the ω-transposase [61]. This endonuclease recognizes a pseudopalindromic 30-bp homing site, and approximately 50% of single-base changes in this site block endonuclease activity. The

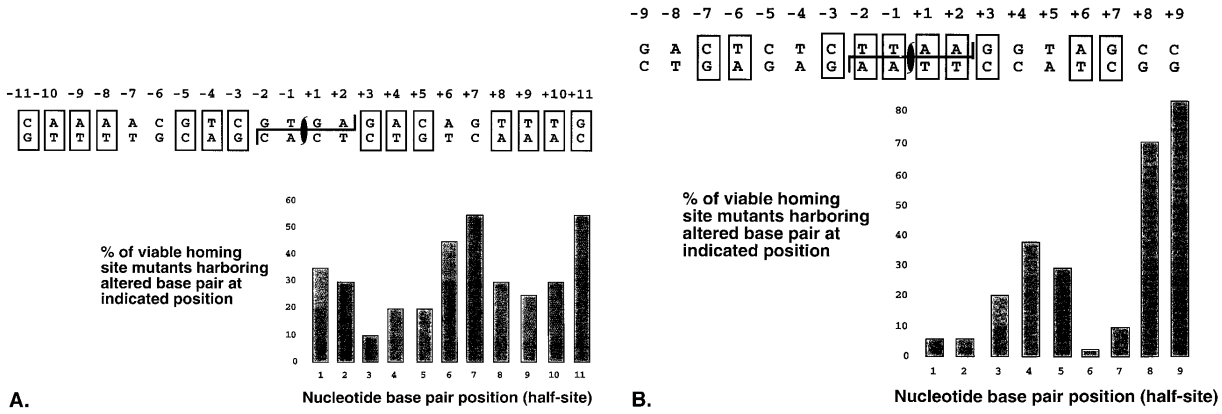

Figure 5. Sequence specificity for two representative homing endonucleases. (*A*) Homing site recognition by I-*Cre*I. The enzyme recognizes a degenerate palindrome 22 bp in length and cleaves across the minor groove to liberate cohesive 3′ ends of length 4 bases each. Top: The homing site found in the rDNA target gene. The positions that are conserved between half-sites are indicated by boxes. Bottom: Frequency of mutations observed in homing site randomization studies that select active, cleavable DNA variants (data compiled from Argast et. al. [15]). Individual mutations are tolerated at all positions; in many active mutants two to four sites are altered. Mutational frequencies in these studies are generally lower at positions that are not conserved between half-sites in the rDNA gene as expected for a homodimeric endonuclease. (*B*) Homing site recognition by I-*Ppo*I. This homodimeric enzyme also recognizes a degenerate palindrome of 16 bp and cleaves to generate 4 nt, 3′ overhangs. The frequencies of mutations in viable homing site variants are from the same experiments described in panel *A*.

estimated cleavage frequency for I-*Sce*I across a random DNA sequence is approximately once in every 10^7 bp. This endonuclease recognizes one of the longest homing sites among the free-standing LAGLIDADG enzymes. In contrast, the I-*Dmo*I endonuclease recognizes a much shorter 14-bp minimal homing site, with up to 20 bp being in contact with the protein in the cleavage complex [17]. For both enzymes, a majority of contacts are made in the major groove, as shown by alkylation interference assays.

Of the free-standing LAGLIDADG endonucleases, the I-*Cre*I enzyme has perhaps been the best characterized in terms of recognition specificity and flexibility (fig. 5A) [15, 41, 62–64]. The enzyme recognizes a 22-bp pseudopalindrome and cleaves across the minor groove to give 4 nt, 3′ overhangs. The overall symmetry between half-sites is greater for this enzyme than for I-*Sce*I and I-*Dmo*I (as expected, since I-*Cre*I is a homodimer), with the palindrome broken at bp $+/-$ 1, 2, 6 and 7. In vitro enrichment results have been reported for this homing site, in which the DNA target was randomly mutated at anywhere from 1 to 10 bp per site, and viable mutants in a cleavage-dependent screen were recovered and sequenced. The results indicate that any one base pair may be mutated while maintaining at least a minimal degree of binding and activity, and that in some mutants 2–4 sites may be altered. This gives an approximate estimate of random cleavage frequency of one in every 4×10^6 bp, which is specific enough to prevent unwanted DSBs across the chloroplast genome where this enzyme is expressed. Interestingly, the positions that are most flexible in this screen (that are most often mutated in viable homing sites) correspond to the positions that are not symmetrically conserved between half-sites. These positions also correlate well with the structural details of the protein/DNA contacts as described below.

Finally, the LAGLIDADG endonucleases that are not free-standing, but are instead translated in frame with a surrounding intein domain, exhibit similar site-recognition properties, with the additional feature of longer-range contacts that are made by a small region of the intein itself. Of these proteins, the PI-SceI intein/endonuclease has been particularly well characterized [42, 65]. Genetic and biochemical studies of this system reveal that the enzyme makes numerous base-specific and phosphate-backbone contacts with its 31-bp asymmetrical site. This site can be divided into two regions, both of which contain nucleotides that are essential for cleavage by PI-SceI. Region I contains the cleavage site, and is contacted throughout the major groove by residues in the endonuclease domain, whereas region II includes an adjacent sequence 3′ of the cleavage site that is contacted by several residues in the intein domain. This sequence is critical for high-affinity binding and activity, and the contacts are made primarily along the phosphate backbone. Binding induces significant distortions to the DNA site in both regions, in a manner that remains to be fully characterized structurally.

The His-Cys box endonucleases: I-PpoI. Befitting its status as the first known homing endonuclease in this family, I-PpoI has been the primary target of studies investigating its pattern of site recognition [15, 47–51]. The endonuclease binds and induces a bend in a partially symmetric, 14-bp minimal homing site and makes additional nonspecific contacts that extend the length of the protein-DNA interface out to 20 bp (fig. 5B). A genetic screen to assess homing site specificity was carried out for this enzyme identical to that described above for I-CreI. The results are strikingly similar, with the exceptions that specificity extends only to bp $+/-$ 7 on each half-site and that one position ($+/-$ 6) appears to be relatively inviolate. In addition, the central six base pairs of the homing site (between and immediately flanking the scissile phosphates) display a greater degree of sequence specificity than that observed for I-CreI, and also greater than expected solely on the basis of nucleotide/side chain contacts. It is probable that this additional sequence specificity is related to the distortions of the DNA substrate across this region of the homing site. The estimated overall frequency of cutting for this site is approximately one in every 2.5×10^6 bp.

The GIY-YIG endonucleases: I-TevI. Studies of the I-TevI system have provided the basis for understanding a homing endonuclease that exhibits more promiscuous protein/DNA interactions [20, 52, 66]. Although I-TevI-mediated cleavage and intron homing are precise, the endonuclease is extremely tolerant of base substitutions. No single base within a 40-bp region is essential for enzyme cleavage, and in fact the base pairs in this site are among the most flexible of all known homing endonucleases. As a result, a nonconventional randomization strategy was used to probe specificity. A substrate with 15% degeneracy over 48 bp was used in a selective-enrichment screen to identify hypomutable regions involved in homing site recognition. These studies, along with DNAse I and hydroxyl radical footprinting, have identified a primary region flanking the intron insertion site as the primary binding domain, with a secondary region of contact that approaches and directly borders the scissile phosphates. In addition, homing site variants containing methylated nucleotides are cleaved with wild-type efficiency, indicating that I-TevI is tolerant of these major groove modifications. These data, together with interference studies, indicate that I-TevI binds DNA in a sequence-tolerant manner across the minor groove in both of these contacted regions, in contrast with the LAGLIDADG and His-Cys box endonucleases.

Relatives of intron-encoded homing endonucleases

The sequence motifs that are the hallmarks for intron-encoded homing endonucleases are also found in several other protein families. Before proceeding to a review of recent structural studies of group I intron-encoded homing endonucleases, it is worthwhile to briefly summarize these families and their relationships to intron-encoded homing endonucleases. Included are proteins involved in RNA splicing (maturases), in the genetic mobility of a novel type of intervening sequence (inteins), and in phenotype conversion via a nuclear gene conversion event (yeast HO endonuclease). The presence of homologues of homing endonucleases outside of introns, in some cases displaying biochemical activities other than DNA cleavage, implies that these proteins have been coopted successfully by some organisms for purposes different from their role in intron mobility.

Maturases. In 1980, mutational analyses suggested that excision of a group I intron (cob-I3) from the yeast mitochondrial cytochrome b gene required expression of an intron-encoded protein [67–69]. These proteins, usually termed maturases, appear to act as allosteric cofactors that stabilize RNA structure in a reactive conformation for splicing [70–73]. Maturases induce and/or accelerate the formation of secondary and tertiary structures of the self-splicing intron. Interestingly, many of these factors appear to serve double duty as enzyme catalysts in addition to their role in RNA splicing. These proteins show sequence similarity with intron-encoded homing endonucleases, particularly members of the LAGLIDADG family, and in most cases are also active endonucleases [5, 74, 75].

There is strong experimental evidence that for those maturases that are homologous to homing endonucleases; the enzymatic and splicing activities of these proteins are very closely associated. An intron ORF in *Saccharomyces capensis* encodes a bifunctional protein with both activities [75]; an ORF in a mobile intron in *Schizosaccharomyces pombe* (presumed to encode an endonuclease) encodes a maturase [74]; and a single point mutation in the *Saccharomyces cerevisiae* cox1-I4α homing endonuclease activates latent maturase activity [76]. Recent biochemical characterization of the I-*Ani*I protein from *Aspergillus nidulans* directly demonstrates both site-specific endonuclease activity and RNA splicing activity in vitro [77]. These studies clearly indicate that maturase and homing endonuclease activities are closely linked evolutionarily and in some cases have only very recently diverged.

Group II introns have also been observed to contain ORFs encoding proteins involved both in splicing and in mobility (fig. 2B). Unlike group I intron-encoded proteins, the group II encoded-proteins tend to be large and multifunctional, with independent maturase, endonuclease and RT domains [6, 78–81]. The maturase domains bear little or no resemblance to the group I-encoded homing endonucleases or maturase splicing

factors described above. However, the activity of the maturase domain is critical for successful transfer of the intron, because homing proceeds through reverse splicing of the splicing intron and formation of a covalently linked RNA-DNA species [27]. As compared with the maturases associated with mobile group I introns, relatively little structural detail has been gathered for the group II maturase protein domain.

Inteins: self-splicing protein domains and homing endonucleases. Protein splicing was first described in 1990 by Kane et al., who reported the unusual processing of the 69-kDa yeast vacuolar H^+ ATPase catalytic subunit from a larger, 120-kDa precursor protein encoded by the yeast *TFP1* (*VMA1*) gene [82]. Sequencing of the gene encoding this protein, and comparison with a *Neurospora* homologue, indicated the presence of a novel intervening sequence encoding a 50-kDa protein sequence flanked by the N- and C-terminal sequences of the ATPase. Studies of this system arrived at the startling conclusion that this novel type of intervening sequence (subsequently termed intein for internal protein) is transcribed and translated in a proper reading frame with the flanking sequences of the ATPase (termed exteins for external proteins). Rather than being excluded from the final protein product by posttranscriptional processing of the messenger RNA, as in intron splicing, the intein is posttranslationally removed from the protein host by a peptide splicing event, liberating the functional ATPase and the excised intein. Self-splicing intein sequences have since been demonstrated to exist in a variety of organisms and host protein genes (including DNA and RNA polymerases, helicases, metabolic enzymes and DNA gyrase A subunits (gyrA), and to be spliced from translated protein products with different exteins and in heterologous systems [83, 84].

The biology of inteins is complicated by the observation that many of the known inteins are themselves interrupted by genetic inserts encoding homing endonucleases (fig. 1) [83–85]. For example, the VMA1 intein contains a monomeric LAGLIDADG endonuclease domain and is now called PI-*Sce*I. The endonuclease is translated in frame with the surrounding intein domains and is present as an insert bulged out from a surface loop of the intein. Removal of the endonuclease domain from the intein by genetic methods does not generally eliminate protein splicing activity, although splicing may be reduced [86, 87]. However, the C-terminal intein domain in PI-*Sce*I makes important DNA contacts downstream of the endonuclease cleavage site, demonstrating that the endonuclease domains of this and perhaps other inteins do not solely determine their site specificity [65, 88]. It seems that for inteins, as for group I and group II introns, homing endonucleases mediate highly specific transfer of their own host sequences to cognate alleles lacking that specific genetic element. The actual number of inteins that are mobile in genetic crosses, and the mobility of these sequences in vivo, has not been studied thoroughly. However, the high degree of homology of the DnaB intein, cloned from the *Rhodothermus marinus* and *Synechocystis* genomes, supports the recent horizontal transfer of this element between remotely related organisms [89]. This observation indicates that homing of at least some inteins may be as efficient as that of mobile introns.

Gene conversion enzymes from free-standing genes (yeast HO endonuclease). HO endonuclease (HO) initiates a mating-type switch in *S. cerevisiae* by making a site-specific DSB in the mating-type locus *MAT*. Laboratory strains of yeast have an inactive *ho* allele and maintain a stable mating type. HO is a member of the dodecamer (LAGLIDADG) family of homing endonucleases [90]. Unlike its intron-encoded cousins, however, HO is encoded by a nuclear gene and contains an inactive, vestigial intein. The final intein motif is replaced in HO by a 120-residue sequence at the C-terminus that appears to encode a novel zinc-finger domain. This motif is implicated in DNA binding, again indicating that the endonuclease domain alone does not fully define the sequence specificity for the endonuclease [91]. Analysis of the HO protein, and of the intein-encoded endonucleases, indicates the propensity of these proteins for domain swaps and exchanges, and the routine use of additional structural 'cassettes' (such as the HO C-terminal domain) to provide additional levels of activity and/or specificity to the protein.

Structural studies

Structural studies of group I homing endonucleases from three of the four general families described above have been reported recently. These include the LAGLIDADG enzymes I-*Cre*I, PI-*Sce*I and I-*Dmo*I, the His-Cys enzyme I-*Ppo*I and the GIY-YIG enzyme I-*Tev*I. These studies allowed investigators to describe how the enzymes have found both related and unique mechanisms to address several common aspects of their role in intron homing: recognition of long sequences by relatively small proteins, relaxed sequence specificity and nucleolytic cleavage across the minor groove of DNA. We review the five structural examples with these issues in mind by first summarizing the overall structural features of the enzymes. This summary is followed by a review of their interactions with DNA, and a description of their nuclease active sites and possible catalytic mechanisms. Finally, a general comparison with restriction endonucleases provides a broader context for the important structural features of site-specific cleavage of DNA.

Overall structural motifs: short protein sequences yield elongated protein folds and novel subunit packing strategies

The high-resolution crystal structures of three LAGLIDADG and one His-Cys box endonuclease have been reported recently (fig. 6). Perhaps the most interesting feature of these structural studies is the evolution of stable, extended protein folds that allow relatively small protein subunits or domains to form long interfaces across lengthy DNA homing sites. These proteins also exhibit novel strategies of dimerization or domain interface packing, and display preformed DNA binding mo-

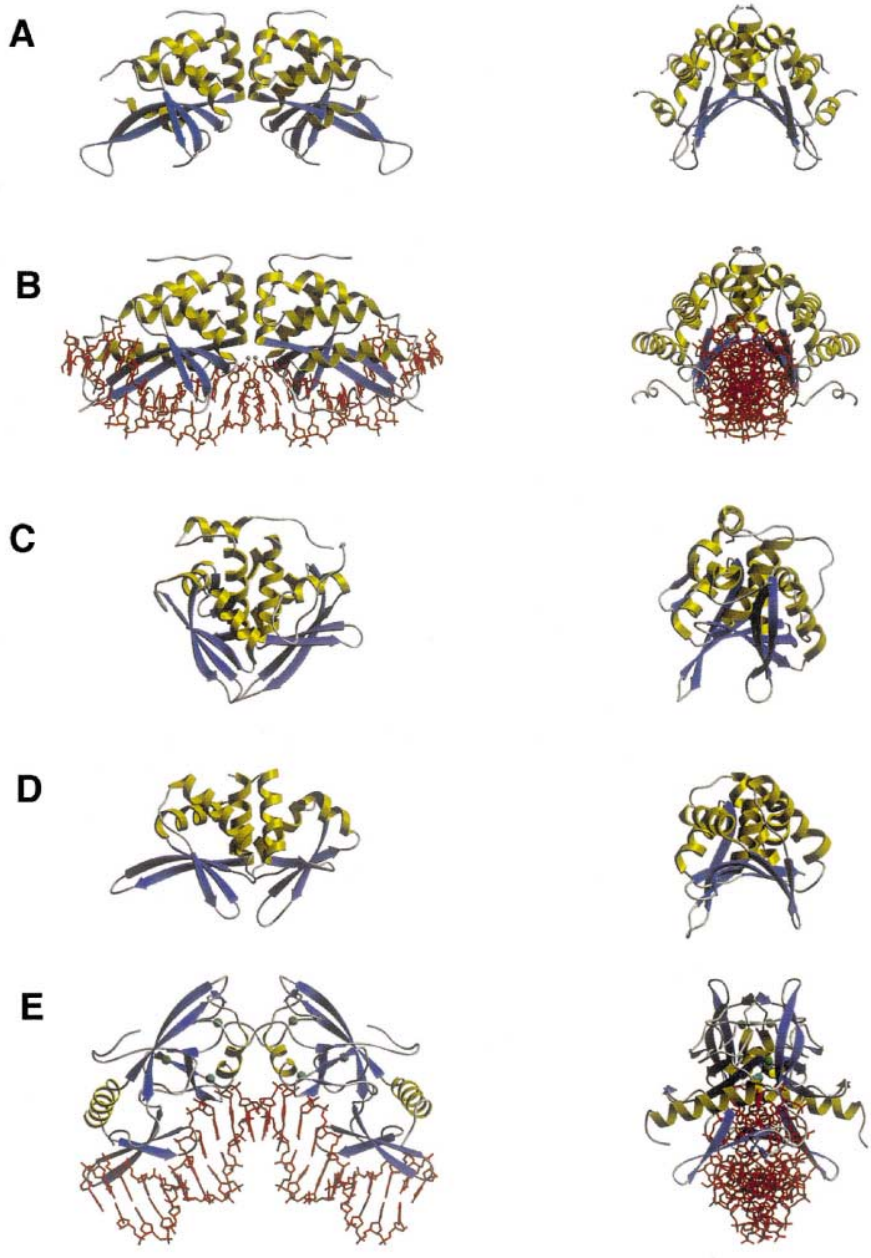

Figure 6. Ribbon diagrams for I-CreI apo enzyme (*A*), I-CreI complexed to DNA (*B*), the endonuclease domain of PI-SceI (*C*), I-DmoI (*D*) and I-PpoI complexed to DNA (*E*). The left images of each panel are looking perpendicular to the DNA binding axis and to the dimer/domain symmetry axis, while the right panels are looking down the protein symmetry axis.

tifs, consisting of antiparallel β-ribbons, capable of making extended major groove contacts with the DNA target site.

LAGLIDADG endonucleases. The structures of two LAGLIDADG enzymes, I-*Cre*I (a free-standing homodimer, fig. 6A) and PI-*Sce*I (an intein-associated monomer, fig. 6C), provided the first atomic view of homing endonucleases [40, 42]. The subsequent structures of the I-*Cre*I/DNA complex (fig. 6B) and I-*Dmo*I (a free-standing, thermostable monomer, fig. 6D) add a new level to our understanding of this largest family of homing endonucleases [41, 43]. These studies also provide some important clues for delineating the phylogeny of these enzymes that have greatly diverged in terms of sequence but not basic structure. I-*Cre*I is a 163-residue protein encoded by an ORF located in a group I intron of the *Chlamydomonas reinhardtii* chloroplast 23S rRNA gene [23, 92–94]. The endonuclease of PI-*Sce*I is a distinct domain (228 residues) in an intein encoded within the vacuolar adenosine triphosphate (ATP) synthase of *S. cerevisae* [95, 96]. Both the intein and endonuclease domains are visualized in the crystal structure. The I-*Dmo*I ORF encoding a 188-residue enzyme is found in an archaeal intron in the 23S rRNA gene of *Desulfurococcus mobilis* [97, 98]. All three enzymes cleave their DNA targets to yield 4 nt, 3′ overhangs.

The long dimension of the I-*Cre*I dimer is approximately 80 Å along the DNA binding axis. The topology of an I-*Cre*I monomer is $\alpha\beta\beta\alpha\beta\beta\alpha\alpha\alpha$ (fig. 6A). All but the two C-terminal α-helices of this topology are repeated twice in the PI-*Sce*I endonuclease domain and in I-*Dmo*I, resulting in a pseudo-twofold symmetry in these monomeric endonucleases. The first α-helix of this topology contains the LAGLIDADG sequence and forms the primary contacts within the dimer interface. The four β-strands form an extended antiparallel β-sheet that is curved to form an extended groove that represents the binding site for DNA. The structure of this β-groove, particularly the curvature and twist of the strands, is stabilized by packing interactions in the protein core between the top of this β-sheet and the second two α-helices of the topology. Strands $\beta 1$ and $\beta 2$ are shaped and positioned to follow the major groove of the DNA along the exceptionally long target site. In the I-*Cre*I/DNA complex these β-strands are observed to make contacts over eight consecutive base pairs. The β–strands extend every other side chain into the major groove to mediate DNA recognition, whereas basic residues on the loops flanking these strands contact the DNA phosphate backbone. In I-*Cre*I, dimerization symmetrically doubles the length of this groove to extend over the 22-bp I-*Cre*I homing site. The I-*Cre*I recognition site exhibits a corresponding, though imperfect, twofold symmetry.

The DNA binding groove described above is present in the monomeric PI-*Sce*I and I-*Dmo*I structures, but in contrast to I-*Cre*I there is a pronounced asymmetry in both structures. Although both monomeric enzymes display the internal pseudo-symmetry described above, one endonuclease domain is larger, and its corresponding β-sheet longer, than its sister domain. This corresponds to a more pronounced asymmetry in sequence of the homing sites of these enzymes than is observed for I-*Cre*I. Additionally, the order of these domains differs: in PI-*Sce*I the larger domain is C-terminal, whereas the larger domain of I-*Dmo*I is N-terminal. One possible conclusion from these observations is that the monomeric LAGLIDADG proteins have arisen from homodimeric ancestors by two or more distinct gene fusion events [43]. The fusion of two LAGLIDADG domains into a monomeric enzyme allowed the domains to evolve in an asymmetric manner. This event has also caused the active sites of the enzyme monomers to exhibit heightened structural diversity, as described below.

The I-*Cre*I DNA binding groove is approximately 75-Å long and makes sequence-specific contacts over a 22-bp homing site. The I-*Dmo*I groove is 50 Å in length, corresponding to its shorter recognition sequence of 14 bp. The endonuclease groove in PI-*Sce*I is approximately 60 Å long, but the intein portion of the protein is also involved in DNA recognition, increasing the length of the specific homing site to 31 bp. The relative positions of the endonuclease and intein domains of PI-*Sce*I cannot accommodate this recognition site as unbent B-form DNA, and gel-shift data indicate a bend of up to 60° for DNA bound to the full protein [65]. Modeling studies of DNA on the structure of I-*Dmo*I also indicate that the substrate must be significantly perturbed in order to make reasonable contacts with the β-sheets of both enzyme domains [43]. In contrast, similar experiments indicate little substrate bending when I-*Cre*I binds to DNA (R. Monnat, personal communication). This is confirmed in the protein/DNA complex structure.

The conserved LAGLIDADG motif is central to the structure and function of these endonucleases, and plays a role in protein folding, subunit or domain packing, and formation of the active sites of the enzyme. In the I-*Cre*I homodimer, the single LAGLIDADG sequence from each monomer form a pair of N-terminal α-helices that pack in parallel to form a majority of the dimer interface. These helices also contribute a pair of catalytic residues, Asp20 and Asp20′, into the center of the DNA binding interface directly flanking the domain interface. The separation of these aspartate residues is approximately 10 Å, similar to the

spacing between the scissile phosphates of the DNA substrate. In the structures of the PI-*Sce*I and I-*Dmo*I monomers, the two copies of the LAGLIDADG motif (separated by about 100 residues) pack in the same manner as in the I-*Cre*I dimer to form a domain interface and closely packed active sites. The packing of the LAGLIDADG helices is somewhat unusual in these structures, as it is not created by 'knobs into grooves' packing of side chains, but by direct van der Waals contacts between the protein backbone of the helices along the interface. This close packing, plus the tight turn facilitated by the C-terminal glycine at the end of the helices, positions the highly conserved acidic residue from each helix (either an aspartic or glutamic acid) in close proximity. Each of these residues coordinates a divalent metal ion in two closely spaced active sites that are positioned appropriately for cleavage across the central minor groove to liberate 4 nt, 3′ overhangs.

His-Cys box endonucleases. The 18-kDa I-*Ppo*I homing endonuclease is encoded by a mobile group I intron in the nuclear 23S rRNA gene from *Physarum polycephalum* [24, 47]. This enzyme also cleaves its substrate to yield 4 nt, 3′ overhangs. The structure of this His-Cys box enzyme bound to its homing site DNA demonstrates how a very different protein architecture utilizes some of the same principles seen in the LAGLIDADG enzymes to recognize its 14-bp homing site. Unlike the LAGLIDADG endonucleases, which display extended protein folds surrounding long β-ribbon platforms, I-*Ppo*I possesses a far more loosely packed tertiary core structure stabilized by bound zinc atoms. The conserved His-Cys box motifs of this enzyme do not play a direct role in DNA binding or dimerization, but are instead important in creating zinc binding sites that stabilize the overall fold of the protein. I-*Ppo*I lacks a tightly packed hydrophobic core, and the tightly bound zinc ions substitute as 'ionic staples' to hold the tertiary fold in place. The first bound zinc ion is coordinated by a Cys_3-His_1 ligand cluster (C-X_{58}-C-X_4-C-X_4-H) from the N-terminal end of the protein. The first cysteine is contributed by $\beta 2$ and the remaining three side chains by a short loop between $\beta 7$ and $\beta 8$. The second zinc ion is coordinated by a cluster of four closely spaced side chains (C-X_6-C-X_1-H-X_3-H), all donated from a short buried protein loop. All eight residues involved in zinc coordination are conserved as Cys or His residues among the three nuclear homing endonucleases that comprise the known His-Cys family. The strictly conserved tetrapeptide sequence $Ser_{97}His_{98}Leu_{99}Cys_{100}$ (SHLC) contains one zinc-coordinating residue (Cys 100) that is part of the first zinc binding motif and one side chain (His 98) that is located in the enzyme active site.

Unlike the LAGLIDADG enzymes, which utilize a closely packed subunit interface to cleave across the narrow DNA minor groove, I-*Ppo*I displays a very loosely packed dimer interface, and more widely separated active sites. The central dimer interface of the protein buries a very small surface area (700 Å^2) and is highly solvated. Dimerization is stabilized by a domain-swapped C-terminal tail of 17 residues that extends across the opposing monomer in a molecular 'hug' that buries another 900 Å^2 across the surface of each subunit. In order to allow cleavage across the minor groove, distortion of the DNA in the active site is necessary, as described below.

I-*Ppo*I significantly distorts bound DNA to widen the minor groove at the cleavage site and make the scissile phosphates accessible to the enzyme's pair of active sites. The distortion results in an overall 55° bend of the DNA across the homing site, as compared with the 38° predicted by gel-shift analysis for bending [48]. The enzyme places antiparallel β-strands into the major groove of DNA to either side of the minor groove cleavage site to mediate recognition, similar to the LAGLIDADG endonucleases. The β-sheet created by these strands does not form the extended groove seen in those enzymes, but still conforms to the major groove. In the interface, sequence-specific contacts are made to a shorter length of the DNA homing site (14 bp), but the elongated dimer extends 80 Å to cover this entire recognition sequence with an additional 3 bp at either end of the substrate.

GIY-YIG endonucleases. Studies of the 28-kDa GIY-YIG endonuclease I-*Tev*I reveal yet another unique approach to DNA recognition and cleavage. The homing endonuclease is encoded by a T4 bacteriophage group I *td* intron [66]. The endonuclease has a bipartite structure with separate catalytic and DNA binding domains joined by a flexible linker [53]. This has been concluded from limited proteolysis experiments, independent expression of the stable folded domains and footprinting studies of the protein on DNA [52, 53]. The DNA binding domain recognizes a sequence of about 20 bp, which also contains the insertion site for the *td* intron. This sequence is separated from the upstream cleavage site by 23 to 25 bp. The protein thus binds a span of DNA 37 bp in length. Limited insertions and deletions between the recognition and cleavage sites are tolerated [20, 99]. Preliminary nuclear magnetic resonance (NMR) studies of the N-terminal catalytic domain reveal that its folded structure extends to residue 92 [100]. The domain has mixed α/β topology, with a single three-stranded antiparallel β-sheet. This β-sheet contains the conserved GIY-YIG residues. Mutation of either tyrosine residue or the final glycine of this motif results in loss of catalytic activity, as well as destabilization of the folded structure of the domain.

Sequence alignment of the related GIY-YIG enzymes reveals three additional invariant residues within the catalytic domain. Two of these, Arg27 and Glu75, are present at undetermined positions of α-helices in the domain. Mutation of either of these residues results in a catalytically inactive but well-folded domain [100].

DNA binding and recognition

The complex structures of I-*Cre*I and I-*Ppo*I with their DNA-binding sequences provide insight into the recognition of relatively long homing sites with a specificity that allows for some variability within the site (fig. 7). This somewhat relaxed specificity is likely necessary for dealing with differences between homologous alleles of the host gene for the mobile intron. Both enzymes employ the flexibility of β-strands to follow the DNA major groove along an extended sequence. The spacing between the alternating side chains extended from one side of a β-strand is nearly equal to that between every other DNA base pair [101]. Thus, a staggered two-stranded sheet paralleling the major groove can make contacts to each consecutive base pair. I-*Ppo*I extends four side chains from the β-strands sitting in the major groove to contact five base pairs in each recognition half-site, while I-*Cre*I uses a long loop in conjunction with the β-strands to contact nine consecutive base pairs (fig. 7).

Sequence specificity for DNA binding is mediated by the unique pattern of hydrogen bond donors and acceptors presented by base pairs in the major groove. Each base-pair type in a specific orientation presents a different combination of three or four hydrogen bond donors and acceptors. Protein contacts with two of these are usually sufficient to specify the preferred base pair at a certain position in the protein-DNA interface. A single contact is less specific, and often two different base-pair identities can satisfy the hydrogen bond contact. This principle is reflected in the base contacts and specificities exhibited by both I-*Cre*I and I-*Ppo*I (fig. 8). Neither protein makes saturating contacts with all possible hydrogen bond donors and acceptors presented by the homing sequence throughout the major groove of the DNA. I-*Cre*I makes two or more contacts per base pair at three positions in a recognition half-site. These positions are conserved between the two half-sites of the pseudo-palindromic DNA target and have been shown to be least tolerant to mutation [15]. In addition to these relatively specific contacts, the enzyme makes a single base-pair contact at five other positions within the half-site. For example, Gln26 acts as a hydrogen bond donor to the N7 of adenine in the A-T base pair at position $+6$. In the opposite half-site the same residue in the other half of the dimer, Gln26′, donates a proton to the N7 of guanine in the G-C base pair at position -6. Neither a T-A or C-G base pair at that position would present the N7 hydrogen bond donor. In randomization studies of the recognition site, these two identities (T-A or C-G) at this position are not recovered [15]. Similar results are seen at the other positions with a single protein contact. I-*Ppo*I shows the same behavior: base pairs that make two or more contacts to the protein are highly conserved, whereas positions with a single protein contact are found with two of the four possible base-pair identities. This is illustrated in the structure at positions that either hold or break the palindrome of the recognition sequence, as well as by homing site mutation studies [15].

Based solely on protein DNA contacts, the specificities of these enzymes can be calculated as follows: I-*Cre*I, $(1/4)^6 \times (1/2)^{10}$, which corresponds to finding a homing site once in every 4.2 million base pairs; I-*Ppo*I, $(1/4)^6 \times (1/2)^6$, which corresponds to finding a homing site once in every 2.6 million base pairs. The infrequency of these homing sites is mediated by their length, but is balanced by the flexibility created with subsaturating contacts.

Active site structures and cleavage mechanisms

The mechanism of nucleolytic cleavage requires the positioning of an activated nucleophile, usually a deprotonated water molecule, for inline attack of the electrophilic 5′ phosphate. Positive charge is necessary to stabilize the pentacoordinate intermediate, as is a catalytic acid for donation of a proton to the 3′ oxygen leaving group. The role of divalent metal in this reaction is usually thought to alter the pK_a of a bound water for easy deprotonation, and to position it for nucleophilic attack. The positive charge of the metal can also serve to stabilize the increased negative charge at the phosphate in the reaction transition state, and can increase the stability of a leaving group prior to protonation. Most nuclease catalysts are described in terms of their use of metals in the reaction, with different enzymes utilizing either single-metal or two-metal mechanisms to facilitate cleavage. Analysis of the two homing endonuclease/DNA cocrystal structures currently available (I-*Cre*I and I-*Ppo*I) clearly indicates that the two enzyme families represented by these proteins utilize quite different catalytic mechanisms. Furthermore, the LAGLIDADG family exhibits a striking degree of structural divergence within the active sites of its most well characterized members.

Both I-*Cre*I and I-*Ppo*I cleave DNA to yield 4nt, 3′ overhangs, which corresponds to cleaving the antiparallel strands at positions that sit across the minor groove. In unbent B-form DNA the scissile phosphates are spaced less than 9 Å apart. When one considers the catalytic machinery necessary for nucleolytic cleavage,

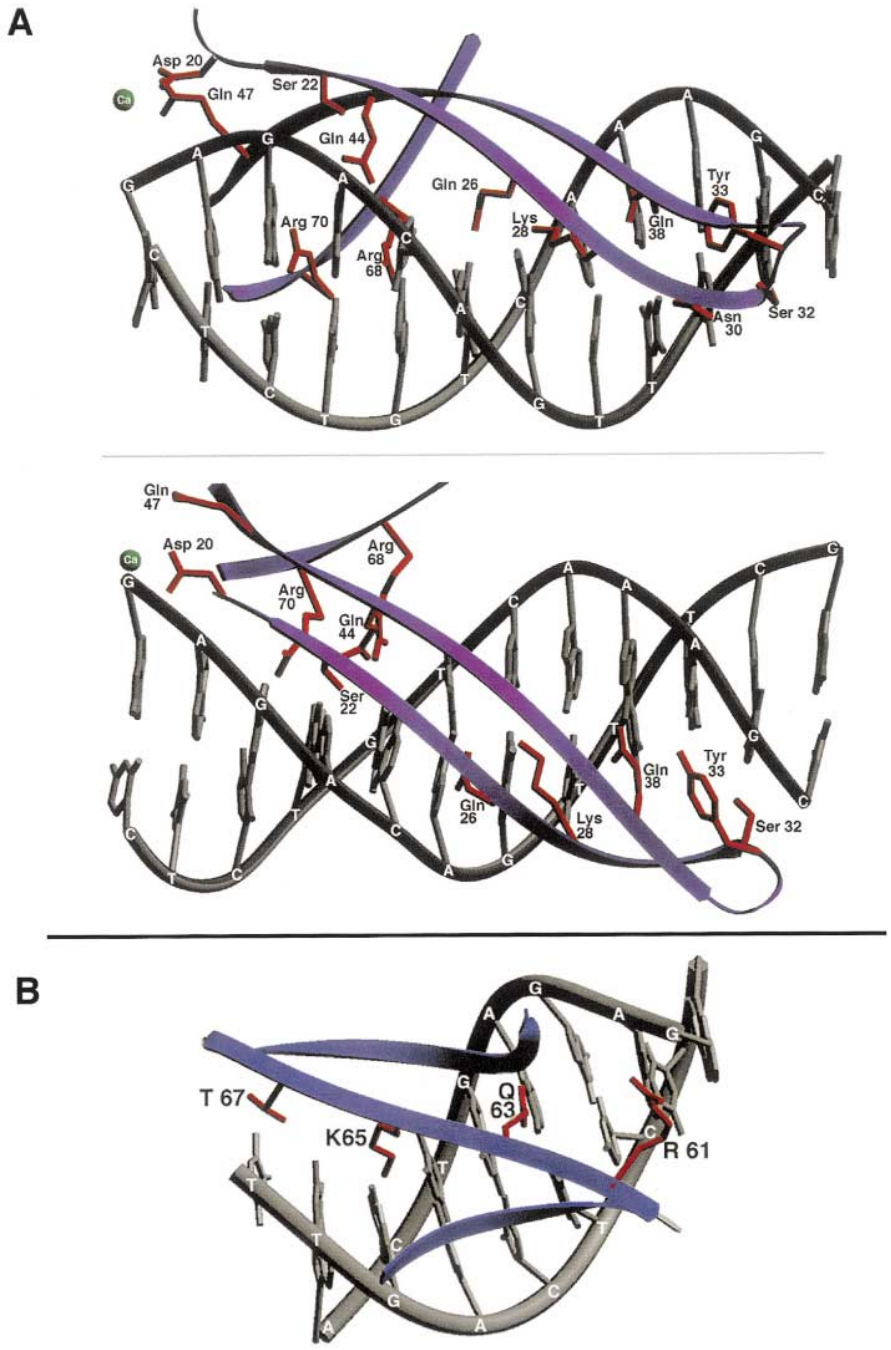

Figure 7. Endonuclease-DNA interfaces. (*A*) Contacts between the I-*Cre*I endonuclease and its homing site (contacts are shown for a protein monomer and DNA half-site). Two views of the contacts between nucleotide bases in a homing half site and side chains from the antiparallel β-ribbon formed by strands $\beta 1$, $\beta 2$ and $\beta 4$ of a single protein subunit are shown. The $\beta 3$-strand also contributes side chains to the DNA interface, but none make direct contact with nucleotide bases, and are therefore not shown to improve clarity. (*B*) Contacts between the His-Cys box endonuclease I-*Ppo*I and a DNA homing half-site. The majority of contacts are made by side chains from the $\beta 4$-strand in the center of the antiparallel β-ribbon.

the close spacing of the scissile bonds presents something of a steric challenge. In contrast, enzymes that cleave across the major groove have a space of over 20 Å between the cleavage sites in which to fit a set of two

complete active sites. The LAGLIDADG and His-Cys box homing endonucleases appear to have adopted radically different solutions to the steric problem of cleavage across the DNA minor groove. I-*Cre*I addresses this issue through the close packing of the LAGLIDADG helices at the dimer interface described earlier. This arrangement of tightly packed helices, with their sharp turns leading away from the dimer interface, positions the catalytic side chains Asp20 and Asp20′ into the DNA binding groove, where they each independently coordinate a divalent cation. Each divalent cation is also coordinated by an oxygen from one of the scissile phosphates. Thus the enzyme places the catalytic metals of two separate active sites within 10 Å of one another. This observation once again underlines the structural importance of the LAGLIDADG motif for protein folding, subunit or domain packing, and for catalysis. I-*Ppo*I takes a very different approach to this problem. The DNA bound to the protein is severely bent and distorted, resulting in a dramatic widening of the minor groove at the central cleavage site. This results in separating the scissile phosphates to nearly 20 Å apart. No special packing is necessary to form the two independent active sites.

The LAGLIDADG endonuclease family. The I-*Cre*I/DNA complex gives a complete view of the active site (fig. 9A) with the caveats that the Ca^{2+} ions used for crystallization inhibit cleavage, and the 3.0 Å resolution limit prevents visualizing solvent molecules. Superposition of this structure with the structures of the I-DmoI and PI-SceI endonuclease (both determined in the absence of bound DNA) aids in identifying important catalytic residues (fig. 9B). Biochemical and genetic studies that identify mutations affecting catalysis in these LAGLIDADG enzymes, as well as sequence alignments of over 130 LAGLIDADG-containing sequences also aid the characterization of the active sites and catalytic mechanisms of these endonucleases. Yet even with this large amount of information, a well-conserved active site structure for this enzyme family is not

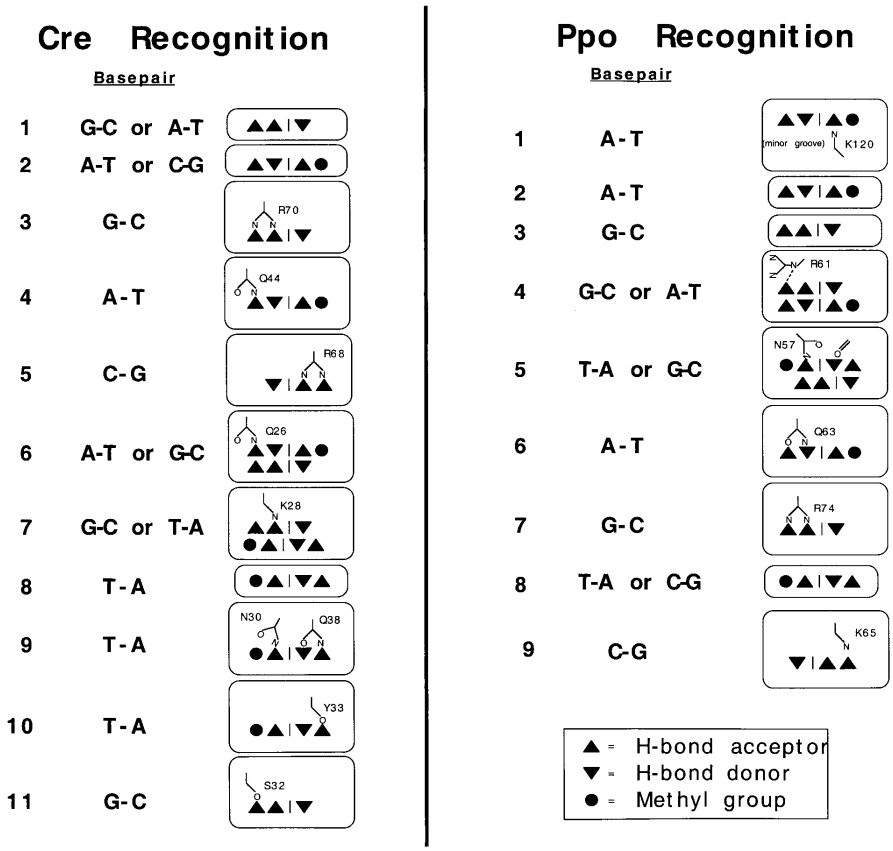

Figure 8. Schematic representation of sequence-specific contacts made by I-*Cre*I (left) and I-*Ppo*I (right). The sequences shown represent a recognition half-site with the sequence specific hydrogen bond donors and acceptors presented in the major groove. Where the palindrome of the recognition sequence is broken both nucleotide identities are shown.

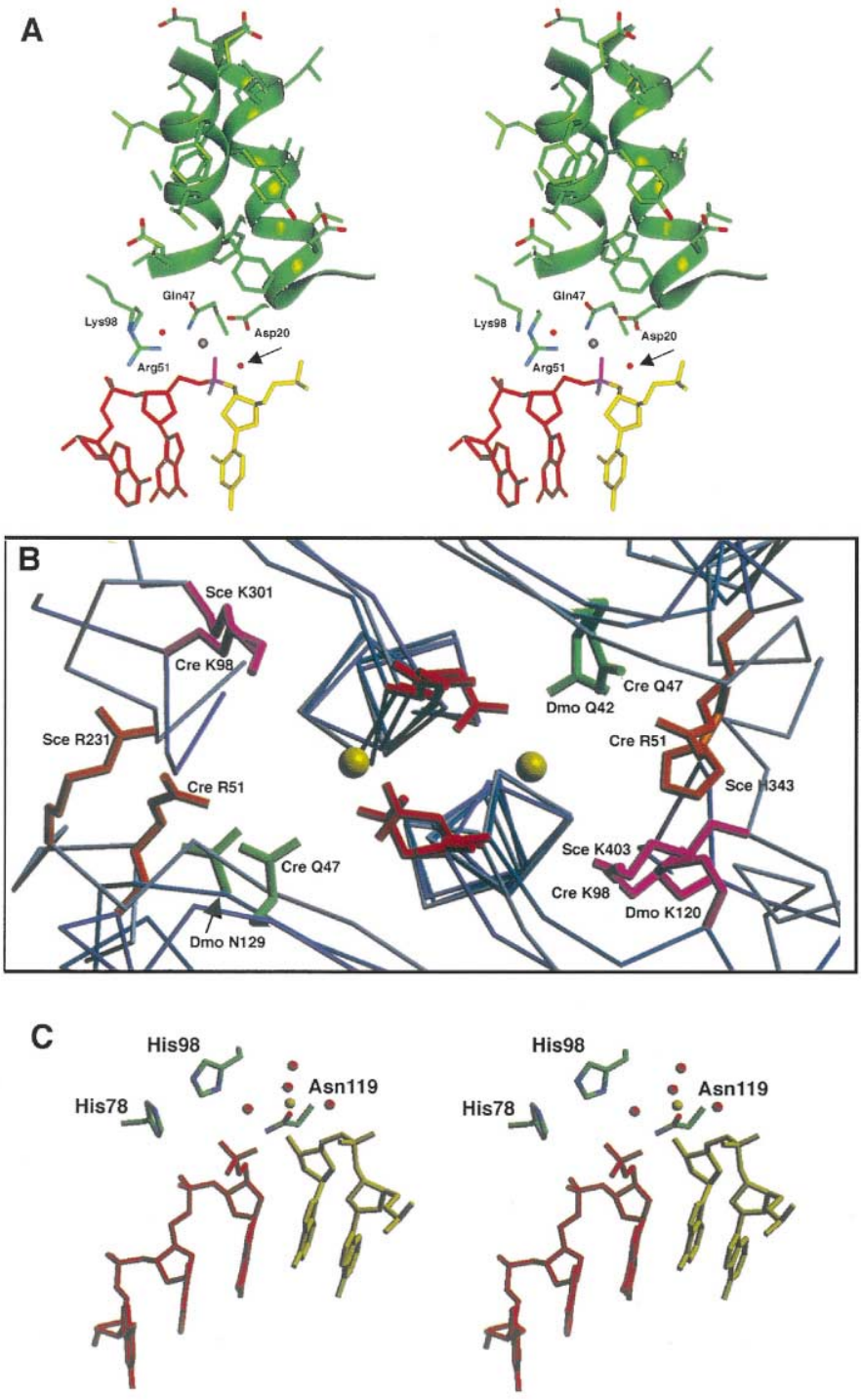

Figure 9. Homing endonuclease active sites. (*A*) Stereo view of the active site for I-*Cre*I/DNA complex. Atoms are shown for a single DNA strand in the half-site complex, the LAGLIDADG-containing protein helices and the most likely catalytic side chains. The four nucleotide bases that form the single-stranded overhang after cleavage are colored red; the scissile phosphate group is purple and the adjoining nucleotide base (Gua3, that retains the cleaved phosphate) is yellow. The bound calcium ion is a gray sphere, and is coordinated as described. Two water molecules are shown, and the potential nucleophilic water is pointed out. (*B*) Superposition of active sites of I-*Cre*I, PI-*Sce*I endonuclease domain and I-*Dmo*I. Residues conserved in the symmetric active sites are colored the same. (*C*) Stereo view of the active site of I-*Ppo*I product complex. The nucleotide bases that form the single-stranded 3′ overhang after cleavage are colored red; the 5′ nucleotide bases are yellow. Water molecules are red, and the bound magnesium ion is a yellow sphere and is coordinated as described.

apparent; instead the active sites of the LAGLIDADG endonucleases appear to be widely diverged.

In the I-*Cre*I/DNA structure there are a number of side chains within the vicinity of the scissile bond that are potential catalytic groups. Only the acidic side chain of the LAGLIDADG sequence that coordinates the divalent cation in the I-*Cre*I/DNA structure (Asp20) is conserved in all three structures. Pairwise comparisons between structures show that a glutamine residue (Gln47) is conserved between I-*Cre*I and both I-*Dmo*I active sites (Gln42, Gln129). Mutation of this residue in I-*Cre*I and the homologous residue in another LAGLIDADG enzyme, I-*Ceu*I (Gln93), abolishes catalytic activity [63, 102]. No homologous residue is present in either of the two PI-*Sce*I active sites. Conversely, a lysine residue is conserved between I-*Cre*I (Lys98 and 98′) and both PI-*Sce*I active sites (Lys301, Lys403), but is not conserved in I-*Dmo*I. There is a lysine in I-*Dmo*I from another region of secondary structure that could act as a base at one active site, but there is nothing that could substitute at the other site. Mutation of the lysine in I-*Cre*I (Lys98) abolishes catalytic activity [63]. Mutation of the homologous residue in either PI-*Sce*I active site also decreases catalytic activity, although perhaps not as dramatically for the second active site (Lys 403) [88]. The position of Arg51 in I-*Cre*I, a side chain also identified in mutation screens, is near that of Arg231 in one of the PI-*Sce*I active sites. His343 is at the other pseudodimeric position in PI-*Sce*I and is also sensitive to mutation.

The sequences of the LAGLIDADG enzymes are highly divergent. This motif has been identified in over 130 sequences, but the extreme divergence of the sequences has made multiple sequence alignment very challenging. A recent alignment of the entire sequence family by Dalgaard et al. aligns the LAGLIDADG sequences very well, but does not align all the active site residues that superimpose in the structures [39]. The alignment does reveal that lysine residues are conserved in most (but not all) family members at or near the position of Lys98 of the I-*Cre*I sequence. Gln47 is conserved in a large number of the sequences as well. Sequence analysis of the LAGLIDADG domains within inteins has focused mainly on identification of conserved blocks [59]. Four blocks have been identified that in the PI-*Sce*I structure correspond to the two LAGLIDADG repeats (EN1, EN3), to the end of $\alpha 3$ leading into the region linking the two 'halves' of the pseudo-dimeric monomer (EN2) and to the $\alpha 2'$ helix (EN4). Lys301 is at the second position of the EN2 block and is highly conserved.

Higher-resolution structures of these enzymes complexed with DNA will help in a more detailed characterization of the catalytic mechanism. Sequence alignments that utilize restraints derived from structural superpositions may help to generalize this mechanism to other family members. The existing structures underscore the possibility that the active sites of LAGLIDADG enzymes have diverged to employ different residues for the role of proton donor and general base in the reaction. A majority of the LAGLIDADG enzymes are monomeric and contain two copies of the motif. The PI-*Sce*I and I-*Dmo*I structures display opposite arrangements of large and small domains along their peptide sequences, which suggests that these enzymes arose from different gene fusions of separate, ancestral monomers. After a fusion event, activity of the resulting enzymes could be regained and optimized through subsequent random mutations. These mutations would differ among groups of enzymes derived from separate monomeric ancestors, and lead to modern LAGLIDADG endonucleases that display large diversity across their active sites. Such mutations would be accommodated and allowed more frequently in these mobile genes than in essential reading frames because their endonuclease products are under no selective pressure to maintain function for host survival. It appears that these genes are more rapidly diverging under different evolutionary constraints than those that act on essential host genes.

The single cocrystal structure available (I-*Cre*I complexed with DNA and calcium, which inhibits cleavage) indicates two metal ions bound at the protein dimer interface. Each ion is independently coordinated by a single aspartic acid residue (Asp 20 from the LAGLIDADG), a single oxygen from a scissile phosphate and an additional oxygen atom from the phosphate located directly across the minor groove. The main-chain carbonyl of residue 19 from the opposing monomer and at least one water molecule complete the coordination sphere around the cations. This water molecule is loosely contacted by Gln47, a catalytically important residue as shown by mutagenesis studies. The direct oxygen contacts to the bound metal exhibit normal calcium bond distances of 2.2–2.5 Å, although the coordination is not strictly octahedral. There is no indication of a bound metal between the symmetry-related aspartate residues. Two different side chains, Arg51 and Lys98, are located in the enzyme active site and are candidates to either stabilize the pentacoordinate transition state or to activate a proton donor in the cleavage reaction. A water molecule coordinated to the bound metal ion would be the most appropriately positioned candidate to act as an inline nucleophile in this reaction. Based on this structure, we favor a mechanism involving single metal ion-assisted hydrolysis of the phosphodiester bond, with separate active sites responsible for each cleavage event. The architecture of the enzyme active site, in complex with DNA, is reminiscent of type II restriction endonucleases. In these en-

zymes, acidic side chains are positioned near the scissile phosphate to bind a divalent cation and water molecules in an activated complex, and a positively charged side chain is positioned to stabilize the pentacoordinate phosphoanion transition state. This strategy appears to have been loosely reproduced in the LAGLIDADG homing endonucleases, with an aspartate residue from the LAGLIDADG motif (Asp 20) and an essential glutamine residue (Gln47) participating in metal binding, and a basic residue (probably Lys 98) interacting with the labile phosphate.

The His-Cys box family (I-PpoI). The His-Cys box enzymes lack an extensive published body of biochemical data and large sequence family to aid characterization of the active site. However, the high resolution of the I-*Ppo*I structures complexed with both substrate and product DNA conformations allow for a detailed model of catalysis. These structures reveal the coordination sphere of the divalent cation in the active site, the position of several solvent molecules, and either the presence of the scissile phosphodiester bond in the substrate structure or its absence in the product structure (fig. 9C). The substrate conformation was trapped in the crystal by using an oligonucleotide in which the 3′ leaving group oxygen was replaced with a sulfur atom, a much poorer leaving group.

The current model for catalysis by I-*Ppo*I is a unique single-metal mechanism for inline attack in which a bound water molecule is activated by a pair of conserved histidine residues, His98 and His78, one of which (His98) also participates in transition-state stabilization. His98 is absolutely conserved in the SHLC sequence found in the His-Cys box endonucleases. The leaving group would be stabilized through a direct interaction with a bound metal ion, as visualized in the structures. The metal ion is positioned by the enzyme via an interaction with Asn119. The current questions regarding the structural mechanism of the enzyme are the exact identity of the base that activates the nucleophilic water, the proton donor for the leaving group, and the role of substrate bending and deformation during the cleavage reaction.

The use of an asparagine (or glutamine) residue as a metal-binding side chain in an endonuclease active site, and the use of a histidine residue to activate a bound water or to act as a Lewis acid to stabilize a phosphooxyanion transition state are not unique. A catalytic mechanism proposed for a nonspecific nuclease from *Serratia* also employs a histidine side chain to deprotonate and activate the nucleophile water and an asparagine to help coordinate the divalent cation [103]. The apo structure of this enzyme has been solved, and while the tertiary fold is completely unrelated to that of I-*Ppo*I [104], 23 residues with the regions of 93 to 119 in I-PpoI and 84 to 119 in Serratia nuclease can be superimposed with a root mean squared difference (rmsd) of 1.14 Å. The two most important catalytic residues of Serratia nuclease, H89 and N119, have counterparts in H98 and N119 in I-PpoI [105]. It thus appears that I-*Ppo*I and the *Serratia* nuclease have independently arrived at this use of side-chain chemistries and very similar active site architectures. Recent enzymological studies lend support for this novel mechanism for both proteins [106].

In addition, previous mutational studies of the 3′,5′ exonuclease of DNA polymerase I have demonstrated that mutation of a metal-binding aspartate side chain (residue 501) to an asparagine does not significantly reduce enzyme activity [107]. The data indicate an acidic side chain may be substituted by a carboxyamide group, provided that only one oxygen atom is involved in metal binding and catalysis. Finally, the use of a conserved histidine residue in ribonuclease A to abstract a proton from the 2′ ribose oxygen to activate it for nucleophilic attack argues that a histidine residue could almost certainly deprotonate and activate a bound water molecule in an endonuclease [108].

Restriction endonucleases versus homing endonucleases

Site-specific cleavage of DNA plays an important role in a number of biological processes. The most widely studied system involving site-specific DNA cleavage is the phenomenon of restriction, as exemplified by the type II restriction endonucleases. A comparison of homing endonucleases with the type II restriction endonucleases illustrates how different evolutionary pressures are reflected in structure. Restriction endonucleases work in a biological defense system and must quickly cleave invading phage DNA before it can start its destructive biological program. The host genome is protected by a complementary methylation system that allows the endonuclease to discriminate between self and host. Restriction endonucleases usually recognize, with high specificity, short and often palindromic sequences generally 4–8 bp long. Under physiological conditions, mutation of any base pair in the restriction site virtually eliminates cleavage by the endonuclease [109, 110]. This specificity is important for directing cleavage to unprotected foreign DNA. The short unmethylated recognition sites have a high probability of being found in the foreign DNA. The function of homing endonucleases, on the other hand, is to facilitate the self-propagation of the intron that encodes them while ensuring nontoxicity to the host genome. These enzymes display sequence-specific recognition for much longer target site, and tolerate individual changes of many base pairs within the site. This flexibility in recognition must be important for recognizing the homing site in cognate alleles of the host gene that are not

isomorphous. Even with the degeneracy allowed in the recognition site, the length of the site ensures that the endonuclease will not cleave the host genome extensively, and possibly only once at the homing site.

The difference between high specifity for a short recognition site versus less specificity for a long recognition site can be visualized when comparing the cocrystal structure of these endonucleases with their cognate DNA. The first difference is in the overall dimension and shape of proteins as they contact DNA. Restriction endonucleases are larger and highly folded globular proteins that engulf the DNA as they bind [109, 110]. The smaller homing endonucleases, which must extend over a long recognition site, have a lower profile on the DNA and a less compact fold. This is likely related to the number of contacts made to base pairs in both the major and minor grooves. The restriction endonucleases nearly saturate the hydrogen bond donors and acceptors in the major groove of their recognition sites. They often make additional contacts in the minor groove as well overspecify the recognition sequence for the enzyme. As described earlier, the homing endonucleases make subsaturating contacts to the DNA over a longer distance. They tolerate different base-pair identities at many of the positions within the site to balance specificity with site variability. The homing enzymes seem to use a minimum of structure to accomplish their task, relying on β-sheets for extended major groove recognition. There may be evolutionary pressure limiting the size of these proteins, presumably because the ORFs that encode them are located in loops of self-splicing introns or inteins. Restriction endonucleases, in contrast, are less size restricted but must instead work to fit a larger number of side chains into the major groove to satisfy nearly all the available hydrogen bond donors and acceptors. Most of the enzymes do this via side chains from α-helices and loops that fit into the major groove. It may take a more extensive scaffold to position these structures.

Future directions

The group I intron-encoded homing endonucleases have proved to be amenable and fascinating systems for structural studies. The many mechanistic, biophysical and evolutionary issues surrounding these proteins should be addressed through continued biochemical and structural studies. Within the group I intron families, the GIY-YIG proteins are only beginning to be structurally characterized. The bipartite protein organization (consisting of separable catalytic and binding domains) and long connecting linker exhibited by I-*Tev*I is of interest for the design of chimeric proteins with unique combinations of DNA-recognition and cleavage activity. As mentioned above, the HNH family is as yet structurally uncharacterized. This family may display a divergence of catalytic structure and mechanism that parallels or exceeds the LAGLIDADG family. Additionally, splicing factors, or maturases, displaying homology to homing endonucleases (and in many cases bearing both activities) are excellent targets for structural analysis because of their fascinating combination of specific binding and biochemical activities, and because of widespread interest in proteins involved in RNA processing and structural organization. Finally, it is clear that the protein factors involved in group II intron mobility will also be strong candidates for structural studies. In particular, the maturase domains of these proteins promise to be completely unique from their group I cousins and may aid structural studies of group II introns. Additionally, the formation of an RNP between the endonuclease domain and the spliced RNA intron, and the use of both RNA and protein atoms in DSB formation, should richly reward a detailed structural analysis.

The structural mechanisms of catalysis displayed by these endonucleases (at least as modeled on the basis of structural studies of the single LAGLIDADG and His-Cys box endonucleases visualized bound to DNA) appear to be unique variations of the metal-dependent nuclease mechanisms described for polymerases and restriction endonucleases. For these enzymes, it will be important to analyze different liganded states that approximate individual reaction steps in the cleavage pathway and compare them with other nucleases. Such comparative crystallographic studies, for example those of cleaved and uncleaved I-*Ppo*I complexes and similarly of cleaved and uncleaved *Bam*HI complexes, have allowed detailed analysis of the role and dynamic motion of individual side chains, metal groups and solvent molecules in DNA cleavage.

Finally, these endonucleases, by virtue of their rare cutting activity, are potentially useful molecules for the development of tools for analyzing and manipulating genomes. Such tools could be used in mapping, cloning and gene targeting. These proteins are therefore being studied for their potential targeting to human cells and as systems that can be rationally or combinatorially redesigned for alternative sequence recognition patterns. As with all forms of 'genetic engineering' (a term once in vogue that is less commonly invoked now), such applications will depend heavily on continued genetic, biochemical and structural studies of these versatile enzyme catalysts.

Acknowledegments. The authors thank several of the cited authors for sharing data and results prior to publication (particularly Alfred Pingoud, Marlene Belfort, Victoria Derbyshire and Patrick VanRoey) and the Van Roey laboratory for providing coordinates

for the I-*Dmo*I endonuclease prior to release. We also thank Vicki Derbyshire, Monica Parker and Ray Monnat for critical reading of this review prior to submission. B.L.S. is funded for this work by the NIH (GM49857); M.S.J. is funded through a PHS National Research Grant T32 GM07270.

1 Dujon B. (1989) Group I introns as mobile genetic elements: facts and mechanistic speculations – a review. Gene **82:** 91–114
2 Belfort M. and Roberts R. J. (1997) Homing endonucleases – keeping the house in order. Nucleic Acids Res. **25:** 3379–3388
3 Belfort M. and Perlman P. S. (1995) Mechanisms of intron mobility. J. Biol. Chem. **270:** 30237–30240
4 Belfort M., Reaban M. E., Coetzee T. and Dalgaard J. Z. (1995) Prokaryotic introns and inteins: a panoply of form and function. J. Bacteriol. **177:** 3897–3903
5 Lambowitz A. M. and Belfort M. (1993) Introns as mobile genetic elements. Annu. Rev. Biochem. **62:** 587–622
6 Lambowitz A. M., Caprara M. G., Zimmerly S. and Perlman P. S. (1998) Group I and group II ribozymes as RNPs: clues to the past and guides to the future. In: RNA World II, vol. 37, pp. 451–485 Gesteland, R. F., Cech, T. R. and Atkins, J. F. (ed.), Cold Spring Harbor Press, Cold Spring Harbor, NY
7 Mueller J. E., Brysk M., Loizos N. and Belfort M. (1993) Homing endonucleases. In: Nucleases, vol. 2, pp. 111–143, Linn S. M., Lloyd R. S. and Roberts R. J. (eds), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
8 Coen D., Deutsch J., Netter P., Petrochilo E. and Slonimski P. P. (1970) Mitochondrial genetics. I. Methodology and phenomenology. Symp. Soc. Exp. Biol. **23:** 449–496
9 Dujon B. (1980) Sequence of the intron and flanking exons of the mitochondrial 21S rRNA gene of yeast strains having different alleles at the omega and rib-1 loci. Cell **20:** 185–197
10 Bos J. L., Heyting C., Borst P., Arnberg A. C. and Van Bruggen E. F. (1978) An insert in the single gene for the large ribosomal RNA in yeast mitochondrial DNA. Nature **275:** 336–338
11 Zinn A. R. and Butow R. A. (1985) Nonreciprocal exchange between alleles of the yeast mitochondrial 21S rRNA gene: kinetics and the involvement of a double-strand break. Cell **40:** 887–895
12 Colleaux L., d'Auriol L., Betermier M., Cottarel G., Jacquier A., Galibert F. et al. (1986) Universal code equivalent of a yeast mitochondrial intron reading frame is expressed into *E. coli* as a specific double strand endonuclease. Cell **44:** 521–533
13 Macreadie I. G., Scott R. M., Zinn A. R. and Butow R. A. (1985) Transposition of an intron in yeast mitochondria requires a protein encoded by that intron. Cell **41:** 395–402
14 Jacquier A. and Dujon B. (1985) An intron-encoded protein is active in a gene conversion process that spreads an intron into a mitochondrial gene. Cell **41:** 383–394
15 Argast G. M., Stephens K. M., Emond M. J. and Monnat R. J. J. (1998) I-*Ppo*I and I-*Cre*I homing site sequence degeneracy determined by random mutagenesis and sequential in vitro enrichment. J. Mol. Biol. **280:** 345–353
16 Aggarwal A. K. and Wah D. A. (1998) Novel site-specific DNA endonucleases. Curr. Opin. Struct. Biol. **8:** 19–25
17 Aagaard C., Awayez M. J. and Garrett R. A. (1997) Profile of the DNA recognition site of the archaeal homing endonuclease I-*Dmo*I. Nucleic Acids Res. **25:** 1523–1530
18 Lykke-Andersen J., Thi-Ngoc H. P. and Garrett R. A. (1994) DNA substrate specificity and cleavage kinetics of an archaeal homing-type endonuclease from *Pyrobaculum organotrophum*. Nucleic Acids Res. **22:** 4583–4590
19 Perrin A., Buckle M. and Dujon B. (1993) Asymmetrical recognition and activity of the I-*Sce*I endonuclease on its site and on intron-exon junctions. EMBO J. **12:** 2939–2947
20 Bryk M., Belisle M., Mueller J. E. and Belfort M. (1995) Selection of a remote cleavage site by I-*Tev*I, the *td* intron-encoded endonuclease. J. Mol. Biol. **247:** 197–210
21 Quirk S. M., Bell-Pedersen D. and Belfort M. (1989) Intron mobility in the T-even phages: high frequency inheritance of group I introns promoted by intron open reading frames. Cell **56:** 455–465
22 Quirk S. M., Bell-Pedersen D., Tomaschewski J., Ruger W. and Belfort M. (1989) The inconsistent distribution of introns in the T-even phages indicates recent genetic exchanges. Nucleic Acids Res. **17:** 301–315
23 Lemieux B., Turmel M. and Lemieux C. (1988) Unidirectional gene conversions in the chloroplast of *Chlamydomonas* inter-specific hybrids. Mol. Gen. Genet. **212:** 48–55
24 Muscarella D. E. and Vogt V. M. (1989) A mobile group I intron in the nuclear rDNA of *Physarum polycephalum*. Cell **56:** 443–454
25 Wenzlau J. M., Saldanha R. J., Butow R. A. and Perlman P. S. (1989) A latent intron-encoded maturase is also an endonuclease needed for intron mobility. Cell **56:** 421–430
26 Michel F. and Ferat J.-L. (1995) Structure and activities of group II introns. Ann. Rev. Biochem. **64:** 435–461
27 Curcio M. J. and Belfort M. (1996) Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. Cell **84:** 9–12
28 Bell-Pedersen D., Quirk S. M., Aubrey M. and Belfort M. (1989) A site-specific endonuclease and co-conversion of flanking exons associated with the mobile *td* intron of phage T4. Gene **82:** 119–126
29 Clyman J. and Belfort M. (1992) *Trans* and *cis* requirements for intron mobility in a prokaryotic system. Genes Dev. **6:** 1269–1279
30 Mueller J. E., Clyman J., Huang Y. J., Parker M. M. and Belfort M. (1996) Intron mobility in phage T4 occurs in the context of recombination-dependent DNA replication by way of multiple pathways. Genes Dev. **10:** 351–364
31 Lazowska J., Meunier B. and Macadre C. (1994) Homing of a group II intron in yeast mitochondrial DNA is accompanied by unidirectional co-conversion of upstream-located markers. EMBO J. **13:** 4963–4972
32 Moran J. V., Zimmerly S., Eskes R., Kennell J. C., Lambowitz A. M., Butow R. A. et al. (1995) Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. Mol. Cell. Biol. **15:** 2828–2838
33 Kennell J. C., Moran J. V., Perlman P. S., Butow R. A. and Lambowitz A. M. (1993) Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. Cell **73:** 133–146
34 Zimmerly S., Guo H., Perlman P. S. and Lambowitz A. M. (1995) Group II intron mobility occurs by target DNA-primed reverse transcription. Cell **82:** 545–554
35 Zimmerly S., Guo H., Eskes R., Yang J., Perlman P. S. and Lambowitz A. M. (1995) A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. Cell **83:** 529–538
36 Yang J., Mohr G., Perlman P. S. and Lambowitz A. M. (1998) Group II intron mobility in yeast mitochondria – target DNA-primed reverse transcription activity of al1 and reverse splicing into DNA transposition sites in vitro. J. Mol. Biol. **282:** 505–523
37 Yang J., Zimmerly S., Perlman P. S. and Lambowitz A. M. (1996) Efficient integration of an intron RNA into double-stranded DNA by reverse splicing [see comments]. Nature **381:** 332–335
38 Cousineau B., Smith D., Cavanagh S. L., Mueller J. E., Yang J., Mills D. et al. (1998) Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. Cell **94:** 451–462
39 Dalgaard J. Z., Klar A. J., Moser M. J., Holley W. R., Chatterjee A. and Mian I. S. (1997) Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-spe-

cific endonuclease of the HNH family. Nucleic Acids Res. **25:** 4626–4638
40 Heath P. J., Stephens K. M., Monnat R. J. and Stoddard B. L. (1997) The structure of I-*Cre*I, a group I intron-encoded homing endonuclease. Nature Struct. Biol. **4:** 468–476
41 Jurica M. S., Monnat R. J. and Stoddard B. L. (1998) DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-*Cre*I. Mol. Cell **2:** 469–476
42 Duan X., Gimble F. S. and Quiocho F. A. (1997) Crystal structure of PI-*Sce*I, a homing endonuclease with protein splicing activity. Cell **89:** 555–564
43 Silva G. H., Dalgaard J. Z., Belfort M. and Roey P. V. (1999) Crystal structure of the thermostable archaeal intron-encoded endonuclease I-*Dmo*I. J. Mol. Biol. **286:** 1123–1136
44 Johansen S., Embley T. M. and Willassen N. P. (1993) A family of nuclear homing endonucleases. Nucleic Acids Res. **21:** 4405–4411
45 Muscarella D. E., Ellison E. L., Ruoff B. M. and Vogt V. M. (1990) Characterization of I-*Ppo*, an intron-encoded endonuclease that mediates homing of a group I intron in the ribosomal DNA of *Physarum polycephalum*. Mol. Cell. Biol. **10:** 3386–3396
46 Flick K. E., McHugh D., Heath J. D., Stephens K. M. Jr., R. J. M. and Stoddard B. L. (1997) Crystallization and preliminary X-ray studies of I-*Ppo*I: a nuclear, intron-encoded homing endonuclease from *Physarum polycephalum*. Protein Sci. **6:** 2677–2680
47 Ellison E. L. and Vogt V. M. (1993) Interaction of the intron-encoded mobility endonuclease I-*Ppo*I with its target site. Mol. Cell. Biol. **13:** 7531–7539
48 Whittmayer P. K. and Raines R. T. (1996) Substrate binding and turnover by the highly specific I-*Ppo*I endonuclease. Biochemistry **35:** 1076–1083
49 Wittmayer P. K., McKenzie J. L. and Raines R. T. (1998) Degenerate DNA recognition by I-*Ppo*I endonuclease. Gene **206:** 11–21
50 Lowery R., Hung L., Knoche K. and Bandziulis R. (1992) Properties of I-*Ppo*I: a rare-cutting intron-encoded endonuclease. Promega Notes **38:** 8–12
51 Flick K. E., Jurica M. S., Monnat R. J. and Stoddard B. L. (1998) DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-*Ppo*I. Nature **394:** 96–101
52 Bryk M., Quirk S. M., Mueller J. E., Loizos N., Lawrence C. and Belfort M. (1993) The *td* intron endonuclease I-*Tev*I makes extensive sequence-tolerant contacts across the minor groove of its DNA target. EMBO J. **12:** 4040–4041
53 Derbyshire V., Kowalski J. C., Dansereau J. T., Hauer C. R. and Belfort M. (1997) Two-domain structure of the *td* intron-encoded endonuclease I-*Tev*I correlates with the two-domain configuration of the homing site. J. Mol. Biol. **265:** 494–506
54 Shub D. A., Goodrich-Blair H. and Eddy S. R. (1994) Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. Trends Biochem. Sci. **19:** 402–404
55 Gorbalenya A. E. (1994) Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. Protein Sci. **3:** 1117–1120
56 Eddy S. R. and Gold L. (1991) The phage T4 nrdB intron: a deletion mutant of a version found in the wild. Genes Dev. **5:** 1032–1041
57 Goodrich-Blair H. and Shub D. A. (1996) Beyond homing: competition between intron endonucleases confers a selective advantage on flanking genetic markers. Cell **84:** 211–221
58 Matsuura M., Saldanha R., Ma H., Wank H., Yang J., Mohr G. et al. (1997) A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. Genes Dev. **11:** 2910–2924
59 Pietrokovki S. (1998) Modular organization of inteins and C-terminal autocatalytic domains. Protein Sci. **63:** 64–71
60 Gorbalenya A. E. (1998) Non-canonical inteins. Nucleic Acids Res. **26:** 1741–1748
61 Colleaux L., D'Auriol L., Galibert F. and Dujon B. (1988) Recognition and cleavage site of the intron-encoded omega transposase. Proc. Natl. Acad. Sci. USA **85:** 6022–6026
62 Thompson A. J., Yuan X., Kudlicki W. and Herrin D. L. (1992) Cleavage and recognition pattern of a double-strand-specific endonuclease (I-*Cre*I) encoded by the chloroplast 23S rRNA intron of *Chlamydomonas reinhardtii*. Gene **119:** 247–251
63 Seligman L. M., Stephens K. M., Savage J. H. and Monnat R. J. (1997) Genetic analysis of the *Chlamydomonas reinhardtii* I-*Cre*I mobile intron homing system in *Escherichia coli*. Genetics **147:** 1653–1664
64 Wang J., Kim H.-H., Yuan X. and Herrin D. L. (1997) Purification, biochemical characterization and protein-DNA interactions of the I-*Cre*I endonuclease produced in *Escherichia coli*. Nucleic Acids Res. **25:** 3767–3776
65 Gimble F. S. and Wang J. (1996) Substrate recognition and induced DNA distortion by the PI-*Sce*I endonuclease, an enzyme generated by protein splicing. J. Mol. Biol. **263:** 163–180
66 Bell-Pedersen D., Quirk S. M., Bryk M. and Belfort M. (1991) I-*Tev*I, the endonuclease encoded by the mobile *td* intron, recognizes binding and cleavage domains on its DNA target. Proc. Natl. Acad. Sci. USA **88:** 7719–7723
67 Van Ommen G.-J. B., Boer P. H., Goot G. S. P., Haan M. D., Roosendaal E. and Grivell L. A. (1980) Mutations affecting RNA splicing and the interaction of gene expression of the yeast mitochondrial loci cob and oxi-3. Cell **20:** 173–183
68 Perlman P. S., Mahler H. R., Dhawale S., Hanson D. and Alexander N. J. (1980) Alternate forms of the COB/BOX gene: some new observations. In: The Organization and Expression of the Mitochondrial Genome, vol. 2, pp. 161–172, Kroon A. M. and Saccone C. (eds), North-Holland Biomedical Press, New York
69 Lazowska J., Jacq C. and Slonimski P. P. (1980) Sequence of introns and flanking exons in wild-type and box3 mutants of cytochrome b reveals an interlaced splicing protein coded by an intron. Cell **22:** 333–348
70 Guo Q. and Lambowitz A. M. (1992) A tyrosyl-tRNA synthetase binds specifically to the group I intron catalytic core. Genes Dev. **6:** 1357–1372
71 Shaw L. C. and Lewin A. S. (1995) Protein-induced folding of a group I intron in cytochrome b pre-mRNA. J. Biol. Chem. **270:** 21552–21562
72 Weeks K. M. and Cech T. R. (1996) Assembly of a ribonucleoprotein catalyst by tertiary structure capture. Science **271:** 345–348
73 Lambowitz A. M. and Perlman P. S. (1990) Involvement of aminoacyl-tRNA synthetases and other proteins in group I and group II intron splicing. Trends. Biochem. Sci. **15:** 367–382
74 Schafer B., Wilde B., Massardo D. R., Manna F., Del Giudice L. and Wolf K. (1994) A mitochondrial group-I intron in fission yeast encodes a maturase and is mobile in crosses. Curr. Genet. **25:** 336–341
75 Szczepanek T. and Lazowska J. (1996) Replacement of two non-adjacent amino acids in the *S. cerevisiae* bi2 intron-encoded RNA maturase is sufficient to gain a homing-endonuclease activity. EMBO J. **15:** 3758–3767
76 Dujardin G., Jacq C. and Slonimski P. P. (1982) Single base substitution in an intron of oxidase gene compensates splicing defects of the cytochrome b gene. Nature **298:** 628–632
77 Ho Y., Kim S. J. and Waring R. B. (1997) A protein encoded by a group I intron in *Aspergillus nidulans* directly assists RNA splicing and is a DNA endonuclease [published erratum appears in Proc. Natl. Acad. Sci. USA (1997) **94:** 14976]. Proc. Natl. Acad. Sci. USA **94:** 8994–8999
78 Mills D. A., McKay L. L. and Dunny G. M. (1996) Splicing of a group II intron involved in the conjugative transfer of pRS01 in *Lactococci*. J. Bacteriol. **178:** 3531–3538

79 Shearman C., Godon J.-J. and Gasson M. (1996) Splicing of a group II intron in a functional transfer gene of *Lactococcus lactis*. Mol. Microbiol. **21:** 45–53
80 Sugita M., Shinozaki K. and Sugiura M. (1985) Tobacco chloroplast tRNA-Lys (UUU) gene contains a 2.5-kilobase-pair intron: an open reading frame and a conserved boundary sequence in the intron. Proc. Natl. Acad. Sci. USA **82:** 3557–3561
81 Bonitz S. G., Coruzzi G., Thalenfeld B. E. and Tzagoloff A. (1980) Assembly of the mitochondrial membrane system. Structure and nucleotide sequence of the gene coding for subunit 1 of yeast cytochrome oxidase. J. Biol. Chem. **255:** 11927–11941
82 Kane P. M., Yamashiro C. T., Wolczyk D. F., Neff N., Goebl M. and Stevens T. H. (1990) Protein splicing converts the yeast TFP1 gene product ot the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. Science **250:** 651–657
83 Shao Y. and Kent S. B. (1997) Protein splicing: occurrence, mechanisms and related phenomena. Chem. Biol. **4:** 187–194
84 Cooper A. A. and Stevens T. H. (1995) Protein splicing: self-splicing of genetically mobile elements at the protein level. Trends Biochem. Sci. **20:** 351–356
85 Pietrokovski S. (1994) Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. Protein Science **3:** 2340–2350
86 Derbyshire V., Wood D. W., Wu W., Dansereau J. T., Dalgaard J. Z. and Belfort M. (1997) Genetic definition of a protein-splicing domain: functional mini-inteins support structure predictions and a model for intein evolution. Proc. Natl. Acad. Sci. USA **94:** 11466–11471
87 Shingledecker K., Jiang S.-Q. and Paulus H. (1998) Molecular dissection of the *Mycobacterium tuberculosis* RecA intein: design of a minimal intein and of a *trans*-splicing system involving two intein fragments. Gene **207:** 187–195
88 He Z., Crist M., Yen H., Duan X., Quiocho F. A. and Gimble F. S. (1998) Amino acid residues in both the protein splicing and endonuclease domains of the PI-*Sce*I intein mediate DNA binding. J. Biol. Chem. **273:** 4607–4615
89 Liu X.-Q. and Hu Z. (1997) A DnaB intein in *Rhodothermus marinus*: indication of recent intein homing across remotely related organisms. Proc. Natl. Acad. Sci. USA **94:** 7851–7856
90 Jin Y., Binkowski G., Simon L. D. and Norris D. (1997) Ho endonuclease cleaves MAT DNA in vitro by an inefficient stoichiometric reaction mechanism. J. Biol. Chem. **272:** 7352–7359
91 Nickoloff J. A., Singer J. D. and Heffron F. (1990) In vivo analysis of the *Saccharomyces cerevisiae* HO nuclease recognition site by site-directed mutagenesis. Mol. Cell. Biol. **10:** 1174–1179
92 Durrenberger F. and Rochaix J.-D. (1991) Chloroplast ribosomal intron of *Chlamydomonas reinhardtii*: in vitro self-splicing DNA endonuclease activity and in vivo mobility. EMBO J. **10:** 3495–3501
93 Herrin D. L., Chen Y.-F. and Schmidt G. W. (1990) RNA splicing in *Chlamydomonas* cholorplasts: self-splicing of 23S pre RNA. J. Biol. Chem. **265:** 21134–21140
94 Rochaix J.-D., Rahire M. and Michel F. (1985) The choloroplast ribosomal intron of *Chlamydomonas reinhardtii* chloroplast codes for a polypeptide related to mitochondrial maturases. Nucleic Acids Res. **13:** 975–984
95 Gimble F. S. and Thorner J. (1992) Homing of a DNA endonuclease by meitoic gene conversion in *Saccharomyces cerevisiae*. Nature **357:** 301–306
96 Gimble F. S. and Stephens B. W. (1995) Substitutions in conserved dodecapeptide motifs that uncouple the DNA binding and DNA cleavage activities of PI-*Sce*I endonuclease. J. Biol. Chem. **270:** 5849–5856
97 Dalgaard J. Z., Garrett R. A. and Belfort M. (1993) A site-specific endonuclease encoded by a typical archaeal intron. Proc. Natl. Acad. Sci. USA **90:** 5414–5417
98 Dalgaard J. Z., Garrett R. A. and Belfort M. (1994) Purification and characterization of two forms of I-*Dmo*I, a thermophilic site-specific endonuclease encoded by an archaeal intron. J. Biol. Chem. **269:** 28885–28892
99 Mueller J. E., Smith D., Byrk M. and Belfort M. (1995) Intron-encoded endonuclease I-*Tev*I binds as a monomer to effect sequential cleavage via conformational changes in the *td* homing site. EMBO J. **14:** 5724–5735
100 Kowalski J. C., Belfort M., Stapleton M. A., Holpert M., Dansereau J. T., Pietrokovski S. et al. (1999) Configuration of the catalytic GYI-YIG domain of intron endonuclease I-*Tev*I: coincidence of computational and molecular findings. Nucleic Acids Res. (in press)
101 Phillips S. E. V. (1994) The β-ribbon DNA recognition motif. Ann. Rev. Biophys. Biomol. Struct. **23:** 671–701
102 Turmel M., Otis C., Cote V. and Lemieux C. (1997) Evolutionarily conserved and functionally important residues in the I-*Ceu*I homing endonuclease. Nucleic Acids Res. **25:** 2610–2619
103 Friedhoff P., Kolmes B., Gimadutdinow O., Wende W., Krause K. L. and Pingoud A. (1996) Analysis of the mechanism of the *Serratia* nuclease using site-directed mutagenesis. Nucleic Acids Res. **24:** 2632–2639
104 Miller M. D., Tanner J., Alpaugh M., Benedik M. J. and Krause K. L. (1994) 2.1 A structure of *Serratia* endonuclease suggests a mechanism for binding to double-stranded DNA. Nature Struct. Biol. **1:** 461–468
105 Friedhoff P., Franke I., Mesiss G., Wende W., Krause K. L. and Pingoud A. (1999) A similar active site for non-specific and specific endonucleases. Nature Struct. Biol. **6:** 112
106 Friedhoff P., Franke I., Krause K. L. and Pingoud A. (1999) Cleavage experiments with deoxythymidine 3′,5′-bis-(*p*-nitrophenyl phosphate) suggest that the homing endonuclease I-*Ppo*I follows the same mechanism of phosphodiester bond hydrolysis as the non-specific *Serratia* nuclease. FEBS Lett. **443:** 209–214
107 Derbyshire V., Grindley N. D. F. and Joyce C. M. (1991) The 3′-5′ exonuclease of DNA polymerase I of *Escherichia coli*: contribution of each amino acid at the active site of the reaction. EMBO J. **10:** 17–24
108 Beintema J. J. and Kleineidam R. G. (1998) The ribonuclease A superfamily: general discussion. Cell. Mol. Life Sci. **54:** 825–832
109 Pingoud A. and Jeltsch A. (1997) Recognition and cleavage of DNA by type-II restriction endonucleases. Eur. J. Biochem. **246:** 1–22
110 Aggarwal A. K. (1995) Structure and function of restriction endonucleases. Curr. Opin. Struct. Biol. **5:** 11–19