Critical Reviews in Biochemistry and Molecular Biology, 38(3):199–248 (2003)

Structural and Functional Characteristics of Homing Endonucleases

*N. Guhan and K. Muniyappa**

Department of Biochemistry, Indian Institute of Science, Bangalore 560012, India

* **Corresponding author:** Dr. K. Muniyappa, Department of Biochemistry, Indian Institute of Science, Bangalore 560012, India. Tel: (91-80) 293.2235/360 0278. Fax: (91-80) 360.0814/0683. Email: kmbc@biochem.iisc.ernet.in

ABSTRACT: Mobile genetic elements constitute a remarkably diverse group of nonessential selfish genes that provide no apparent function to the host. These selfish genes have been implicated in host extinction, speciation and architecture of genetic systems. Homing endonucleases, encoded by the open reading frames embedded in introns or inteins of mobile genetic elements, possess double-stranded DNA-specific endonuclease activity. They inflict sequence-specific double-strand breaks at or near the homing site in intron- or intein-less allele. Subsequently, through nonreciprocal exchange the insertion sequence (intron or intein) is transferred from an intein- or intron-containing allele to an intein- or intron-less allele. The components of host double-strand break repair pathway are thought to finish the "homing" process. Several lines of evidence suggest that homing endonucleases are capable of promoting transposition into ectopic sites within or across genomes for their survival as well as dispersal in natural populations. The occurrence of inteins at high frequencies serves as instructive models for understanding the mechanistic aspects of the process of homing and its evolution. This review focuses on genetic, biochemical, structural, and phylogenetic aspects of homing endonucleases, and their comparison with restriction endonucleases.

KEY WORDS: homing endonucleases, inteins, introns, mobile genetic elements, transposition, site-specific endonucleases.

I. INTRODUCTION

More than 50 years ago, through a combination of masterful genetic analysis and unique insights, B. McClintock discovered the biologically important phenomenon of transposition. This discovery originated from her studies on broken chromosomes in maize. She observed a pattern of regular breakage on chromosome 9 during development, and she named the breakage site the *Ds* (dissociator) locus. Further analysis of *Ds* locus led to the discovery of a second locus that was required to activate the process of transposition, and she designated this as the *Ac* (activator) locus. As a result of *Ds*/*Ac* system, the progeny plants with broken chromosomes displayed an unusually high frequency of variegation, including sectored coloring of the kernels. After intense genetic analysis, McClintock concluded that the pattern of change resulted from movement of *Ds* to a new location thereby causing genome instability. These experiments were followed by intense research, which led to the identification of mobile genetic elements in organisms, including eubacteria and phage T4. Several lines of evidence suggest that such elements are ubiquitous and are known to alter the integrity of the genome in a number of different ways. For example, the draft human genome sequence (http://www.ncbi.nlm.nih.gov/genome/guide/human) revealed that the genome is populated with vast number (43%) of mobile elements than any other organism whose genome has been analyzed. These insertion elements have been implicated in shaping the human genome. Their insertion has also resulted in human diseases such as muscular dystrophy, hemophilia, and breast cancer (Kobayashi *et al*., 1998; Kazazian *et al*., 1988).

Mobile genetic elements constitute a remarkably diverse group of nonessential selfish genes, which provide no apparent function to the host. The enzyme encoded by the mobile genetic element promotes its self-propagation. Some exist at multiple locations,

1040-9238/03/$.50

whereas others are present at unique sites, in eukarya both in nuclear as well as organelle genomes. Based on a wealth of correlative studies published over the last 3 decades, mobile genetic elements can appropriately be subdivided into four categories: transposons, homing endonucleases, segregation distorters, and heritable microorganisms (Hurst and Werren, 2001). This review contains an overview of the genetic studies, structural organization, enzymology, and phylogenetic analyses of homing endonucleases (HEases, according to Roberts *et al*., 2003). To help clarify the functional differences, HEases are broadly classified into two categories: (a) intron- and (b) intein-encoded endonucleases (ENases, according to Roberts *et al*., 2003). This division is intended to guide the discussion on their structural organization, mechanistic aspects, as well as their regulation.

A. Discovery of Homing

Historically, insights into the phenomenon of "homing" emerged from studies on mitochondrial genetic marker, termed 'ω' of *Saccharomyces cerevisiae*. It was observed that mixing of ω^+ and ω^- strains of *S. cerevisiae* led to the conversion of ω^- to ω^+ in a nonreciprocal manner (Coen *et al*., 1971; Bolotin *et al*., 1971), and co-conversion of flanking sequences (Dujon *et al*., 1976; Jacquier and Dujon, 1985; Macreadie *et al*., 1985). Further analysis led to mapping of the ω locus to the group I intron of mitochondrial 21S rRNA gene (Dujon *et al*., 1976; Bos *et al*., 1978; Heyting and Menke, 1979). Subsequently, it was discovered that expression of an open reading frame (*fit*1) from the intron as well as introduction of a double-strand break within the ω^- allele were required for conversion (Jacquier and Dujon, 1985; Macreadie *et al*., 1985; Zinn and Butow, 1985). An enzymatic activity encoded by *S. cerevisiae* group I intron, termed I-*Sce*I, was required for the cleavage of ω^- allele, and transfer of endonuclease-encoding intron (Colleaux *et al*., 1986, 1988).

For several years following its discovery, the mobility of ω element was thought of as a locus-specific phenomenon. The occurrence of group I introns in diverse organisms such as *S. cerevisiae* mitochondrial *COX1* (aI4α intron or I-*Sce*II, aI3α or I-*Sce*III) (Delahodde *et al*., 1989; Wenzlau *et al*., 1989; Sargueil *et al*., 1991), *Physarum polycephalum* nuclear LSU rRNA gene (I-*Ppo*I) (Muscarella and Vogt, 1989), T4 bacteriophage *td* (I-*Tev*I) and *sunY* (I-*Tev*III) genes (Quirk *et al*., 1989), 23S rRNA (I-*Dmo*I) gene of *Desulfurococcus mobilis* (Kjems and Garrett, 1988; Aagaard *et al*., 1997), and in DNA polymerase gene of *Bacillus subtilis* phages SP01 and SP82 (I-*Hmu*I, I-*Hmu*II) (Goodrich-Blair *et al*., 1990) indicated ubiquitous distribution of HEases. Homing of non-self DNA at an extragenic locus containing VDE (*VMA1* Derived Endonuclease) recognition sequence further strengthened the notion that homing was not locus-specific (Nogami *et al*., 2002).

B. Distinction between Homing Endonucleases and Type II Restriction Endonucleases

HEases share functional and mechanistic similarities with type II restriction endonucleases (REases, according to Roberts *et al*., 2003). Typically, both inflict double-strand breaks in their target DNA substrates. However, they differ in their structure, sequence recognition, and genomic location. HEases are ubiquitous in all three biological kingdoms, whereas REases exist only in prokaryotes. At the sequence level, HEases are identified by the presence of four conserved motifs: LAGLIDADG, GIY-YIG, H-N-H, and His-Cys box. These motifs participate in the coordination of metal ions and hydrolysis of phosphodiester bonds. HEases are notable for their long target sites and a tolerance for sequence polymorphisms in their DNA substrates. This is in contrast to REases, which normally require short recognition sequences. Most REases contain the PDX$_{8-25}$ (E/D)XK motif in the catalytic center and share an important core containing five-stranded mixed β-sheet flanked by α-helices (Kovall and Matthews, 1998, 1999). Nevertheless, such examples are few and should not be considered a general feature of the class of REases. An excellent comprehensive review on type II REases can be found elsewhere (Pingoud and Jeltsch, 2001).

The criterion adapted for naming HEases is analogous to that of REases (Roberts *et al*., 2003). HEases are designated by a three-letter genus-species acronym in which the first letter (in uppercase) is the first letter of the genus and the next two (in lower case) are the first two letters of the species (Belfort and Roberts, 1997). A Roman numerical suffix is used to distinguish multiple enzymes from the same organism in the chronological order of their discovery. However, it was recognized that the genus-species designation is necessary but not sufficient to define a HEase. HEases are also characterized by prefix **F**-, **I**-, or **PI**- for enzymes encoded by free-standing ORFs, introns, and inteins, respectively. For example, the founding members of the intron-, intein-, and free-standing gene encoded HEases are I-*Sce*I, PI-*Sce*I, and F-*Sce*II (HO endonuclease), respectively, from *Saccharomyces cerevisiae*.

C. Occurrence of ORFs Encoding HEases

HEases are encoded by ORFs of group I, group II, group III introns, in-frame protein sequences (inteins) as well as free-standing genes. These have been categorized into two broad classes based on whether self-splicing occurs at the RNA or protein level, and their genomic distribution. While introns belong to the first category, inteins represent the second. Introns have been subdivided further into group I and group II introns based on their RNA secondary structure and splicing mechanisms. Group I introns catalyze their own splicing in a protein-independent manner, whereas group II introns are site-specific retroelements that depend on protein machinery for splicing. However, identification of introns with characteristics different from the above two groups resulted in further classification. For example, the short group II-like introns in the mRNA of plastids in euglenoid protists, designated as group III introns differ from conventional group II introns by lacking secondary structure corresponding to domains II – VI and exhibit relaxed splice-site consensus. Furthermore, results from archaea suggest that they possess a distinct class of introns in their tRNA and rRNA genes, which undergo splicing by an archaeal-specific mechanism. Although introns are not the main focus of this review, the section below highlights basic information regarding structural and mechanistic aspects of introns, as they have contributed immensely to our understanding of intein-encoded HEases. Comprehensive reviews on group I and group II introns have appeared elsewhere (Saldanha *et al*., 1993; Lambowitz and Belfort, 1993; Michel and Ferat, 1995; Bonnen and Vogel, 2001).

II. INTRONS

A. Group I Introns

The founding member of group I intron is the one in the thymidylate synthase (*td*) gene of bacteriophage T4 (Chu *et al*., 1984). Additional examples of group I introns emerged from phages: *nrdB* and *nrdD* of phage T4 (Gott *et al*., 1986; Young *et al*., 1994), phage SP01 DNA polymerase gene (Goodrich-Blair *et al*., 1990) and in the large rRNA gene of *S. cerevisiae* (Jacquier and Dujon, 1985). Since then group I introns have been identified in tRNA, rRNA, and mRNA genes of fungi, protists, plant organelle genomes, as well as in the mRNA and tRNA genes of eubacteria and phages. Recently, group I intron has been identified in the eubacterial protein-coding gene, *recA* of *Bacillus anthracis* (Ko *et al*., 2002).

Group I introns have wide phylogenetic distribution from phages to plants, but frequent in mitochondrial genomes and nuclear tRNA genes of fungi. Also, they are present in large rRNA genes and genes that encode components of the electron transport system. Of over 80 bacterial group I introns identified so far, only in one case (*Simkania negevensis*) was it found in rDNA (Everett *et al*., 1999). Group I introns range in size from 200 to 3000 nucleotide residues. Approximately 30% of group I introns encode a protein that displays endonuclease activity (reviewed by Lambowitz, 1989; Dujon, 1989; Belfort, 1990; Gorbalenya, 1994; Belfort *et al*., 1995; Edgell *et al*., 2000; Bonen and Vogel, 2001). The folding of RNA into characteristic secondary structure results in the generation of the active center for splicing. In principle, group I introns are ribozymes containing P1– P9 motifs that bring the 5' and 3' splice sites and the reactive guanosine cofactor into close proximity. The mechanism of splicing of group I introns involves a series of transesterification reactions with external guanosine cofactor acting as a nucleophile in Mg^{2+}-dependent manner. In some cases, the linear intron is further converted into a circular structure by an additional transesterification reaction (reviewed by Cech, 1990). The transfer of group I intron from intron containing allele to intron-less allele is promoted by the components of double-strand break repair pathway (reviewed by Dujon, 1989; Belfort, 1990).

B. Group II Introns

Group II introns are ubiquitous in the genomes of fungal and plant mitochondria and chloroplasts (Michel and Ferat, 1995). The representative organisms include Cyanobacteria (*Calothrix*, *Anabena* and *Nostoc*), proteobacteria (*E. coli*, *Azotobacter vinelandii*), *Lactococcus lactis* (Mills *et al*., 1996, 1997; Shearman *et al*., 1996), *Clostridium difficile*

(Mullany *et al*., 1996), *Pseudomonas alcaligenes* (Yeo *et al*., 1997), *Serratia marcescens* (Kulaeva *et al*., 1998), *Streptococcus pneumoniae* (Coffey *et al*., 1998), *Sinorhizobium meliloti* (Martinez-Abarca *et al*., 1998), and *Sphingomonas aromaticivorans* (Romine *et al*., 1999). Further evidence for their abundance has emerged from genome projects and genome surveys. The mitochondrial genomes of *A. thaliana* and *Marchantia polymorpha* contain 23 and 25 group II introns, respectively (Gray *et al*., 1998). A classic example for the abundance of introns is the chloroplast genome of *Euglena*: it harbors 80 normal group II introns and 155 degenerate group III introns. Together these two groups correspond to ~40% of the *Euglena* chloroplast genome (Hallick *et al*., 1993). Based on distinct structural characteristics, group II introns are further subdivided into group IIA and IIB (Michel *et al*., 1989), which differ in: (a) 3' splice-site consensus sequence, (b) distance between bulged ‘A’ and 3' splice-site, (c) and tertiary interactions. Also, some group II introns encode polypeptides that are essential for splicing *in vivo*. Sequence analysis of ORFs in subgroups IIA and IIB indicate considerable homology to reverse transcriptases (Michel and Lang, 1985; Xiong and Eickbush, 1990).

Group II introns constitute a large class of catalytic RNAs varying in length from 0.6 to 2 kb. These assume distinct three-dimensional structures, designated domains I–VI. The splicing catalytic center containing the consensus splice sites (5' -GUGYG…AY-3') is assembled from domains I and V. The splicing pathway proceeds via a two-step transesterification reaction with adenosine in bulged domain VI at the 3' end of the intron acting as a nucleophile. Splicing generates a lariat structure in which the 5' end of the intron is linked by a 2'–5' phosphodiester bond to adenine residue at the 3' end of the intron. Although some group II introns can self-splice, they do so under nonphysiological conditions. The splicing mechanism is analogous to the splicing of eukaryotic nuclear mRNAs (Sharp, 1985; Cech, 1986; Hickey and Benkel, 1986). Some group II introns require factors encoded by either host or intron for efficient splicing *in vivo*. These factors are believed to facilitate proper folding of RNA substrate into a catalytically favorable configuration. Sequencing projects have identified subtle structural diversity among group II introns, with some intermediate between group IIA and IIB introns (Toor *et al*., 2001).

C. Mobility of Group I and Group II Introns

Homing of group I intron is initiated by DNA endonuclease (ENase) encoded by the intron itself. The homing sites are long asymmetric sequences, displaying different binding specificities and cleavage frequencies. Most HEases, if not all, inflict a double-strand break at or near the insertion site to generate a staggered break with 4-nucleotide 3' overhangs. The cellular components of the double-strand break repair pathway are crucial for the repair of DSBs inflicted by HEases (Figure 1). The mechanism appears to be similar to the DSBR pathway originally proposed for gene conversion in *S. cerevisiae* (Szostak *et al*., 1983). An analysis of phage T4 *td* intron in crossover resolvase mutants suggested that homing mediated by group I intron can also follow alternate gene conversion pathways such as Synthesis-Dependent Strand Annealing (Mueller *et al*., 1996; reviewed by Dujon, 1989; Edgell *et al*., 2000).

Group II introns use a variety of mechanisms to “home” into cognate or ectopic DNA sites. Studies in *S. cerevisiae* and *L. lactis* have suggested that homing requires both intron-encoded proteins as well as a RNA component (Zimmerly *et al*., 1995a). The observation that mobility was abolished by mutations that either block excision of group II introns, on the one hand, or interactions between catalytic intron RNA with intron-encoded proteins, on the other hand, indicated that homing involves RNA-guided insertion of introns. To distinguish it from the ENase promoted homing of group I introns, the process has been termed “retrohoming” (Curcio and Belfort, 1996), analogous to retrotransposon integration.

The basic features of group II homing has been elucidated using the yeast mitochondrial introns aI1 and aI2 (Zimmerly *et al*., 1995b; Moran *et al*., 1995; Yang *et al*., 1996; 1998), *L. lactis LtrB* intron and *S. meliloti RMInt1* intron (Mills *et al*., 1997; Martinez-Abarca *et al*., 2000). The mechanism of retrohoming (Figure 1, center panel) involves target-primed reverse transcription carried out by the ribonucleoprotein complex comprising intron-encoded RT and excised intron RNA lariat (Zimmerly *et al*., 1995b). This complex binds the homing site DNA in the intron-less allele and the intron RNA reverse splices into the sense strand. Indeed, it has been shown that the efficiency of group II intron homing and reverse transcription is brought about by strong interaction

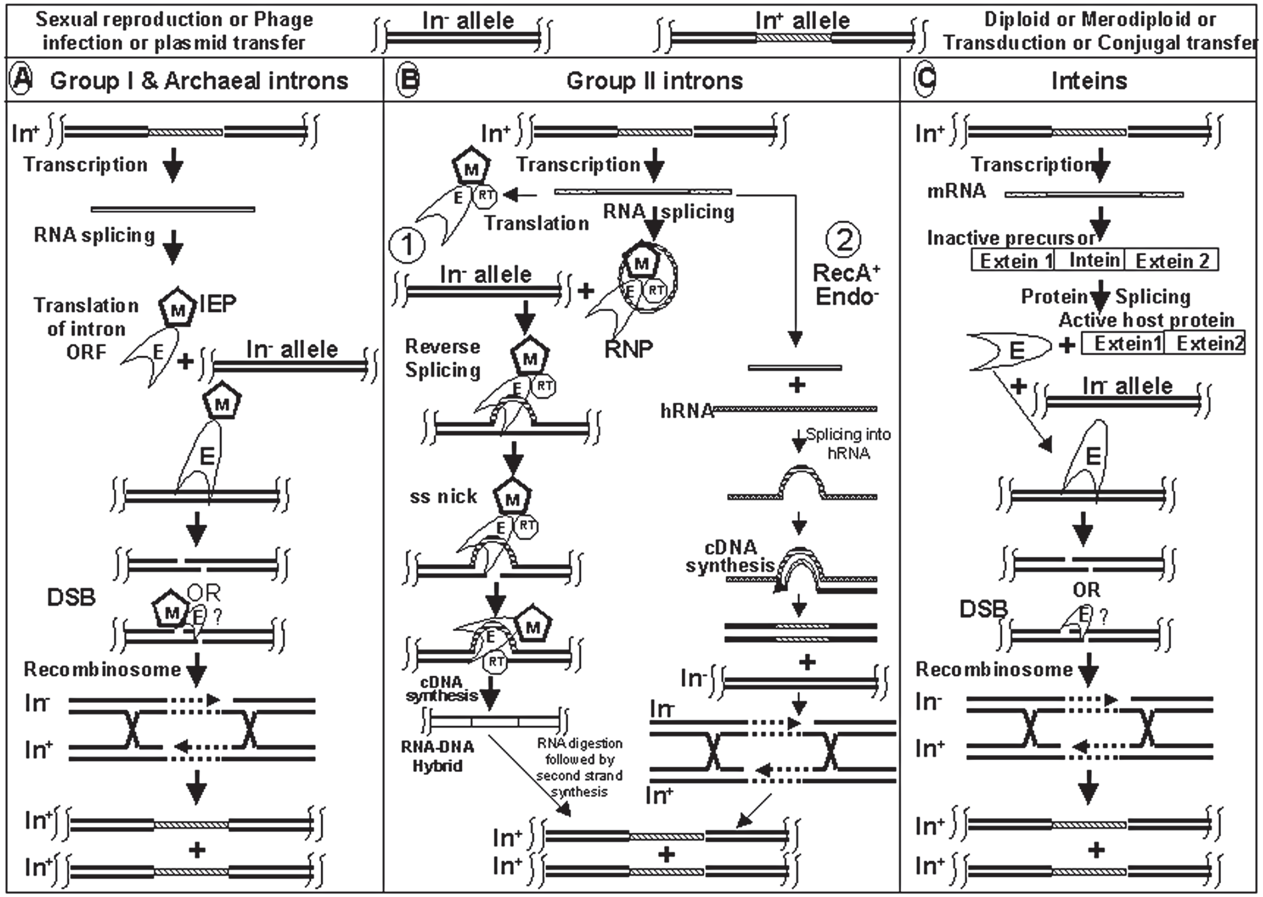

FIGURE 1. Schematic representation of homing mediated by group I, group II, archaeal introns and inteins. Panel A. Mechanism of homing mediated by group I and archaeal introns. Intron containing allele (In^+) specifies the precursor RNA, which undergoes splicing to generate mature RNA and the intron. The intron-encoded protein (IEP) acts as a maturase (M) and further facilitates its own splicing and also functions as an ENase (E). The ENase inflicts a double-strand break (DSB) at or near the intron insertion site in intron-less allele (In^-). Subsequently, the double-strand break is repaired using In^+ allele as the template by the host recombination machinery ($In^- \rightarrow In^+$). **Panel B.** Homing of group II introns. (1) The intron encodes a multifunctional protein: maturase that facilitates its own RNA splicing, an ENase and reverse transcriptase (RT). In RecA-independent, but ENase-dependent pathway the ribonucleoprotein (RNP) complex comprising intron lariat and IEP (M, E, RT) binds the target sequence embedded in the In^- allele. Following binding, the RNA reverse splices into the sense-strand of the DNA. The ENase nicks the anti-sense strand and, subsequently, the RT synthesizes the first cDNA strand using the 3' hydroxyl end as primer and the RNA as template. Following RNA digestion, the second strand synthesis results in the conversion of In^- to In^+ allele. (2) In RecA-dependent, but ENase-independent or retrotransposition pathway, the intron RNA reverse splices into a heterologous RNA (hRNA), followed by cDNA synthesis by RT and recombination to culminate in transposition. **Panel C.** Homing process mediated by inteins. The precursor protein undergoes splicing to generate ENase and the mature host protein. The ENase recognizes the target sequence in intein-less allele (In^-) and inflicts a DSB, which is then repaired by the host recombination machinery resulting in homing. In some cases the ENase has been found to be associated with the DNA ends postcleavage and the '?' mark indicates whether it enhances the assembly of the ensuing recombination process. (Modified after Chevalier and Stoddard, 2001b.)

between intron RNA and DNA target sites (IBS1-EBS1, IBS2-EBS2, and δ-δ'). For example, deletion of domain V of intron RNA reduces the specificity of interaction between RNA and target DNA (Morozova *et al*., 2002). In addition to sequence-specific interaction provided by its RNA component, intron-encoded protein also provides specificity by binding and unwinding the target duplex to abet base pairings between RNA and the target DNA during reverse splicing reaction. Also, interactions between intron-encoded protein and the downstream DNA target site have been shown to be important for cleavage of antisense strand. The zinc-domain of RT cleaves the antisense strand at a specific site and the RT reverse transcribes the inserted intron using the cleaved DNA as the primer. The entire process requires catalytic activities of both the RT and intron RNA at each step of the reaction.

Some group II introns are also capable of transposition into ectopic sites at low frequencies (Mueller *et al*., 1993; Sellem *et al*., 1993; Cousineau *et al*., 2000; Dickson *et al*., 2001; Morozova *et al*., 2002). The mechanism of insertion of group II introns at ectopic sites is similar to that at cognate sites. Those introns that are devoid of the ENase domain are dependent on *recA* for their homing. Although transposition of *S. meliloti* intron has been shown to be *recA*-independent (Martinez-Abarca *et al*., 2000), *L. lactis* intron is *recA*-dependent (Cousineau *et al*., 2000). Interestingly, ectopic transposition of *S. meliloti RmInt*1 has been shown to occur in natural field populations (Munoz *et al*., 2001).

D. Archaeal Introns

Archaeal introns exist in tRNA and rRNA genes and are spliced by an archaeal-specific mechanism (reviewed by Lykke-Andersen *et al*., 1997a). The archaeal introns are unrelated to the eukaryotic or bacterial introns, but share significant structural and functional similarities. These are relatively small in size, varying from 15 to 110 nucleotide residues. However, some of these harbor ORFs comprising of about 600 nucleotide residues (Garrett *et al*., 1991; Dalgaard and Garrett, 1992; Burggraf *et al*., 1993) and encode proteins that contain the LAGLIDADG motif, which is normally found in the enzymes encoded by group I introns and inteins. Archaeal introns are catalytically inert in self-splicing, and this seems to be the rule rather than exception. The mechanism of archaeal intron excision involves the formation of a bulge-helix-bulge motif at the intron-exon junction. This motif is recognized by an endoribonuclease and cuts at symmetrical positions within the three-nucleotide bulges (Thompson and Daniels, 1988, 1990; Kjems and Garrett, 1988). Once the cleavage is complete, the ends of the exons are ligated to yield 3' to 5' phosphodiester bonds resulting in circularization of the intron.

III. INTEINS

Sequence comparison of *Neurospora crassa* vacuolar membrane ATPase with its Ca^{2+}-dependent ATPase (Shih *et al*., 1988) disclosed the presence of an intervening sequence. Consistent with this, two groups independently reported the existence of intervening sequence in the 69-kDa subunit of the vacuolar membrane H^+-ATPase encoded by *VMA1* of *S. cerevisiae* (Hirata *et al*., 1990; Kane *et al*., 1990). The sequence at the amino- and carboxyl terminal portions of vacuolar H^+-ATPase of *S. cerevisiae* was found to be highly homologous to that of carrot and *N. crassa* ORFs except for a central nonhomologous region comprised of 454 amino acids. Genetic and biochemical evidence suggested that the *VMA1* gene product (69 kDa) is generated by posttranslational excision of the nonhomologous region followed by the ligation of amino- and carboxyl terminal domains. Further results indicated that the two proteins (H^+-ATPase and intein) were encoded by *VMA1* gene in *S. cerevisiae.*

Most inteins contain two distinct domains each harboring protein splicing or endonuclease activities. Protein splicing is an autocatalytic process, which results in self-excision of **in**ternal pro**tein** (INTEIN) fragment from a protein precursor followed by ligation of **ex**ternal pro**tein** (EXTEIN) fragments. The ligation of exteins resulting in a normal peptide bond (Cooper *et al*., 1993) differentiates protein splicing from other auto-proteolytic processes (Paulus, 2000). Most importantly, the information necessary for protein splicing reaction resides within the intein sequence itself and the first residue of the C-extein (Xu *et al*., 1993). Many excellent reviews have covered the organization of inteins, mechanism of protein splicing and its applications (Perler *et al*., 1994; Cooper and Stevens, 1995; Shao and Kent, 1997; Anraku, 1997; Perler *et al*., 1997b; Gimble, 1998; Paulus, 2000; 2001; Perler and Adam, 2000; Liu, 2000; Giriat, 2001; Evans and Xu, 2002).

A. Distribution of Inteins

Over 130 inteins have been identified to date from the entire taxonomy (eukaryotes, bacteria, archaea, viruses, and bacteriophages) and are embedded in 34 different types of host proteins (proteins involved in DNA and RNA metabolism, proteases, vacuolar-type ATPases, etc.) (Perler, 2002; Pietrokovski, 2001). These are mostly inserted into highly conserved motifs of the host protein with no apparent size limitation. For example, although *Mja* RFC1-3 is four times the size of host protein, *Fne* pRP8 is ten times smaller (Gogarten *et al*., 2002). Although inteins have been identified in several organisms, they are undetectable in 32 of 85 different genomes that have been sequenced so far, including *Escherichia coli*, *Arabidopsis thaliana*, *Drosophila melanogaster,* and *Homo sapiens***.** Among the eubacterial inteins that are known to date, 22 of 32 inteins exist in mycobacteria. While most mycobacteria harbor one or two inteins, *Mycobacterium leprae* harbors four in *gyrA*, *recA*, *dnaB,* and *pps1* genes (Cole *et al*., 2001; Davis *et al*., 1991, 1992) and *Mycobacterium tuberculosis* contains three in *recA*, *dnaB,* and *pps1* genes (Cole *et al*., 1998). Intein-coding sequences disrupting *recA* and *pps*1 from *M. tuberculosis* have been found to be specific for *M. tuberculosis* complex (Saves *et al*., 2002a). Intein sequences in RecAs have been identified in both pathogenic (*M. tuberculosis* and *M. leprae*) (Davis *et al*., 1994) and nonpathogenic species (*Mycobacterium chitae*, *Mycobacterium fallax*, *Mycobacterium flavescens*, *Mycobacterium gastri*, *Mycobacterium shimodei,* and *Mycobacterium thermoresistible*) (Saves *et al*., 2000b), and in slow- and fast-growing species of mycobacteria (Blackwood *et al*., 2000). Depending on the location of inteins in RecA protein, they are classified into two groups: (1) RecA-b site, where intein is located downstream of Gly^{205} (as in *M. leprae*) and whose equivalent in *E. coli* (Gly^{204}) corresponds to the DNA-binding Loop 2, and (2) RecA-a site, where intein is located downstream of Lys^{251} whose equivalent in *E. coli* is a region for which no direct function is implicated. Biochemical characterization of the endonuclease activity of Pps1 intein from *M. gastri* (PI-*Mga*I) and *M. tuberculosis* has been reported (Saves *et al*., 2002a, 2001b). The occurrence of inteins at such high frequencies in mycobacteria posits a probable physiological role. A catalog of inteins in all the three biological kingdoms and their insertion sites is given in Table 1 and also available at: **http://www.neb.com/inteins.html or http://bioinfo.weizmann.ac.il/~pietro/inteins**

B. Structural Organization of Inteins

Inteins range in size from 134 (*Mth* RIR1) to 608 (*Pab* RFC-2) amino acids and are organized into 10 conserved motifs designated as A to H, N2 and N4 (Pietrokovski, 1994, 1998a; Perler *et al*., 1997b; Perler, 2002). While the amino- (blocks A, N2, B, and N4) and carboxyl terminal (blocks F and G) domains are responsible for protein splicing, the central domain (blocks C, D, E and H) harbors the ENase activity (Figure 2). However, in PI-*Sce*I part of the DNA-binding domain is embedded within the amino-terminal region involved in splicing of N-terminal extein (Grindl *et al*., 1998; Hall *et al*., 1997; Moure *et al*., 2002; Werner *et al*., 2002). Comparative sequence analysis suggested that such an organization is very rare among inteins (Dalgaard *et al*., 1997a, 1997b; Pietrokovski 1998a). The consensus sequence in the individual blocks of intein and their involvement in two different activities are listed in Table 2.

Inteins with unique structural and biochemical properties have been identified in different organisms. One group designated as “mini-inteins” contain 130 to 200 amino acid long polypeptides lacking the ENase domain (*Porphyra purpurea dnaB* — Pietrokovski, 1996; *Mycobacterium xenopi gyrA* — Telenti *et al*., 1997). In *M. xenopi gyrA* intein, the central ENase domain was replaced by a small (24 amino acids) nonhomologous spacer, followed by nine amino acid GyrA intein sequence (Telenti *et al*., 1997; Klabunde *et al*., 1998). Because mini-inteins lack ENase domains, the mechanism of their dispersal remains a fascinating area for further investigation. Additionally, noncanonical inteins were also identified by Hidden-Markov analysis in the genomes of few eubacteria (Gorbalenya *et al*., 1998). One such intein was found in the genome of *Methanococcus jannaschii* where ‘Ala’ is substituted for Ser/Cys at the N-terminus of intein (Bult *et al*., 1996), and three in *Synechocystis* sp. PCC6803 (Pietrokovski, 1996; Chen *et al*., 2002a; 2002b; Wu *et al*., 1998a, 1998b; Evans *et al*., 2000; Martin *et al*., 2001; Perler, 2002; reviewed by Gogarten *et al*., 2002).

IV. HOMING ENDONUCLEASES

HEases are believed to play key roles in genome rearrangements and shuffling of protein domains during evolution (Dalgaard *et al*., 1997a). For example, HO endonuclease (F-*Sce*II) of *S. cerevisiae*, a free-standing HEase, initiates mating-type switching by inflicting a double-strand break at the *MAT* locus,

TABLE 1
Inteins from Eukarya, Eubacteria, and Archaea
Intein from different organisms are listed here with their size, location in the host gene, and amino- and carboxyl-terminal splice junction amino acid residues. Inteins demonstrated as endonucleases by experimentation are designated according to the current nomenclature. DOD, dodecapeptide motif; HNH, His-Asn-His motif; ND, not detectable; ?, uncertain.

Eukarya

Intein	Size	Endo activity	Endo motif	Location in Extein	N-terminal Splice junction	C-terminal splice junction	Insertion site
Ceu ClpP	456	ND	DOD	E447	GCHVMIHQPE/C	GN/SSIQCQASDI	Endopeptidase active site, clpP-a
ClV RIR1	339	ND	DOD	L271	KGLTIKQSNL/C	GQ/CSEIILPTDS	*RIR1*-b
Ctr VMA	471	ND	DOD	G283	SNSDVIIYVG/C	HN/CGERGNEMAE	*VMA*-a
Fne PRP8 (*Cne*PRP8)	172	ND	None	A?	TWEGLFWEKA/C	HN/SGFEESMKNK	*PRP8*-a
Gth DnaB	160	ND	None	G376	PILSDLKESG/C	HN/SIEQDADVVL	*dnaB*-a
Ppu DnaB	150	ND	None	G361	PLLSDLRESG/C	HN/SIEQDADLVI	*dnaB*-a
Sce VMA	454	PI-*Sce*I	DOD	G283	SNSDAIIYVG/C	HN/CGERGNEMAE	*VMA*-a

Table 1. (contd.)

Eubacteria

Intein	Size	Endo activity	Endo motif	Location in Extein	N-terminal Splice junction	C-terminal splice junction	Insertion site
Aae RIR2	346	ND	DOD	L229	QIKYINRDEL/C	GN/CHVTLFRNII	*RIR2*-a, Metal binding site
*APSE*1 dpol	306	ND	None	S608	ERLKTYGGKS/C	HN/CENICQAAAR	*dpol*-a
Bsu RIR1	385	ND	DOD	L?	HISKVKFSNL/C	GN/CSEVLQSSQV	*RIR1*-b
Cau Hyp	276	ND	DOD	A278	MTKILNEGWA/C	HN/SYWHSTIMTQ	*Cau* hyp-a
Cth Hyp	333	ND	DOD	K84	YVEIPKKNGK/Q	HN/SELAAAVALY	*Cth* hyp-a
Dha RIR1		ND	DOD	P272	QLGEIEATNP/C	HN/CGEQPLLPNE	*RIR1*-b
Dra-RF18410 Snf2	343	ND	DOD	K693	ILADDMGLGK/A	HN/TLQT	*snf2*-a, ATP/GTP-binding site motif A (P-loop)
Dra-RF78101 RIR1	366	ND	DOD	P464	ERYEIRSTNP/C	HN/CGEIPLTVGE	*RIR1*-b
Dra-RF78101 Snf2	343	ND	DOD	K693	ILADDMGLGK/A	HN/TLQTLAHLLK	*snf2*-a, ATP/GTP-binding site motif A (P-loop)
Mav DnaB	337	ND	DOD	K232	IVAARPGVGK/A	HN/STLGLDFLRS	*dnaB*-b, ATP/GTP-binding site motif A (P-loop)
Mch RecA	364	ND	DOD	G?	REKIGVMFG/C	HN/SPETTTGGKA	*recA*-b
Mfa RecA	363	ND	DOD	G?	REKIGVMFG/C	HN/SPETTTGGKA	*recA*-b
Mfl GyrA	421	ND	DOD	Y?	GNDPPAAMRY/C	HN/TEARLTPLAM	*gyrA*-a, After active site Tyr
Mfl RecA	364	ND	DOD	G205	LREKIGVMFG/C	HN/SPETTTGGKA	*recA*-b
Mfl-ATCC14474 RecA	364	ND	DOD	G?	REKIGVMFG/C	HN/SPETTTGGKA	*recA*-b
Mga GyrA	420	ND	DOD	Y?	GNDPPAAMRY/C	HN/TEARLTPLAM	*gyrA*-a, After active site Tyr
Mga Pps1	378	PI-*Mga*I	DOD	E?	VAAQYE/C	NN/SEVVY	*pps1*-c
Mga RecA	368	ND	DOD	G?	RDKIGVMFG/C	HN/SPETTTGGKA	*recA*-b
Mgo GyrA	420	ND	DOD	Y?	GNDPPAAMRY/C	HN/TEARLTPLAM	*gyrA*-a, After active site Tyr
Min DnaB	335	ND	DOD	K233	IVAARPGVGK/A	HN/STLGLDFLRS	*dnaB*-b, ATP/GTP-binding site motif A (P-loop)
Mka GyrA	420	ND	DOD	Y?	GNVPPAAMRY/C	HN/TEARLTPLAM	*gyrA*-a, After active site Tyr
Mle DnaB	145	ND	ND	K233	IVAARPGVGK/A	HN/STLGLDFMRS	*dnaB*-b, ATP/GTP-binding site motif A (P-loop)
Mle GyrA	420	ND	DOD	Y130	GNDPPAAMRY/C	HN/TEARLTPLAM	*gyrA*-a, After active site Tyr

Table 1. (contd)

Mle Pps1	386	ND	DOD	G201	ALNTAVWSGG/C	HN/SFIYVPPGVH	*pps1*-a
Mle RecA	365	ND	DOD	G205	LREKIGVMFG/C	HN/SPETTTGGKA	*recA*-b, Disordered loop 2 that interacts with DNA
Mma GyrA	420	ND	DOD	Y?	GNNPPAAMRY/C	HN/TEAR	*gyrA*-a, After active site Tyr
Msh RecA	364	ND	DOD	G?	REKIGVMFG/C	HN/SPETTTGGKA	*recA*-b
Msm DnaB-1	139	ND	None	K237	IIAARPGVGK/A	HN/STLGLDFMRS	*dnaB*-b, ATP/GTP-binding site motif A (P-loop)
Msm DnaB-2	425	ND	DOD	G543	PQVSDLRESG/C	HN/SLEQDADMVM	*dnaB*-a
Mth RecA	365	ND	DOD	G?	REKIGVMFG/C	HN/SPETTTGGKA	*recA*-b
Mtu Pps1	359	ND	DOD	G252	EGSYVHYVEG/C	HN/CTAPIYKSDS	*pps1*-b
Mtu-CDC1551 DnaB	416	ND	DOD	G399	PMLADLRESG/C	HN/SLEQDADVVI	*dnaB*-a
Mtu-H37Rv DnaB	416	ND	DOD	G399	PMLADLRESG/C	HN/SLEQDADVVI	*dnaB*-a
Mtu-H37Rv RecA	440	PI-*MtuI*	DOD	K251	RTRVKVVKNK/C	HN/CSPPFKQAEF	*recA*-a
Mxe GyrA	198	ND	None	Y?	GNDPPAAMRY/C	HN/TEAPLTPLAM	*gyrA*-a, After active site Tyr
Npu DnaB	429	ND	DOD	G388	PMLSDLRESG/C	HN/SIEQDADLVI	dnaB-a
Npu DnaE	102+36	ND	None	Y776	FEQMLKFAEY/C	SN/CFNKSHSTAY	*dnaE*-a, Beta and tau binding domains
Npu GyrB	322	ND	HNH	G77	IPAAVGVDIG/C	HN/CGMSAIKTSF	*gyrB*-b
Nsp DnaB	429	ND	DOD	G388	PMLSDLRESG/C	HN/SIEQDADLVI	*dnaB*-a
Nsp DnaE	102+36	ND	None	Y775	FDDMLKFAEY/C	SN/CFNKSHSTAY	*dnaE*-a, Beta and tau binding domains
Nsp RIR	407	ND	DOD	R275	VTIVAGNIRR/C	HN/SAGMRQFISD	*RIR*-a
Rma DnaB	428	ND	DOD	G420	PQLSDLRESG/C	HN/SIEQDADVVL	*dnaB*-a
Spb RIR1	385	ND	DOD	L380	HISKVKFSNL/C	GN/CSEVLQSSQV	*RIR1*-b
Ssp DnaB	429	ND	DOD	G380	PMMSDLRESG/C	HN/SIEQDADLIM	*dnaB*-a
Ssp DnaE	123+36	ND	None	Y774	FDQMVKFAEY/C	AN/CFNKSHSTAY	*dnaE*-a, Beta and tau binding domains
Ssp DnaX	430	ND	DOD	E219	CRYKVYVIDE/C	HN/CHMLSTAAFN	*dnaX*-a, putative helicase DEAD-box motif

Ssp GyrB	435	ND	HNH	G436	FIVEGDSAGG/C	HN/SAKQGRDRRF	*gyrB*-a, Signature for DNA topoisomerase II
Ter GyrB	244	ND	HNH	G439	YLVEGDSASG/C	HN/SAKQGRDRRF	*gyrB*-a, Signature for DNA topoisomerase II
Ter Snf2	469	ND	DOD	K?	CLADDMGLGK/C	HN/TIQTIAFLLK	*snf2*-a, ATP/GTP-binding site motif A (P-loop)
Tfu RecA-1	422	ND	DOD	K95	EIYGPESSGK/C	HN/TTVALHAVAN	*recA*-c, ATP/GTP-binding site motif A (P-loop)
Tfu RecA-2	357	ND	DOD	G469	LREKVGVMFG/C	HN/SPETTSGGRA	*recA*-b

Archaea

Intein	Size	Endo activity	Endo motif	Location in Extein	N-terminal Splice junction	C-terminal splice junction	Insertion site
Ape Hyp	468	ND		Q175	LVSQYAITTQ/S	HN/SAFGWGLEHV	*Ape* hyp-a
Fac Pps1	356	ND	DOD	G242	EGAKVHYIEG/C	HN/CTAPKYNTSS	*pps1*-b
Fac RIR1	366	ND	DOD	P437	NIGYIESTNP/C	HN/CGEQPLLPYE	*RIR1*-b
Hsp-NRC1 CDC21	182	ND	None	K282	LLIGDPGTGK/C	HN/SQMISYVQNI	CDC21-a, ATP/GTP-binding site motif A (P-loop)
Hsp-NRC1 Pol II	195	ND	None	N925	PYFHAAKRRN/C	GQ/CDGDEDCVML	*pol* II-a
Mja GF-6P	499	ND	DOD	H74	DIDGNIGIGH/C	HN/SRWATHGNVC	GF6P-a
Mja Helicase	501	ND	DOD	L337	IKVICCTPTL/C	HN/SAGLNLPCRR	helicase-a, Motif V of DEAD and DEAH box helicases
Mja Hyp-1	392	ND	DOD	H128	DVGGLIGPAH/C	HN/CFTPWTSLYK	*Mja* hyp1-a
Mja IF2	546	ND	DOD	K30	CVLGHVDHGK/C	HN/TTLLDKIRKT	IF2-a, ATP/GTP-binding site motif A (P-loop)
Mja KlbA	168	ND	None	G404	LVAMNTGHDG/A	SN/CSGTLHANSA	*klbA*-a
Mja PEP	412	ND	DOD	T410	AIVTDEGGLT/C	FN/CHAAIVSREL	PEPsyn-a, Prosite signature PS00370
Mja Pol-1	369	ND	DOD	R425	FEDIISMDFR/C	HN/SLYPSIIISY	*pol*-a, Pol Motif A

Table 1. (contd)

Mja Pol-2	476	ND	DOD	N882	EQKSLKILAN/S	HN/SVYGYLAFPR	*pol*-b, Pol Motif B
Mja r-Gyr	494	ND	DOD	L866	IAQELFELGL/C	HN/CTYHRTSSTR	r-*gyr*-a, Putative active site Tyr
Mja RFC-1	548	ND	DOD	K53	LFSGPPGVGK/C	HN/TTAALCLARD	*RFC*-a, ATP/GTP-binding site motif A (P-loop)
Mja RFC-2	436	ND	DOD	A626	WRDNFLELNA/S	HN/SDERGIDVIR	*RFC*-b
Mja RFC-3	543	ND	DOD	S1124	YSDVCRFILS/C	HN/CNYPSKIIPP	*RFC*-c
Mja RNR-1	453	ND	DOD	Q337	NQMYVARGGQ/S	HN/TIFSSINLEL	*RNR*-a
Mja RNR-2	533	ND	DOD	S1058	WTVTQTPAES/S	HN/TAGRFARLDY	*RNR*-b
Mja Rpol A″	471	ND	DOD	M75	QSIGEPGTQM/S	HN/TMRTFHYAGV	*RpolA″*-a
Mja Rpol A′	452	ND	DOD	V463	YRTFRHNLCV/C	GN/CPPYNADFDG	*RpolA′*-a
Mja RtcB (*Mja* Hyp-2)	488	ND	DOD	N97	SPGGVGFDIN/C	HN/CGVRLIRTNL	*rtcB*-a
Mja TFIIB	335	ND	DOD	Y99	RCRVGAPMTY/S	HN/TIHDKGLSTV	TFIIB-a
Mja UDP GD	454	ND	DOD	S260	LNAGIGYGGS/C	HN/CFPKDVKALI	UDP GD-a, Active site
Mkn ATPase	394	ND	DOD	E634	PEVLSKWVGE/S	HN/SEKKIREIFQ	ATPase-a
Mkn NtpB	517	ND	DOD	Y260	LVILTDMTNY/C	SN/CEALREISAA	*ntpB*-a
Mkn RFC	305	ND	DOD	A82	WRDNFLELNA/S	HN/SDERGIDVIR	*RFC*-b
Mth RIR1	134	ND	None	P265	QLGRIEATNP/C	HN/CGEQPLLTHE	*RIR1*-b
Pab CDC21-1	164	ND	None	K334	LLVGDPGVAK/C	HN/SQLLRYIANL	CDC21-a, ATP/GTP-binding site motif A (P-loop)
Pab CDC21-2	268	ND		L525	SGKSSSAAGL/C	HN/TAAVVRDEFT	CDC21-b
Pab IF2	394	ND	DOD	K20	AVLGHVDHGK/C	HN/TTLLDRIRKT	IF2-a, ATP/GTP-binding site motif A (P-loop)
Pab KlbA	196	ND	None	G453	FTAMNTGHDG/A	SN/CMGTIHSNSA	*klbA*-a
Pab Lon	333	ND	DOD	Q220	LGDVRHDPFQ/C	KN/SGGLGTPAHL	*lon*-a
Pab Moaa	455	ND	DOD	Y155	TNRCNLNCWY/C	SN/CFFYAREGEP	*Moaa*-a
Pab Pol II	185	ND	None	N954	PYFHAAKRRN/C	HQ/CDGDEDAVML	*pol* II-a
Pab RFC-1	499	ND	DOD	K61	LFAGPPGVGK/C	HN/TTAALALARE	*RFC*-a, ATP/GTP-binding site motif A (P-loop)
Pab RFC-2	608	ND	DOD	S647	FSSNVRFILS/C	HN/CNYSSKIIEP	*RFC*-c
Pab RIR-1	399	ND	DOD	G301	VAMIQKMGGG/C	HN/TGLNFSKLRP	*RIR1*-a

Table 1 (contd.)

Pab RIR-2	438	PI-*Pab*I	DOD	A722	DIVGTTTGAA/C	SN/SGPVSFMHLI	*RIR1*-c
Pab RIR1-3	382	PI-*Pab*II	DOD	P1297	KGGPIRATNP/C	HN/CGEEPLYEYE	*RIR1*-b
Pab RtcB (Pab Hyp-2)	436	ND	DOD	N97	SPGGIGYDIN/C	HN/CGVRLIRTNL	*rtcB*-a
Pab VMA	429	ND	DOD	K240	AIPGPFGSGK/C	HN/TVTQHQLAKW	*VMA*-b, ATP/GTP-binding site motif A (P-loop)
Pfu CDC21	367	ND	None	L364	SGKSSSAAGL/C	HN/TAAAVRDEFT	CDC21-b
Pfu IF2	387	ND	DOD	K22	AVLGHVDHGK/C	HN/TTLLDRIRKT	IF2-a, ATP/GTP-binding site motif A (P-loop)
Pfu KlbA	522	ND	DOD	G463	FTAMNTGHDG/A	SN/CMGTIHANSA	*kblA*-a
Pfu Lon	401	ND	DOD	Q209	LGDVRHDPFQ/C	KN/SGGLGTPAHE	*lon*-a
Pfu RFC	525	ND	DOD	K59	LFAGPPGVGK/C	HN/TTAALALARE	*RFC*-a, ATP/GTP-binding site motif A (P-loop)
Pfu RIR-1	454	PI-*Pfu*I	DOD	G301	VAMIQKMGGG/C	HN/TGLNFSKLRP	*RIR1*-a
Pfu RIR-2	382	PI-*Pfu*II	DOD	P914	KGGPIRATNP/C	HN/CGEEPLYEYE	*RIR1*-b
Pfu RtcB (*Pfu* Hyp-2)	. 380	ND	DOD	N97	SPGGIGYDIN/C	HN/CGVRLIRTNL	*rtcB*-a
Pfu TopA	373	ND	DOD	F314	IAQSLYEKGF/C	HN/CSYPRTESQK	r-*gyr*-a, 2 aa before the catalytic Y residue
Pfu VMA	425	ND	DOD	K240	AIPGPFGSGK/C	HN/TVTQHQLAKW	*VMA*-b, ATP/GTP-binding site motif A (P-loop)
Pho CDC21-1	168	ND	None	K334	LLVGDPGVAK/C	HN/SQLLRYVANL	CDC21-a, ATP/GTP-binding site motif A (P-loop)
Pho CDC21-2	260	ND	Unknown	L529	SGKSSSAAGL/C	HN/TAAVVRDEFT	CDC21-b
Pho IF2	444	ND	DOD	K22	AVLGHVDHGK/C	HN/TTLLDKIRKT	IF2-a, ATP/GTP-binding site motif A (P-loop)
Pho KlbA	520	ND	DOD	G451	FTAMNTGHDG/A	SN/CMGTIHSNSA	*kblA*-a
Pho LHR	475	ND	DOD	V346	LKRGELRAVV/C	KN/SSTSLELGID	*LHR*-a
Pho Lon	474	ND	DOD	Q210	LGDVRHDPFQ/C	KN/SGGLGTPAHL	*lon*-a
Pho Pol I	460	ND	DOD	N492	RQRAIKILAN/S	HN/SYYGYYGYAK	*pol*-b, Pol Motif B
Pho Pol II	166	ND	None	N954	PYFHAAKRRN/C	HQ/CDGDEDAVML	*pol* II-a
Pho r-Gyr	410	ND	DOD	L953	LAQDLFEAGL/C	HN/CTYHRTDSIH	r-*gyr*-a, 2 aa before the catalytic Y residue

Table 1 (contd.)

Pho RadA	172	ND	None	K152	EVFGEFGSGK/C	HN/TQLAHTLAVM	*recA*-c, ATP/GTP-binding site motif A (P-loop)
Pho RFC	525	ND	DOD	K61	LFAGPPGVGK/C	HN/TTAALALSRE	*RFC*-a, ATP/GTP-binding site motif A (P-loop)
Pho RIR1	385	ND	DOD	P467	KGGPIRATNP/C	HN/CGEEPLYEYE	*RIR1*-b
Pho RtcB (*pho*-Hyp-2)	390	ND	DOD	N97	SPGGIGYDIN/C	HN/CGVRLIRTNL	*rtcB*-a
Pho VMA	376	ND	DOD	K240	AIPGPFGSGK/C	HN/TVTQHGLAKW	*VMA*-b, ATP/GTP-binding site motif A (P-loop)
Psp-GBD Pol	537	PI-*Psp*I	DOD	N492	RQRAIKILAN/S	HN/SYYGYYGYAK	*pol*-b, Pol Motif B
Tac-ATCC25905 VMA	173	ND	None	K235	RVPGPFGSGK/C	HN/TVIQHQLAKW	*VMA*-b, ATP/GTP-binding site motif A (P-loop)
Tac-DSM1728 VMA	174	ND	None	K235	AVPGPFGSGK/C	HN/TVIQHQLAKW	*VMA*-b, ATP/GTP-binding site motif A (P-loop)
Tag Pol-1 (Tsp-TY Pol-1)	360	ND	DOD	R409	WENIAYLDFR/C	HN/SLYPSIIVTH	*pol*-a, Pol Motif A
Tag Pol-1 (Tsp-TY Pol-2)	538	ND	DOD	N854	RQRAVKLLAN/S	HN/SYYGYMGYPK	*pol*-b, Pol Motif B
Tag Pol-1 (Tsp-TY Pol-3)	157	ND	None	D1441	KFGFKVLYAD/S	HN/TDGFYATIPG	*pol*-c, Pol Motif C
Tfu Pol-1	360	PI-*Tfu*I	DOD	R406	WENIAYLDFR/C	HN/SLYPSIIISH	*pol*-a, Pol Motif A
Tfu Pol-2	389	PI-*Tfu*II	DOD	D900	KFGFKVLYAD/S	HN/TDGFFATIPG	*pol*-c, Pol Motif C
Thy Pol-1	537	ND	DOD	N458	RQKAIKILAN/S	HN/SYYGYYGYAR	*pol*-b, Pol Motif B
Thy Pol-2	389	PI-*Thy*I	DOD	D1044	RFGFKVLYAD/S	HN/TDGFFATIPG	*pol*-c, Pol Motif C
Tko Pol-1 (Pko Pol-1)	360	PI-*Pko*I	DOD	R406	WENIVYLDFR/C	HN/SLYPSIIITH	*pol*-a, Pol Motif A
Tko Pol-2 (Pko Pol-2)	536	PI-*Pko*II	DOD	N851	RQRAIKILAN/S	HN/SYYGYYGYAR	*pol*-b, Pol Motif B
Tli Pol-1	538	PI-*Tli*II	DOD	N494	RQRAIKLLAN/S	HN/SYYGYMGYPK	*pol*-b, Pol Motif B
Tli Pol-2	390	PI-*Tli*I	DOD	D1081	KRGFKVLYAD/S	HN/TDGFYATIPG	*pol*-c, Pol Motif C
Tsp-GE8 Pol-1	535	ND	DOD	N491	RQRAIKILAN/S	HN/SYYGYYGYAK	*pol*-b, Pol Motif B
Tsp-GE8 Pol-2	389	ND	DOD	D1075	KFGFKVLYAD/S	HN/TDGFFATIPG	*pol*-c, Pol Motif C
Tvo VMA	186	ND	None	K235	AVPGPFGSGK/C	HN/TVIQHQLAKW	*VMA*-b, ATP/GTP-binding site motif A (P-loop)

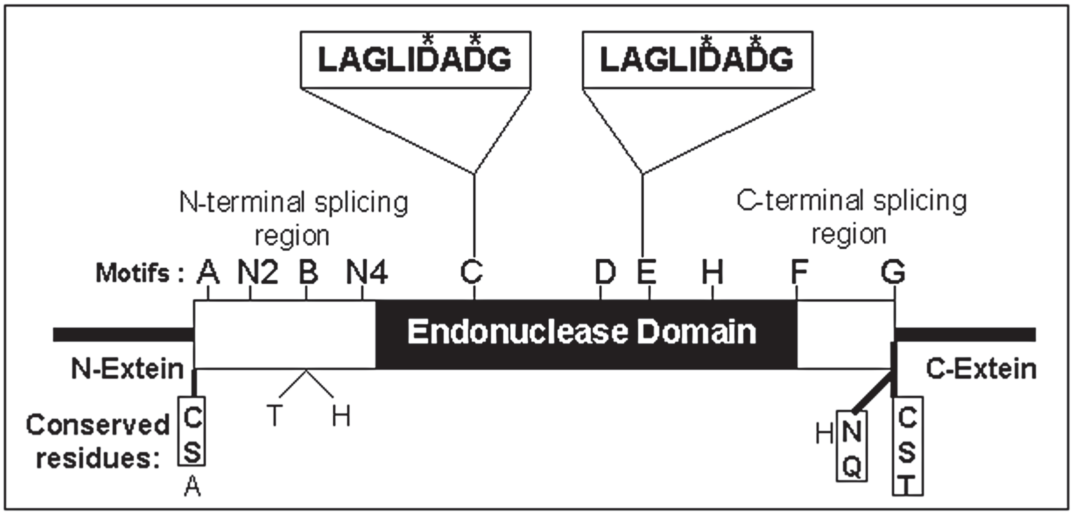

FIGURE 2. Structural organization of LAGLIDADG family of inteins. Linear representation of the LAGLIDADG family of inteins showing a central (closed and open) region corresponding to ENase and protein splicing domain flanked by N- and C-Extein sequences (thick lines). The conserved amino acid residues at the amino- and carboxyl-terminal splice sites involved in protein splicing are in bold single-letter codes and boxed. Inteins contain 10 conserved motifs designated as A, B, C, D, E, H, F, G, N2, and N4. Of these, the central four (blocks C, D, E, H) contribute to ENase function, whereas the remaining to protein splicing activity. The dodecapeptide motif containing the active site aspartate or glutamate (indicated by asterisk) is normally located in blocks C and E.

TABLE 2
Consensus Sequence Motifs of LAGLIDADG Family of Inteins

BLOCK	CONSENSUS SEQUENCE	MOTIF
A	ChXXDpXhhhXXG	N-terminal splicing
B	GXXhXhTXXHXhhh	N-terminal splicing
C	LhGXXhhaG	DOD
D	XKXIPXXh	DNA binding
E	XLXGhFahDG	DOD
H	pXSXXhhXXhXXLLXXhGI	DNA binding
F	rVYDLpV[1-3 residues]aXX[H or E]NFh	C-terminal splicing
G	NGhhhHNp	C-terminal splicing

(Uppercase letters represent the single-letter amino acid residues and lowercase letters represent amino acid groups: X, any residue; h, hydrophobic residues (G, A, V, L, I, M); p, polar residues (S, C, T); a, acidic residues (D or E); r, aromatic residues (F, Y, W); DOD – Dodecapeptide motif) (reviewed by Perler *et al.*, 1997a; Gogarten *et al.*, 2002).

thus changing the phenotype of the yeast cell (Strathern *et al*., 1982; Kostriken *et al*., 1983). The double-strand break is repaired by the components of homologous recombination using sequences from either of the two silent loci (*HML* or *HMR*) that flank the *MAT* locus. As a consequence of switching, daughter cells display mating-type opposite to that of their mother cells (Klar, 1987; Bobola *et al*., 1996; Long *et al*., 1997). Similar to mating-type switching, "intron/intein homing" is also initiated by a site-specific ENase.

Initial studies on intron- and intein-encoded ENases showed that these enzymes contain a conserved LAGLIDADG motif and, consequently, are grouped under the family of LAGLIDADG HEases (Hensgens *et al*., 1983). Later, characterization of phage T4 intron-encoded HEases revealed a new GIY-YIG motif (Bell-Pedersen *et al*., 1990), and sequence analysis using Hidden-Markov algorithm identified H-N-H motif in HEases (Ellison and Vogt, 1993). Based on the type of conserved sequence motifs, HEases are categorized into four families:

(a) LAGLIDADG

(b) GIY-X$_{(10\text{-}11)}$-YIG

$$\left.\begin{array}{l}\textbf{(c) H-N-H}\\[1em] \textbf{(d) His-Cys}\end{array}\right\}\ \beta\beta\alpha\textbf{-Me}$$

An array of 22 amino acids containing three of the four secondary structural elements is shared between H-N-H and His-Cys box enzymes. Interestingly, the catalytic metal-ions are located at identical positions in the crystal structures in both families of HEases. Because of such remarkable structural similarities, the His-Cys box and H-N-H ENases have been grouped under the category of *ββα*-**Me** (three secondary structural elements and a metal-ion) family (Kuhlmann *et al*., 1999). However, little or no structural homology exists outside this central core (reviewed by Chevalier and Stoddard, 2001b). Indeed, one additional feature of His-Cys box family of HEases, which distinguish them from H-N-H family, is the presence of tightly bound and catalytically active zinc in the core-fold. In this review, we focus on struructual and functional characteristics of each HEase family with relevant examples.

A. LAGLIDADG Family of HEases

The LAGLIDADG family of HEases (with >200 members known) is the much-studied family of HEases. Members of this family contain one or two copies of the conserved LAGLIDADG motif. The two motifs in the monomer, referred to as P1 and P2, are separated by ~80 to 150 amino acid residues (Hensgens *et al*., 1983; Michel and Cummings, 1985; Perlman and Burtow, 1989). The HEases, which harbor one LAGLIDADG motif bind as homodimers to a pseudo-palindromic target DNA sequence with two-fold symmetry (I-*Cre*I: Chevalier *et al*., 2001a; I-*Ceu*I: Turmel *et al*., 1997), whereas those with two motifs bind as monomers to asymmetric target DNA sequences (I-*Dmo*I: Dalgaard *et al*., 1994; PI-*Sce*I: Christ *et al*., 1999). However, both types bind long recognition sites (20 to 30 bp) and inflict a DSB in their homing site resulting in cohesive ends, usually with 4 nucleotide 3' overhangs.

Although LAGLIDADG enzymes differ significantly in their primary structure, they display similar topological arrangement of helices in their core structure (α1-β1-β2-α2-β3-β4-α3) (e.g., PI-*Pfu*I, Figure 3). The α1 helix in each subdomain of a monomeric or dimeric HEase contains the dodecapeptide motif. The two motifs (in a monomer or dimer) are juxtaposed by van der Waals interaction related by a pseudo-twofold symmetry (Figure 3). The conserved Gly at the bottom of the helices induces a sharp turn and positions the acidic residue (Asp^{149}/Glu^{250}, Figure 3) at the protein interface to place the divalent cation at its active site. Whereas Asp or Glu embedded in the LAGLIDADG motif of the α-helix is implicated in hydrolysis of the phosphodiester bond, the saddle formed by the β-strands offer specificity for binding of LAGLIDADG HEases to the entire length of the target recognition sequence (Figures 5 and 6).

S. cerevisiae TFP1 intein (later termed as *VMA*1-subunit of V̲acuolar M̲embrane A̲TPase) was the first LAGLIDADG intein HEase to be identified and characterized. Sequence analysis of *VMA1* intein showed a high degree of homology (34% identical) to *S. cerevisiae* HO ENase (Gimble and Thorner, 1992; Dalgaard *et al*., 1997a; reviewed in Gogarten *et al*., 2002). Analogous to intron-encoded ENases, VDE

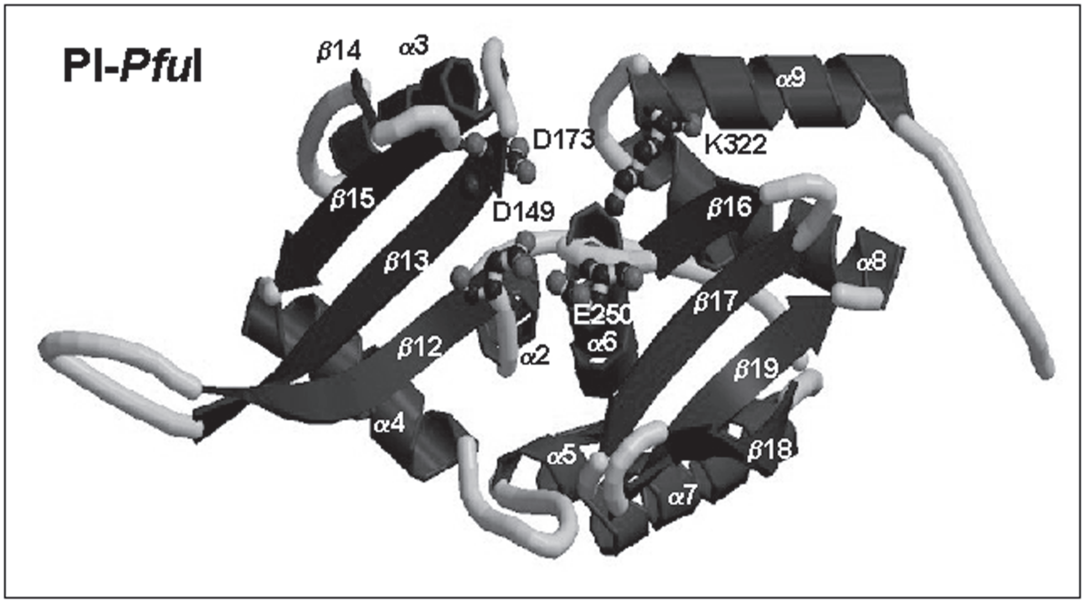

FIGURE 3. Ribbon diagram of Archaeal intein PI-*Pfu*I endonuclease domain. The putative active site acidic amino acid residues Asp[149] and Glu[250] from the two dodecapeptide motifs present in α2 and α6 helices, and the Asp[173] and Lys[322] presumed to facilitate catalysis are indicated. This figure was prepared by MOLSCRIPT using the coordinates obtained from Protein Data Bank (1DQ3.pdb) and reproduced with permission from Elsevier Press.

(Vacuolar membrane ATPase Derived Endonuclease; PI-*Sce*I according to current nomenclature) cleaves inteinless allele efficiently in the presence of Mg^{2+} (Gimble and Thorner, 1992). In addition, PI-*Sce*I cleaves at a single site in bacteriophage λ DNA in the presence of Mn^{2+}. Interestingly, cleavage efficiency of PI-*Sce*I is enhanced (2- to 2.5-fold) if the inteinless allele is embedded in a supercoiled plasmid, suggesting a role for the topological context of the homing site (Wende *et al*., 1996). By primer extension analysis, the precise location of cleavage by PI-*Sce*I was mapped in both VMA*Δvde* and λ-DNA. In both the cases, PI-*Sce*I generates 3' extended cohesive ends with four nucleotide 3' hydroxyl overhangs and 5' phosphates exactly at the position where PI-*Sce*I coding sequence was located in the *VMA*1 gene. The cleavage site sequence between VMA*Δvde* and λ-DNA were substantially similar (over a 37-bp stretch) except for a C•G → T•A transition in λ-DNA. Subsequent analysis of the cleavage site revealed that PI-*Sce*I might promote homing of its coding element *in vivo* (Gimble and Thorner, 1993). In general, the solution conditions required for most LAGLIDADG HEases to display ENase activity include 1 to 10 m*M* of divalent cation (Mg^{2+}or Mn^{2+}), alkaline pH, and monovalent cations (Na^+ or K^+). The presence of Mn^{2+} induces relaxation in the specificity of cleavage sequence by PI-*Sce*I and PI-*Pfu*I (Gimble and Thorner, 1992; Komori *et al*., 1999a).

B. Common Tri-Partite Structures of Intein-Encoded LAGLIDADG HEases

Structural studies revealed that the tripartite structure of PI-*Pfu*I and PI-*Sce*I contain, in addition to the protein splicing and ENase domains, an additional domain called the stirrup domain or DNA recognition region (DRR) (Ichiyanagi *et al*., 2000; Duan *et al*., 1997). Although both enzymes appeared to be topologically similar in their three-dimensional structures, but display distinct differences: (1) β-strands in PI-*Sce*I are twice as long as PI-*Pfu*I and, consequently, the DRR in PI-*Sce*I shows extended conformation; (2) DRR domain of PI-*Pfu*I makes contact with the other two domains, while it is not in PI-*Sce*I, and finally (3) DRR domain is inserted in the linker region between ENase and splicing domains, whereas it is present in an internal loop within splicing domain (Figure 4). The DRR domain (~9 kDa) has no sequence similarity to any other protein in the sequence database. Surprisingly, the protein-splicing domain of inteins (PI-*Pfu*I, PI-*Sce*I, *Mxe* GyrA) displayed structural similarities to the developmentally important hedgehog proteins. Therefore, the fold in the

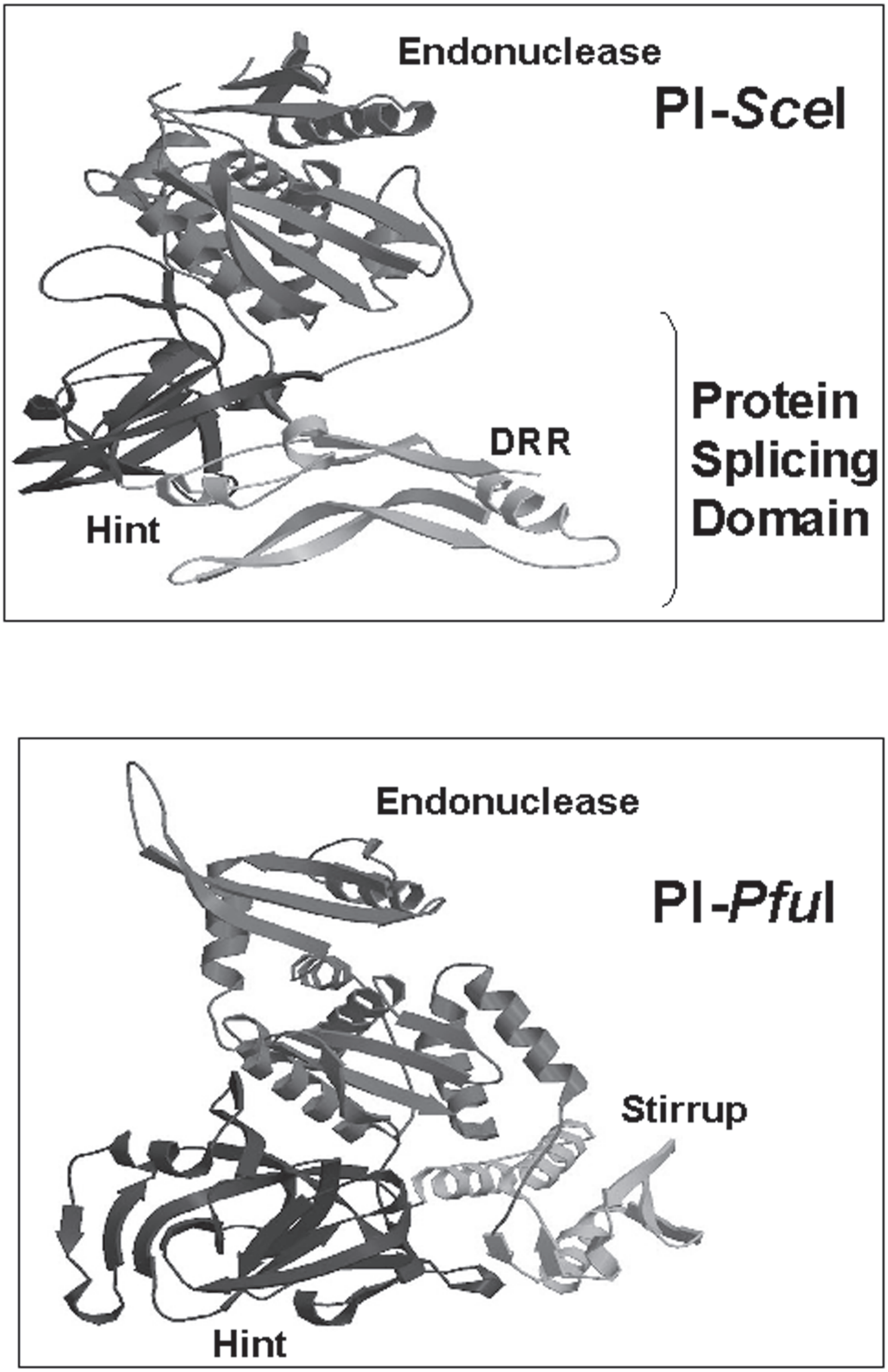

FIGURE 4. Ribbon diagram representing tripartite structure of PI-*Sce*I and PI-*Pfu*I. The 'Hint' domain, 'ENase' domain and the DRR or Stirrup domain represent the tripartite structure of the two-motif HEases. For further details see the text. (Reproduced with permission from Ichiyanagi *et al.*, 2000; Duan *et al.*, 1997 and, Elsevier Press and Nature publishing group.) This figure was prepared by MOLSCRIPT using the coordinates obtained from Protein Data Bank (1VDE.pdb for PI-*Sce*I, and 1DQ3.pdb for PI-*Pfu*I).

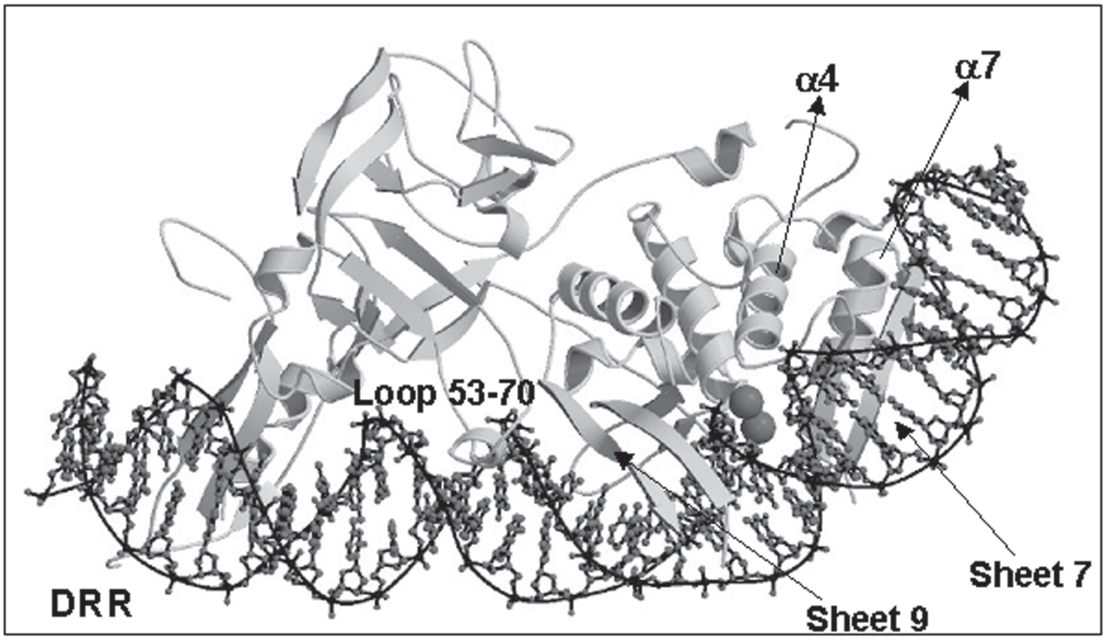

FIGURE 5. Co-crystal structure of PI-*Sce*I with its cognate DNA. Ribbon diagram of PI-*Sce*I bound to its homing site DNA (ball and stick model). Loop 53 – 70, sheets 7 and 9, and the loop in DRR are involved in DNA recognition. Helices (α4 and α7) containing the LAGLIDADG motifs are labeled correspondingly, and the two calcium ions are indicated as grey spheres near the point of cleavage in the DNA. The structure is reconstructed from the coordinates using MOLSCRIPT (1LWS.pdb) and reproduced with permission from Nature publishing group.

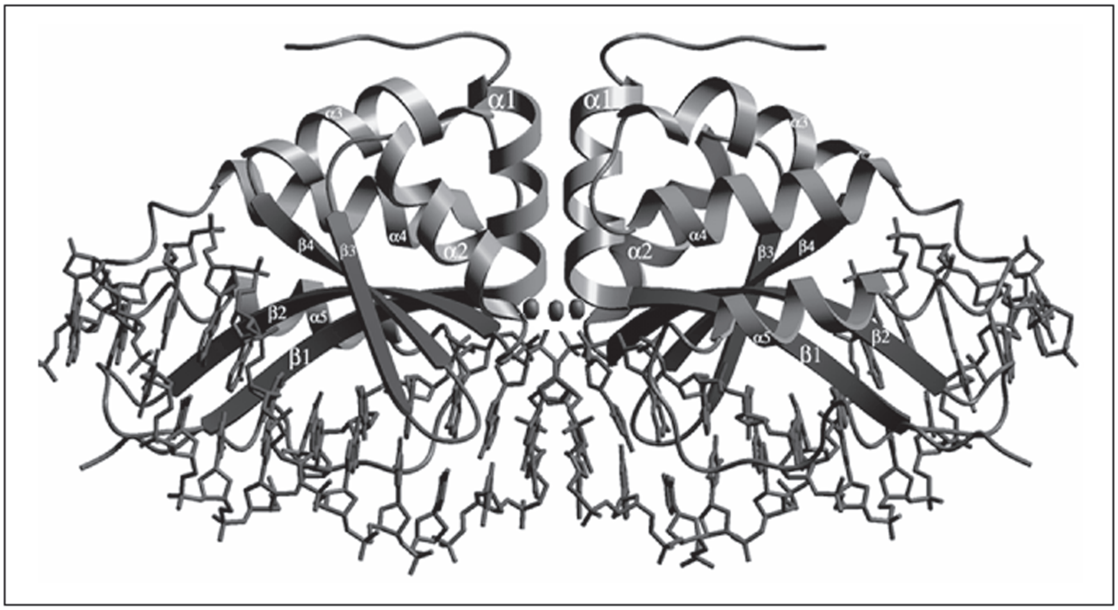

FIGURE 6. Crystal structure of I-CreI complexed with its DNA homing site. The Ca^{2+} ions in the substrate bound I-*Cre*I inactive complex is coordinated by Asp^{20} from LAGLIDADG motif, oxygen atom from scissile phosphate and another phosphate directly across the minor groove of the cleavage site. For further details see the text. (This figure is reproduced with permission from Jurica *et al*., 1998 and Elsevier Press.)

protein-splicing domain is designated as 'HINT' module (Hedgehog INTein module) (Hall *et al*., 1997; Perler, 1998; Amitai *et al*., 2003).

C. Binding and Distortion of DNA by LAGLIDADG HEases

Studies of many DNA-binding proteins, particularly REases, have provided evidence that the mechanism of search for target DNA sequence involves binding first in a sequence non-specific manner, followed by sliding along the DNA. It is known that many LAGLIDADG HEases show high affinity ($K_D \approx 5$ to 30 n*M*) to DNA, but the mechanism by which they locate target DNA sequences remains to be elucidated. Consistent with biochemical analysis, PI-*Sce*I was shown to bind Region I (proximal to cleavage site) of target DNA sequence via the major groove (Gimble and Wang, 1996; Pingoud *et al*., 1998, 1999; Moure *et al*., 2002). The complementary groove is laced with positively charged amino acids of the ENase domain, which are required for protein-DNA contacts as well as to confer stability to the complex (Moure *et al*., 2002). The β-sheets in DNA recognition region of splicing domain shows ordered structure. This domain establishes contact with Region II, which is downstream of the cleavage site via the major groove (Gimble and Wang, 1996; Figure 5). The loops formed by residues 53 to 70 (between splicing and ENase domain) and 269 to 284 (specific only for PI-*Sce*I) contact DNA through minor grooves. Together these findings suggest that DNA-binding causes several regions of PI-*Sce*I to become more ordered when compared with the apo-enzyme. Interestingly, Lys^{340} is the only residue in the ENase domain that showed decreased DNA-binding when mutated, suggesting the interaction occurs between the two domains during sequence recognition (Moure *et al*., 2002).

In the co-crystal structure of I-*Cre*I bound to the 24-bp homing site, extended and concave β-ribbons formed by four anti-parallel β-strands from each monomer interact with the homing half-sites, with the longest two strands (β1 and β2) contributing eight amino acids to the interface (Figure 6). The β-ribbons from each monomer is part of a saddle of ~18 Å in diameter and display helical twist to maintain direct protein-DNA contacts across nine consecutive base pairs (3 to 11) in each homing half-site. The length and flexibility of these β-strands facilitates the recognition of extended homing site sequence motifs by providing 6 to 7 Å spacing between side chains on the protein interface, parallel to that of alternative base pairs and phosphate groups in B-DNA (Phillips, 1994). The large surface area created by extensive molecular interactions confers stability to I-*Cre*I - DNA complex (Jurica *et al*., 1998). I-*Cre*I displays three substantial conformational changes after binding to the homing site DNA. First, the amino acids in the loop (Asn^{30}, Ser^{32}, and Tyr^{33}) connecting the strands β1 and β2 provide a "twist" in the β-ribbon, while contacting nucleotides at the end of homing site. Secondly, the last 16 residues (138 to 153) in I-*Cre*I make a number of nonspecific phosphate backbone contacts and hence become ordered after binding DNA. Finally, the amino acids 113 to 123, which are disordered in the unbound protein (Heath *et al*., 1997) become ordered and clearly visible in the DNA-bound form (Jurica *et al*., 1998). In addition, interaction of the synthetic H-*Dre*I with its hybrid target site (fusion of I-*Cre*I and I-*Dmo*I half-sites) is also mediated by the extended β-sheets through the major groove. Interestingly, the contacts made across the I-*Cre*I half-site resembled that of wild-type protein co-crystal complex. As seen in other HEases, the hydrogen bonds at the protein-DNA interface was undersaturated (48 out of 96 potential bonds) (Chevalier *et al*., 2002)

Normally, proteins that bind to DNA in a sequence nonspecific manner are characterized by their ability to release counter ions and water molecules from the protein-DNA interface. On the other hand, some proteins that bind in a sequence-specific manner cause distortion in the DNA due to increased protein-DNA contacts. Such distortions are believed to augment catalysis in one of the following three ways: facilitate contact between small proteins and longer DNA sequences, promote transition state formation, position scissile phosphates at active site, or bend DNA in the major groove to widen the minor groove and facilitate catalysis by allowing access to scissile phosphate (Gimble and Wang, 1996; Wende *et al*., 1996). A number of approaches, including biochemical, circular permutation and phasing analyses, and X-ray crystallography, have been used to monitor distortion in the target site DNA induced by HEases. PI-*Sce*I was found to distort DNA at its homing site by two different angles consequent to two different modes of interaction. The result from circular permutation analysis was further corroborated by the co-crystal structure of PI-*Sce*I bound to

31-bp DNA (3.5 helical turns) (Moure *et al.*, 2002). In the binary complex, the ENase domain of PI-*Sce*I distorted the DNA to larger extent (55^{o}) when compared with a minor bend (22^{o}) induced by protein splicing domain. However, in the case of I-*Cre*I-DNA complex the degree of bending is ~10^{o}, which is at the middle of each half-site near the fifth base pair position. This is due to the interaction between β groove of the enzyme and the complementary major groove of the homing site DNA. In addition, W:C base pairs are maintained throughout the homing site, with many exhibiting a measurable propeller twist, particularly near the cleavage site (Jurica *et al*., 1998). Similarly, other LAGLIDADG HEases have also been shown to induce distortion at their respective homing sites as follows: PI-*Pfu*I (73^{o}), PI-*Pfu*II (67^{o}) (Komori *et al*., 1999b), and I-*Sce*I (Beylot and Spassky, 2001).

D. Base-Specific Interactions by LAGLIDADG HEases

The sequence-specific interaction between HEases and bases within the recognition sequence involve an extensive network of hydrogen bonds with donors and acceptors. The evidence from affinity cleavage studies of PI-*Sce*I indicates that Arg^{90} and Arg^{94} are involved in binding to DNA, in a sequence-specific manner (He *et al*., 1998). Similarly, photo-crosslinking of PI-*Sce*I to oligonucleotides has unveiled interactions between DNA Recognition Region (DRR) and distal region of the recognition sequence (Pingoud *et al*., 1999). As suggested, the additional contacts that were not in close proximity to the cleavage site are dispensable for the enzyme's function (Christ *et al*., 2000). Biochemical and structural characterization of PI-*Sce*I bound to the recognition sequence has established the molecular basis for DNA recognition (He *et al*., 1998; Hu *et al*., 2000; Moure *et al*., 2002). Three notable features that emerge from interaction between PI-*Sce*I and DNA include: (1) 3 bp (A/T^{+16}, G/C^{+18}, and A/T^{+19}) in the homing site and Arg^{90} and Arg^{94} are absolutely essential for binding; (2) DNA was distorted (~70^{o}) to fit the binding pocket; and (3) contact between Lys^{376} and Lys^{378} with DNA was found to be located in the major groove near the two guanosine residues on the top strand (G^{+13} and G^{+18}).

The high-resolution cocrystal structures of I-*Cre*I-DNA and H-*Dre*I-DNA shed light on how the HEases make base-specific contacts with DNA. I-*Cre*I makes contact with nine consecutive base pairs (3 to 11) in each of the half-site through the major groove. Interaction involves 12 direct contacts to atoms of the nucleotide bases and three water-mediated interactions. Most of the DNA–protein contacts involve hydrogen bond (15 of 36) donors and acceptors from the major groove of the homing site DNA. Except for the three bases (at position 3, 5, and 9) that make more than one direct atomic interaction with the protein, remaining positions make single side-chain contacts allowing flexibility at these positions. Similar results were obtained by genetic randomization studies and methylation interference analysis (Argast *et al*., 1998; Wang *et al*., 1997). However, the base pairs flanking the scissile phosphate and the final base pair at each end of the homing site are not involved in the interaction (Jurica *et al*., 1998). E-*Dre*I, the synthetic HEase makes 32 direct hydrogen bonds and 16 water-mediated contacts with the nucleotides in the chimaeric DNA target site. Tyr^{29}, Gly^{31}, Ser^{34}, Arg^{33}, Glu^{35}, Arg^{37}, Arg^{81}, Glu^{79}, Asp^{75}, Thr^{76} and Arg^{77} of I-*Dmo*I from the amino-terminal domain, and Gln^{123}, Lys^{125}, Asn^{127}, Tyr^{130}, Gln^{135}, Gln^{141}, Arg^{165}, Arg^{167}, and Asp^{172} of I-*Cre*I at the carboxyl terminal domain of H-*Dre*I are involved in interaction with specific nucleotides corresponding to the chimaeric DNA half-sites (Chevalier *et al*., 2002).

E. Active Site Residues of LAGLIDADG HEases

Site-directed mutagenesis of most conserved acidic amino acids (Asp^{218}, Asp^{229}, and Asp^{326} in PI-*Sce*I; Asp^{149}, Glu^{250} in PI-*Pfu*I; Asp^{156}, Asp^{249} in PI-*Pfu*II; Asp^{122}, Glu^{220} and Asp^{222} in PI-*Mtu*I; Asp$^{20, 20}$ in I-*Cre*I, and Asp^{21} and Glu^{117} of I-*Dmo*I, respectively) in the LAGLIDADG motifs have implicated a role for these residues in DNA cleavage (Gimble and Stephens, 1995; Komori *et al*., 1999b; Schottler *et al*., 2000; Guhan and Muniyappa, unpublished; Lykke-Anderson *et al*., 1997b). Mutation of conserved residues in the LAGLIDADG motif had a modest or no effect on the ability of mutant enzymes to bind DNA in a sequence-specific manner. However, mutation of Asp^{218} (in catalytic center I) or Asp^{326} (in catalytic center II) led to loss of activity in the presence of Mg^{2+} but not Mn^{2+}, indicating that amino acids at the relevant positions are involved in binding metal ions. Furthermore, while mutation of third residue, Asn^{225}, in the active site of PI-*Pfu*I did not overtly affect catalysis, mutation of Lys^{224} reduced its cleavage activity by 50-fold, suggesting that the crucial third residue may play a subtle role in setting the catalytic threshold for these enzymes (Komori *et al*., 1999b).

HEases, including I-*Cre*I, I-*Sce*I, I-*Sce*II, I-*Tli*I, and I-*Ceu*I, bind and cleave their target DNA substrates to generate four nucleotide 3' overhang (Monteilhet *et al*., 1990; Colleaux *et al*., 1988; Durrenberger and Rochaix, 1993; Jurica *et al*., 1998). In these enzymes, in addition to the metal-ion binding acidic amino acids, Lys^{98}, Arg^{51}, Asp^{137}, Gln^{47}, and Gly^{19} are essential for coordinating the hydrogen bond network at the active site for catalysis (Aggarwal and Wah, 1998; Chevalier *et al*., 2001b). Each metal ion in the I-*Cre*I DNA bound complex is independently coordinated by a single Asp^{20} residue. The contacts consist of oxygen from the scissile phosphate, additional oxygen from phosphate across the minor groove, main chain carbonyl group (residue 19) from the opposing monomer, and a water molecule. An interesting feature is that Arg^{51} and Lys^{98} located within a specific distance from the scissile phosphate in the inactive complex might act as Lewis acids in order to stabilize the pentavalent transition state of the cleaved phosphate during catalysis. In this regard, the residues equivalent to Lys$^{98, 98}$ in I-*Cre*I is well conserved in PI-*Sce*I (Lys$^{301, 403}$) (Jurica *et al*., 1998). Three I-*Cre*I allelic intron HEases, I-*Mso*I, I-*Pak*I, I-*Cvu*I, also recognize and cleave DNA sequences in a similar manner, albeit with different efficiencies. However, the five residues implicated to play a role in catalysis in case of I-*Cre*I are conserved, but not the residues that are involved in interaction with DNA. Intriguingly, sequence comparison of single-motif HEases with double-motif HEases revealed that the active site residues in the latter are less well conserved than the former (Lucas *et al*., 2001). The crystal structure of H-*Dre*I has revealed the presence of three Mg^{2+} ions at the active site in the substrate bound complex, with the central metal-ion being shared between both the active sites (D21 and D117). In addition to the normal contacts made by the amino acids (Gln^{42}, Gln^{144}, Arg^{148}, and Lys^{195}) at or near the active site periphery, Trp^{19}, which was designed to stack against Phe^{151}, made additional contacts with the active site residues (Chevalier *et al*., 2002).

F. Mechanism of DNA Cleavage by LAGLIDADG HEases

Several lines of evidence support the notion that ENases cleave double-stranded DNA by two different mechanisms. According to the sequential cleavage mechanism, the ENase makes an incision in one of the strands followed by nicking in the second strand. On the other hand, the concerted cleavage mechanism posits simultaneous cleavage of both the strands by ENase. The cleavage mechanism(s) utilized by LAGLIDADG HEases are diverse, apparently because their active sites are divergent. HEases such as PI-*Pfu*II and PI-*Pab*II have been shown to cleave both strands of duplex DNA at the homing site in a concerted manner (Komori *et al* 1999a; Saves *et al.,* 2002b). While in the case of PI-*Sce*I, although biochemical experiments favor concerted cleavage mechanism (Christ *et al*., 1999), structural analysis suggests an ordered cleavage mechanism (Moure *et al*., 2002). The observation that two Ca^{2+} ions in the PI-*Sce*I–DNA complex are colinear with the top strand scissile phosphate lends itself to the possibility that it might be cleaved first (Figure 5, Moure *et al*., 2002). Alternatively, the differences could arise because the cleavage of the two DNA strands could be tightly coupled.

The homing site recognized and cleaved by I-*Cre*I is 22 bp in length. The high-resolution crystal structures of I-*Cre*I bound to DNA, substrate complex with calcium and a product complex with magnesium have been determined. These studies together with previous biochemical analysis of the cleavage pattern suggest that I-*Cre*I forms most of its sequence-specific contacts in the major groove and cleaves the homing site across the minor groove. The kinetic mechanism of double-strand cleavage follows a concerted mechanism (Chevalier *et al*., 2001a). Other studies have demonstrated that PI-*Pfu*I cleaves DNA by a sequential cleavage mechanism (Komori *et al*., 1999a). Further advances in our understanding of the preference for sequential mechanism emerged after structural analysis, where Asp^{149} assisted by Lys^{322} and Asp^{173} cleave the bottom strand efficiently, and Glu^{250} with no assistance cleaves the top strand inefficiently (Ichiyanagi *et al*., 2000). Interestingly, cleavage efficiencies of the two strands by PI-*Pab*I were notably different (Saves *et al*., 2002b). Other HEase, PI-*Tfu*I cleaves its 16 bp nonpalindromic homing site in negatively supercoiled DNA by a sequential mechanism in the presence of Mn^{2+} or Mg^{2+} (Saves *et al*., 2000a). Our recent studies indicate that PI-*Mtu*I cleaves homing as well as ectopic DNA sites by a sequential mechanism (Guhan and Muniyappa, 2002a, 2002b). It seems safe to conclude that in general, both mechanisms are used by HEases to cleave double-stranded DNA at specific target sites.

In vivo, PI-*Sce*I was shown to cleave the homing site embedded in *SNF3* of *S. cerevisiae* (Bremer *et al*., 1992). Southern analysis of genomic DNA from vegetative cells and meiotic haploids revealed that homing was meiosis specific (Gimble and Thorner, 1992). The possible cause for the absence of PI-*Sce*I promoted homing in vegetatively dividing cells is due to lack of karyopherin-mediated nuclear import of VDE (Nagai *et al*., 2003). In *VMA1Δvde/ VMA1* cells, PI-*Sce*I activity follows super-Mendelian segregation pattern, and it is absolutely essential for homing process. In the case of I-*Cre*I, ectopic expression of its HEase activity resulted in mobilization of the intron sequence into the 23S cDNA in the haploid progeny (Durrenberger *et al*., 1996). Interestingly, gene conversion was observed only when the 23S cDNA and the neighboring 23S gene were in opposite orientations.

G. LAGLIDADG HEases with Unique Biochemical Properties

PI-*Pfu*I cleaves pUC118 DNA at a unique site either in the presence of Mn^{2+} or at low concentrations of KCl. An analysis of the cleavage sequence revealed that 14 of 30 nucleotide residues was identical to its homing site (Komori *et al*., 1999a). Intriguingly, PI-*Pko*I and PI-*Pko*II show extreme thermostability (active at 90°C for 1 h) and salt tolerance (0.5 *M* NaCl or KCl for PI-*Pko*I, 1 *M* KCl, and 0.75 *M* NaCl for PI-*Pko*II) (Nishioka *et al*., 1998). I-*Sca*I cleaves its homing site DNA 2- to 2.5-fold faster when provided with supercoiled plasmid when compared with linear DNA, similar to PI-*Sce*I and PI-*Tfu*I (Wende *et al*., 1996; Saves *et al*., 2000a). Interestingly, I-*Sca*I failed to cleave its DNA substrate fully because it is unstable at temperatures >15°C. However, the addition of specific or nonspecific DNA or RNA diminishes inactivation of I-*Sca*I (Monteilhet *et al*., 2000).

Our results have revealed that PI-*Mtu*I differs markedly from the classic view of LAGLIDADG HEase function. PI-*Mtu*I displayed dual target specificity in the presence of alternative cofactors. In the presence of Mn^{2+} and ATP, it was able to cleave homing site in inteinless *recA* allele, and ectopic DNA sites in the presence of Mg^{2+}alone (Guhan and Muniyappa, 2002a, 2002b). The ability of PI-*Mtu*I to cleave ectopic DNA sites raises the possibility that intein sequences are dispersed in natural populations by allowing the HEase to cleave related target site variants of natural homing site DNA. H-*Dre*I is a synthetic chimaeric ENase generated by the fusion of amino-terminal domain of I-*Dmo*I to the I-*Cre*I, with mutations (8 – 12) at domain interface, and a linker peptide between the domains to form an enzyme monomer. H-*Dre*I cleaved target sites dre3 and dre4 (fusion of a pseudo-palindromic half-site from I-*Cre*I and the asymmetric 3' half-site from I-*Dmo*I), but not either of the native target sites. Conversely, wild-type ENases did not cleave the chimaeric target sites. Although the DNA-binding affinity of H-*Dre*I was two orders of magnitude lower than I-*Cre*I, the single catalytic turnover rate was nearly identical (Chevalier *et al*., 2002).

H. GIY-YIG Family

So far, 60 members of the family of GIY-YIG HEases have been identified from eubacteria, archaea, and fungi. These include *td* (thymidylate synthase) and *sun*Y (split gene, unknown function, why?) intron ORFs of bacteriophage T4 (Shub *et al*., 1988; Gott *et al*., 1988; Quirk *et al*., 1989). Most insights into the function of GIY-YIG Family HEases emerge from work with phage T4 introns. Although introns are present in most T-phages, the finding that these are absent in some T-even phages indicates that they are not essential for viability (Shub *et al*., 1988). The *td* (I-*Tev*I) or *sun*Y (I-*Tev*II) intron encoded HEases display relaxed cleavage specificity. The cleavage sites for these HEases are positioned distal to the intron insertion site: 23 nucleotide residues upstream in the case of I-*Tev*I, whereas 15 nucleotide residues downstream in I-*Tev*II. Strand cleavage results in the generation of two (I-*Tev*II) or three (I-T*ev*I) nucleotide 3' overhangs. Cleavage of intron insertion sites by I-*Tev*I or I-T*ev*II is insensitive to extensive sequence divergence flanking cleavage sites (Bryk *et al*., 1995; Mueller *et al*., 1995; Chu *et al*., 1984; Shub *et al*., 1988). The expression of I-*Tev*I catalytic domain was toxic to host cells, indicating that it might bind to and/or cleave genomic DNA in a sequence non-specific manner (VanRoey *et al*., 2002). The intron core sequence is dispensable for I-*Tev*I-dependent homing and insensitive to large insertions at selected sites within the HEase domain. Further experiments have suggested that the natural cleavage site was not absolutely required for I-*Tev*I binding.

Typically, HEases containing GIY-YIG motifs act as monomers, recognize long homing site sequences (~35 bp) and cleave DNA many bases away from the insertion site generating 2 nucleotide 3' overhangs (Mueller *et al*., 1995; Derbyshire *et al*., 1997). Biochemical analyses have shown that I-*Tev*I has a bipartite structure with distinct catalytic and DNA-binding domains connected by a long flexible linker. It binds DNA primarily through the minor groove and makes phosphate backbone contacts. The carboxyl terminal domain appears to be involved in DNA-binding (recognizes a 20-bp sequence including insertion site), whereas the amino-terminal domain interacts with the cleavage site (Van Roey *et al*., 2001, 2002). Mutational analysis suggested that single base substitution within the I-*Tev*I interaction region fail to alter its catalytic activity. Fine mapping of the I-*Tev*I recognition site revealed that it consists of three distinct domains: domain I, implicated in the formation of cleavage proficient complex; domain II confers stability to the enzyme-DNA complex; and domain III involved in DNA-binding. Although the linker sequence in the homing site is not crucial for cleavage *per se*, it facilitates appropriate spacing between domains I and III. *In situ* cleavage assays revealed that I-*Tev*I nicks the bottom strand (>95% efficiency) in the absence of metal-ion, resulting in slow migrating (U_s) distorted complex, whereas both the strands were intact in the fast migrating (U_f) minimally distorted complex (Loizos *et al*., 1996). Interestingly, I-*Tev*II also formed two catalytically active complexes. Unlike I-*Tev*I, both complexes were associated with measurable distortions of the homing site DNA (Mueller *et al*., 1995).

Based on mutational analysis, a 19 amino acid zinc-finger motif has been deduced as the driving force for the cleavage of homing site at a fixed distance (Bryk *et al*., 1995; Derbyshire *et al*., 1997). After deletion of zinc-finger motif, I-*Tev*I cleaves DNA at a displaced site albeit with lower efficiency. Intriguingly, the capacity of I-*Tev*I to “pull back” its catalytic domain is robust, compared with its ability to “stretch out” during the process of severing the displaced cleavage site. Based on these studies, two models were proposed to explain the mechanism of I-*Tev*I catalyzed cleavage. According to “catalytic-clamp” model, the catalytic domain is positioned at the cleavage site via intramolecular interactions with the zinc-finger motif. On the other hand, the “linker-organizer” model posits that interactions between zinc-finger motif and other linker components control folding of the linker, which, in turn, positions the catalytic domain at the cleavage site. A characteristic feature of both the models is that they require interdomain protein-protein interactions (Dean *et al*., 2002; VanRoey *et al*., 2001).

I-*Bmo*I, a member of the GIY-YIG family, binds equally efficiently to substrates containing intron as well as to the intronless allele. However, it can distinguish the sequences during cleavage. Intriguingly, I-*Tev*I and I-*Bmo*I cleave their respective intronless substrates at identical positions, although their introns are inserted 21 bp apart in the intron containing allele (Edgell *et al*., 2001). I-*Bmo*I lacks the zinc-finger, but contained an additional minor groove-binding subdomain as revealed by sequence analysis, suggesting that the GIY-YIG family of HEases have rapidly evolved new mechanisms to maintain their sequence specificities (VanRoey *et al*., 2001).

I. Structural Analysis of GIY-YIG HEases

The crystal structure of I-*Tev*I DNA-binding domain (130–245 residues) complexed with its recognition site (20 bp) revealed an exceptionally large structure with three subdomains: a zinc-finger motif (149–167 residues), an elongated segment containing minor groove-binding helix (183–194 residues), and a helix-turn-helix motif (204–244 residues) (Figure 7) (Van Roey *et al*., 2001). The protein wraps around DNA along its two full turns, so that 50% of both protein and DNA surfaces are solvent inaccessible. Interestingly, the zinc-finger motif makes contact with DNA through two hydrogen bonds. The base-specific contacts involve residues in the unstructured region of the elongated domain (consistent with methylation interference assay — Bryk *et al*., 1993), and this mode of binding distorts the DNA (phosphate–phosphate distances ~4 Å). The helix-turn-helix (H-T-H) domain makes nine hydrogen bond contacts with DNA phosphate backbone (consistent with ethylation interference — Bryk *et al*., 1993), while its hydrophobic interaction with C^5-methyl group of thymines confers specificity. The central elongated region spans the insertion site, and it is the base-specific contact at this interface that distinguishes intronless allele from the intron-containing allele during cleavage (VanRoey *et al*., 2001).

In contrast to the picture for the I-*Tev*I DNA-binding domain, less is known about the structure of its N-terminal domain. To fully understand the mecha-

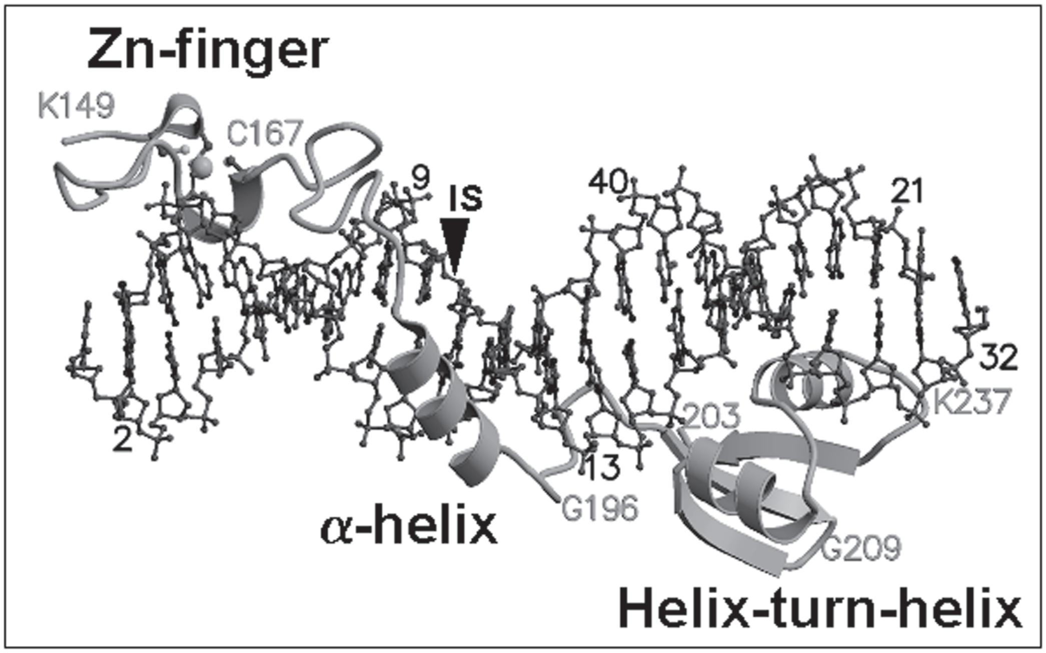

FIGURE 7. Crystal structure of I-*Tev*I DNA binding domain complexed with its *cis*-element. The three domains at the carboxyl-terminus of I-*Tev*I (helix-turn-helix, extended α-helix and the zinc-finger) in complex with DNA is shown in ribbon and stick model, respectively. I-*Tev*I insertion site (IS) is indicated by arrowhead. (Reproduced with permission from VanRoey *et al.*, 2001 and Oxford University Press.)

nistic aspects of DNA-binding and cleavage, a detailed knowledge of co-crystal structure of full-length protein is necessary. Regrettably, no three-dimensional structural information is currently available for the structure of full-length protein. However, an important advance in understanding the structure of full-length protein emerges from the structure of catalytically inactive mutants (R27A and E75A). Using NMR and structure prediction approaches, the three-dimensional structure of I-*Tev*I has been generated. The overall architecture reveals that the catalytic domain (residues 1–92) consists of three α-helices and three-stranded antiparallel β-sheets arranged to form a ββαβα superhelical structure (Kowalski *et al*., 1999). The structural core contains a novel fold with unique helix orientations around the β-strands. Although the helices are loosely packed against the sheet, large side chains in the hydrophobic core of the molecule aids in maintaining the structural integrity (VanRoey *et al*., 2002).

The hallmark of GIY-YIG family of HEases is the presence of GIY-YIG module, and their ability to display promiscuous DNA-binding activity. Normally, the module contains five conserved motifs: motif A (residues 4–19) with GIY-YIG sequence in β-strands 1 and 2; motif B (residues 21–32) in helix 1, motif C (residues 47–62) at the carboxyl-terminal half of helix 2 and all of strand 3, motif D (residues 75–81) in the central part of helix 3, and motif E (residues 85–92) in the carboxyl terminal loop of the domain. Motifs A, B, D, and E contain one invariant residue (Gly^{19}, Arg^{27}, Glu^{75}, and Asn^{90} in I-*Tev*I) in each and two highly conserved residues (Tyr^{6} and Tyr^{17}). Motifs D and B contain the catalytic residues Glu^{75} and Arg^{27}, respectively (VanRoey *et al*., 2002). Interestingly, all of these residues are arranged on one surface of the molecule formed by β2α1α3 and the carboxyl terminal loop. As noted above, GIY-YIG module in motif A plays only a structural role. Similarly, motif C may also play a structural role because Phe^{56} in this motif forms the central hydrophobic core of the molecule. The conserved Tyr residues are proposed to be crucial for either catalysis or substrate binding (VanRoey *et al*., 2002).

The catalytic domain of I-*Tev*I, which is slightly concave, lacks the pronounced groove-shaped DNA-binding clefts. The conserved amino acids at the active site surface, which are implicated in catalysis by computational modeling (Bujnicki *et al*., 2001a), are not within the hydrogen bonding distances to form a single active site. The catalytic residues of I-*Tev*I are similar to the His-Cys box HEase, I-*Ppo*I (Glu^{75} and

Tyr[17] of I-*Tev*I are equivalent of Asn[119] and His[98] of I-*Ppo*I), despite having completely different folds (Galburt *et al*., 1999; VanRoey *et al*., 2002). However, there are differences between the two enzymes, like Arg[27] does not align with Arg[61] of I-*Ppo*I, and Tyr[17] and Glu[75] are not contiguous as in I-*Ppo*I, and their β-strands are oriented in opposite directions (Wu *et al*., 2002). These findings suggest that I-*Tev*I does not contain the ββα-Me structural motif that relate to the His-Cys box and H-N-H HEase. However, mechanistic insights into the catalytic mechanism of GIY-YIG HEases await co-crystal structures of the wild-type catalytic domain with the homing site DNA.

J. The H-N-H Family of Intron-Encoded HEases

The H-N-H family of HEases was first identified in the DNA polymerase genes of *B. subtilis* phages (SP01 and SP82), and independently by Hidden-Markov analysis of bacteriophage SPP1, T3, T7, and T4 genomes (Shub *et al*., 1994). Although the biological significance is unclear, H-N-H proteins are found mostly in phages that infect Gram-positive and Gram-negative bacteria (Crutz-Le Coq *et al*., 2002). The SP01 and SP82 phages contain modified uracil (5-hydroxymethyl) (HMU) instead of thymine. Accordingly, SP01 and SP82 phages are designated as HMU phages and their intron-encoded ENases as I-*Hmu*I and I-*Hmu*II, respectively (Goodrich-Blair *et al*., 1990; Goodrich-Blair and Shub, 1994, 1996). The ENases are neither required for phage propagation nor for their viability and contain only the conserved H-N-H motif. Such a motif was later identified in *nrd*B intron ORF of RB3 bacteriophage (I-*Tev*III), and other HEases such as I-*Cmoe*I, RF253, *yos*Q, *Avi*, *Cpc*, *Pet*D, I-*Two*I, nonspecific ENases (colicins E2, E7, E8, E9, S1, S2), *Lactococcal* bacteriophage bIL170, group II intron ORFs (I-*Sce*V, I-*Sce*VI and I-*Lla*I), and restriction enzymes (McrA). I-*Hmu*I and I-*Hmu*II can distinguish SP01 and SP82 DNA and show preference for heterologous phage DNA. Most importantly, while I-*Hmu*II is required for exclusion of SP01 markers during mixed infection, I-*Hmu*I is not (Goodrich-Blair *et al*., 1996). Furthermore, sequence analysis suggested that the active site residues are conserved in I-*Hmu*I, I-*Hmu*II, and I-*Tev*III ENases. The members of H-N-H family HEases cleave their target sites either as monomers or dimers depending on the substrate. These have been shown to cleave their target sequences either in one (I-*Hmu*I, I-*Hmu*II, I-*Two*I — Landthaler *et al*., 2002) or both the strands (I-*Cmoe*I — Drouin *et al*., 2000; I-*Tev*III — Eddy and Gold, 1991) leading to the generation of 5' overhangs.

Colicins are SOS-induced protein toxins produced by *E. coli* to kill competing *E. coli* strains and closely related bacteria. It was further shown that the mechanism by which colicin exerts its function is via degradation of chromosomal DNA. However, co-expression of a high affinity ($K_d = 10^{-17}$) exo-site inhibitor by the host blocks degradation of host chromosomal DNA (Kleanthous *et al*., 1999; Li *et al*., 1997). The cleavage of negatively supercoiled DNA by colicin E9 requires Mg^{2+} or Ca^{2+}, whereas Ni^{2+} is essential for cleavage of single-stranded or calf thymus DNA. Mutational analysis implicated H-N-H motif of ColE9 in DNA-binding, and residues that coordinate with Mg^{2+} have been identified as crucial for nuclease activity (Walker *et al*., 2002; Ku *et al*., 2002). Comparative genomic and functional analyses suggest that colicins actually do not belong to the conventional H-N-H family of HEases. However, in the absence of structural information for any known H-N-H HEase, it would be informative to consider them in the context of the colicin structure (Ko *et al*., 1999; Sui *et al*., 2002). The relative orientation of the two H-N-H motifs in the dimeric ColE7 nuclease is similar to the two *ββα*-folds of I-*Ppo*I, a member of His-Cys box ENase (see below). The spatial arrangement of ColE7 shows crescent shape with two H-N-H motifs closely located at the dimeric interface. By comparison with the I-*Ppo*I-DNA complex, it was proposed that ColE7 binds DNA through the major groove using α-helices (α2), and the zinc-ions are located close to the scissile phosphate at the central minor groove. The electrostatic surface plot of ColE7 revealed a basic concave surface close to the active sites and hence appropriate for DNA-binding (Cheng *et al*., 2002).

K. His-Cys Box Family of HEases

In slime molds, fungi and amoeba, the group I introns encode His-Cys box family of HEases. These are characterized by the presence of a highly conserved series of histidines and cysteines over a 100 amino acid central region (Johansen *et al*., 1993). The homodimeric I-*Ppo*I, a much-studied member of the His-Cys box HEase, recognizes a 14-bp pseudo-palindromic homing site and inflicts a double-strand break across the minor groove to generate 4 nucleotide 3' overhangs (Muscarella *et al*., 1990; Ellison and Vogt, 1993; Wittmayer and Raines, 1996; Mannino *et al*., 1999). Structural analysis of I-*Ppo*I in complex with its target DNA has revealed three distinct features (Figure 8). First, I-*Ppo*I contains a highly conserved His-Cys box motif that coordinates with two zinc ions to form two zinc-encased folds per monomer. Second, the active sites of I-*Ppo*I monomers separated by 20 Å bend the DNA to widen the minor groove and position the scissile phosphate at its active site. Third, the dimerization of I-*Ppo*I monomers are stabilized by domain-swapped carboxyl terminal tail that wraps around the opposing monomer (Flick *et al*., 1998). Detailed structural analysis show that I-*Ppo*I severely bends (at 50^{o}) and twists its target DNA substrate to pull the scissile phosphates apart. The amino- and carboxyl terminal zinc-binding motif of I-*Ppo*I also exist in I-*Nxx*I (*Naegleria sp*) and I-*Dir*I (*Didymium iridis*). The latter pair has been shown to promote homing in natural hosts (Decatur *et al*., 1995; Vader *et al*., 1999; Elde *et al*., 2000). Despite significant conservation, striking structural differences are likely because (a) I-*Nxx*I yields 5 nucleotide 3' overhang after cleavage, while I-*Ppo*I generates a 4 nucleotide 3' overhang (Elde *et al*., 1999), (b) I-*Ppo*I severely distorts DNA to accommodate the scissile phosphates within the active site (20 Å apart), whereas I-*Nxx*I induces only a minor bend (15 Å apart), and (c) I-*Nxx*I lacks the essential carboxyl terminal tail involved in dimerization (Johansen *et al*., 1993; Flick *et al*., 1998).

To gain mechanistic insights into the formation of double-strand breaks by His-Cys box HEases, I-*Ppo*I-DNA complex was crystallized in the presence of Ca^{2+} and Mg^{2+}(Galburt *et al*. 1999). Structural analysis of the product - I-*Ppo*I complex suggests that

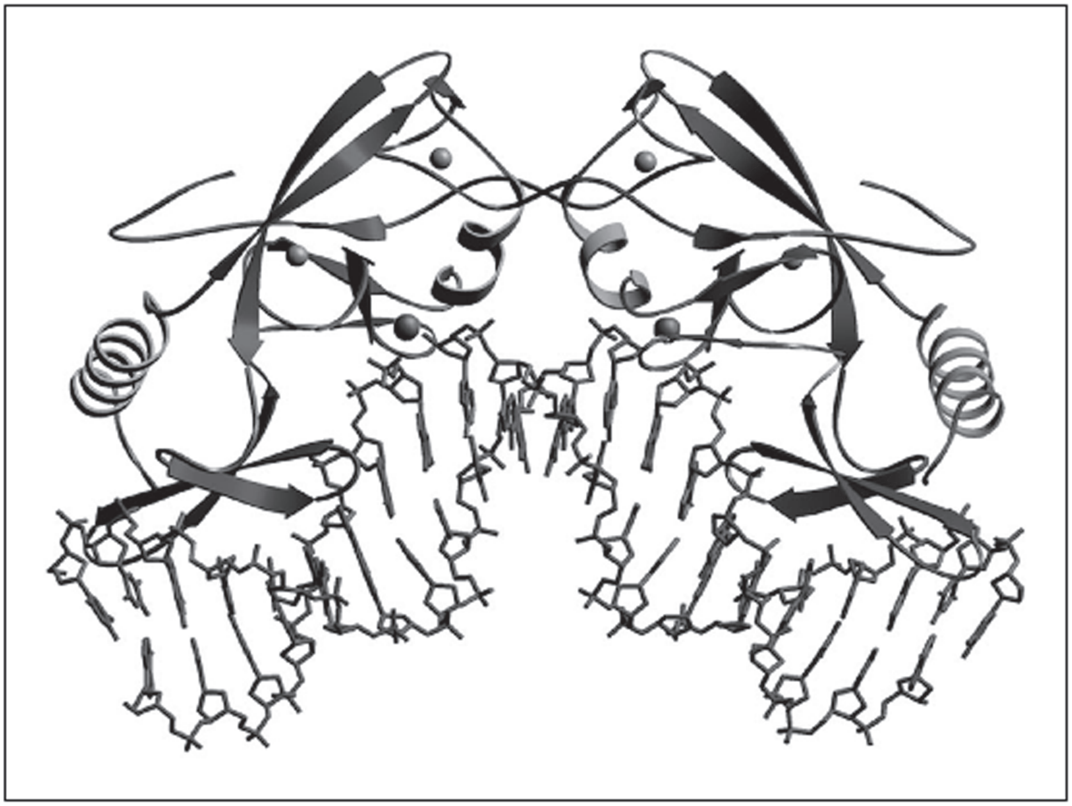

FIGURE 8. Binding of dimeric I-*Ppo*I to the homing site induces distortion in DNA. The three-dimensional structure of each I-*Ppo*I protomer is stabilized by zinc-ions (grey spheres), while the dimeric structure is stabilized by interaction of one monomer with its long carboxyl terminal tail. Dimeric I-*Ppo*I distorts (50^{o}) the B-form homing site DNA, thereby inducing a kink at the cleavage site by opening the minor groove. (Reproduced with permission from Flick *et al*., 1998 and Nature publishing group.)

a conserved Asn, 3' hydroxyl of cleaved DNA, and four water molecules mediate hydrogen bond interactions with the metal ion at the active site. The metal ion is positioned in such a way that it can interact with scissile phosphate but not with the water molecule. Therefore, the metal ion in I-*Ppo*I seem to play three distinct roles during catalysis: stabilization of the phosphoanion intermediate and 3'-hydroxyl leaving group, lessen the pK_a of water molecule and accelerate proton transfer to the 3' hydroxyl leaving group, and geometrically strain DNA and enhance the rate of the reaction (Figure 9) (Mannino *et al*., 1999; Flick *et al*., 1998). Other studies have shown that I-*Ppo*I can tolerate base pair substitutions at several positions within its homing site; however, exhibits strong preference for A:T bp at the central core. In contrast to this, a single point mutation L116A in I-*Ppo*I leads to an inactive conformation, suggesting that Leu^{116} is involved in packing and rotation of the monomers after binding the homing site. However, the position of Leu^{116} in I-*Ppo*I does not allow it to act as a wedge, but is close enough to desolvate the minor groove where adenine is unstacked and stabilized in the complex. Interestingly, Leu^{116} is not a conserved residue (e.g., I-*Nja*I lacks Leu) among the members of HEases, indicating that some have evolved distinct DNA-binding modes and active site geometries to accomplish similar biological function (Galburt *et al*., 2000). The sequence alignment of target sites of representative ENases from eukaryotic, eubacterial and archaeal origin is listed in Table 3.

L. Comparison between HEases and REases

Members of the family of HEases share with REases the ability to inflict site-specific double-strand breaks in their target DNA substrates. Although both require divalent metal ions for their catalysis and utilize similar mechanisms, but show fundamental differences in their biochemical as well as structural properties (Table 4). The differences in sequence-specific DNA-binding displayed by HEases and REases could be gleaned from comparison of their co-crystal structures (reviewed by Pingoud and Jeltsch, 2001; Aggarwal, 1995; Jurica and Stoddard, 1999; Chevalier and Stoddard, 2001b). In general, although HEases display sufficient sequence specificity, they make only limited hydrogen bond contacts in the major groove of the target DNA substrates. In contrast to this, REases are highly sequence specific and saturate nearly all of the hydrogen bond donors and acceptors in the major and minor grooves of their DNA target substrates. Indeed, the inherent differences between these two families of ENases lie at the heart of their *in vivo* roles. Because HEase function has important evolutionary implications, they must function strategically as well as physiologically. To ensure the success of lateral DNA transfer in natural populations, HEases must cleave the host genome once to avoid the accidental cleavage of essential genes within the host genome. This is partly abutted by the requirement of long and often degenerate and asymmetric target DNA substrates. On the other hand, REases must readily identify and eliminate foreign DNA sequences invading the host, thus are able to recognize and cleave short sequences so as to increase the frequency of finding the target DNA substrate. Furthermore, the protection to host genome is conferred by methylation of the restriction site by the corresponding methylase of the R-M system (reviewed by Jurica and Stoddard, 1999; Chevalier and Stoddard, 2001b; Pingoud and Jeltsch, 2001).

Why do HEases engage in self-propagation, while REases are not? Belfort and Roberts (1997) have proposed two alternative possibilities. First, while methylation protects cleavage of host DNA by REases, HEase sites are protected only if the intervening sequence DNA is inserted. Second, unlike most REases, some HEases (I-*Sce*I, I-*Cre*I, F-*Sce*II, I-*Tev*II, and PI-*Tfu*I) tenaciously bind their cleavage products (Mueller *et al*., 1996; Belfort and Roberts, 1997; Saves *et al*., 2000a). Hence, the mobile elements may be inserted as a means to prevent cleavage at ectopic sites (transposition), which, in turn, is likely to confer stability to the genome. The reason for this argument is that, if the homing site is undergoing degeneration (as in the case of *S. cerevisiae* DH1-1A; Gimble, 2001), the HEases could tolerate only to certain extent after which it would enable a mechanism to find alternate (ectopic) site in the genome. However, these speculations have yet to be tested by experimentation.

V. CATALYTIC MECHANISMS OF HOMING ENDONUCLEASES

A. Mechanism of Phosphodiester Bond Hydrolysis by ENases

Under physiological pH, the thermodynamically stable phosphodiester bonds present kinetic barriers

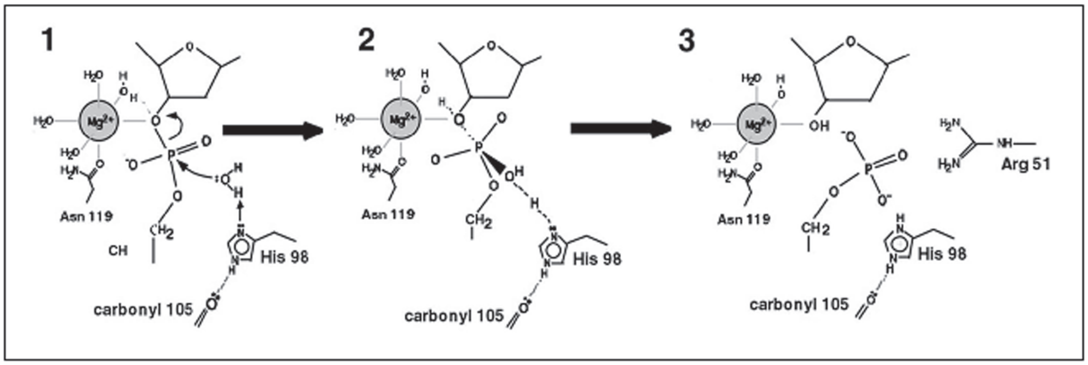

FIGURE 9. A postulated role for Mg^{2+} in the hydrolysis of phosphodiester bond catalyzed by I-*Ppo*I. Mg^{2+} exists in octahedral geometry in I-*Ppo*I and participates in transition state stabilization and protonation of 3' oxygen leaving group. Imidazole ring of 'His' acts as a general base because of its H-bond interaction with backbone carbonyl oxygen and activates the hydrolytic water molecule to perform an in-line attack. (Reproduced with permission from Galburt *et al.*, 1999 and Nature publishing group.)

for hydrolysis. The kinetic barrier arises as a result of electrostatic repulsion between the attacking nucleophile and the negatively charged phosphate group. However, enzymes overcome this partly by using metal ions as cofactors. In addition, nucleases require a general base to activate the nucleophile, a general acid (Bronsted acid) to stabilize the leaving group and a Lewis acid to stabilize the negative charge that develops on phosphorus atom during transition state. In principle, metal ions satisfy most of these requirements in nucleases. By acting as Lewis acid, metal ions lower the pK_a of water molecules bound to them and generate hydroxide ion, which by itself serve as a nucleophile or a general base. Alternatively, the metal ion also aids in the transfer of proton from the bound water molecule. In most instances, metal ions also neutralize the additional negative charge that develops during the transition state.

Direct transfer of phosphoryl group to water molecule during reactions catalyzed by ENases can occur by alternate mechanisms depending on the orientation of the attacking nucleophile with respect to the leaving group, and the degree of bond formation along the reaction coordinates. Regardless of the mechanism, the transition state geometry is trigonal bipyramidal with three equatorial and two axial oxygen atoms. Most importantly, the oxygen atoms that enter or leave the transition state can only be axial as the other angles are sterically hindered. Depending on the order of bond breakage or formation, reaction mechanisms can be classified into **S_N2** (associative) and **S_N1** (dissociative) mechanism. In S_N2 reaction, the nucleophile is associated with the phosphorus atom before dissociation of the leaving group, whereas in S_N1 mechanism, the leaving group dissociates before the association of nucleophile. Consequently, the nucleophile attacks the unstable metaphosphate intermediate in S_N1 mechanism.

In S_N2 reactions, if the attacking nucleophile and the leaving group are in axial positions, the mechanism is defined as '**in-line**'. Alternatively, if their orientation changes during the reaction due to pseudorotation, it is referred to as '**adjacent**' mechanism. However, both of them result in inversion of configuration. Unlike S_N2, S_N1 reactions do not depend on a general base to generate the nucleophile. Rather, catalysis is mediated by the stabilization of the leaving group and transition state metaphosphate. Most phosphoryl group transfer enzymes have been shown to follow S_N2 mediated 'in-line' mechanism during catalysis (Galburt and Stoddard, 2002).

B. Choice of Divalent Metal-Ion for Catalysis

HEases can utilize a variety of divalent metal ions, including Mg^{2+}, Mn^{2+}, Ca^{2+}, Co^{2+}, and Zn^{2+}. Most HEases prefer oxophilic Mg^{2+} as the cofactor for optimal catalysis due to its favorable physicochemical properties. The choice is possibly due to its intracellular availability (free concentration is ~0.5 m*M*), redox inertness, small ionic radius, high charge density, high transport number, and reduced solvent exchange rates (reviewed by Cowan, 1998). The ten-

TABLE 3
Sequence Alignment of Cleavage Sites of Homing Endonuclease
The vertical arrows in the sequence denote intein or intron insertion sites. Symbols "diamond" and "underscore" contiguous with the nucleotide sequence indicate the position of cleavage on the upper and lower strands, respectively. "E-", "I-" and "PI-" denotes engineered hybrid, intron- and intein-encoded ENases. PI-*Mtu*I* indicates cleavage site sequence of *M. tuberculosis pps*1 intein.

H-*Dre*I	CGCGCCGGAAC_TTAC♦GACGTTTTG
I-*Bmo*I	AGAGC_CCG♦T↓AGTAATGACATGGCCTTGGGAAAT
I-*Cre*I	GCTGGGTTCAAAACGT↓C_GTGA♦GACAGTTTGGT
I-*Dmo*I	AATGCCTTGCCGG_GTA↓A♦GTTCCGGCGCGCATG
I-*Ppo*I	GTAACTATGACTCTC_T↓TAA♦GGTAGCCAAATGC
I-*Sca*I	ATTGTCACATTGAGGT↓GCACTAGTTATTACTA
I-*Sce*I	AAGTTACGCTAGGG_AT↓AA♦CAGGGTAATATAGC
I-*Tev*I	CA_AC♦GCTCAGTAGATGTTTTCTTGGGT↓CTACCGTTTAATATTG
I-*Tev*II	TTCCAAGCTTATGAGT↓ATGAAGGTGAACAC_GT♦TATTC
I-*Tev*III	T♦TA_TGTATCTTTTGCGT↓GTACCTTTAACTTCCA
PI-*Mga*I	CGTAGCTGCCCA_GTAT♦GAG↓TCAGAGGTGG
PI-*Psp*I	CAAAATCCTGGCAAAC↓AGCTATTATGGGTATT
PI-*Sce*I	TATCTATGTCGG_GTGC♦↓GGAGAAAGAGGTAATG
PI-*Tli*I	GGTTCTTTATGCGG_AC↓AC♦TGACGGCTTTTATG
PI-*Tli*III	TAAATTGCTTGCAAAC↓ AGCTATTACGGCTATA
PI-*Pko*I	TAGATTTT_AGAT♦CCCTGTACCCC
PI-*Pko*II	AACAGCTA_CTAC♦GGTTACTA
PI-*Pfu*I	TACAGAAGATGGGAGGA_GGG↓A♦CCGGACTCAACTTCTCAAA
PI-*Pfu*II	CGATAAGGGCAACGAATCCA↓TGTG_GAGA♦AGAGCCTCTATA
PI-*Tfu*I	CTTATTTAGATTTT_AGG↓T♦CGCTATATCCTTCGATT
PI-*Tfu*II	AAAGTGCTGTACGCGG_AT↓AC♦AGACGGCTTTTTCGCAAC
PI-*Pab*I	GGGGGCAGC_CAGT♦↓GGTCCCGTTTCG
PI-*Pab*II	ACCCC↓TGTG_GAGA♦GGAGCCC
PI-*Mtu*I*	ACGTGCACTACGTAGAGGGC↓ TGCACCGCACCGATCTACAA
PI-*Mtu*I	AA_CGCGGTCGGCAACCGCACC♦CGGGTCACGGTCGTCAAGAACAAG↓TGT

TABLE 4
Comparison of Structural and Biochemical Properties of HEases with REases

Properties	HEases	REases
Nature	Genomic parasites.	Selfish-symbionts that provide protection to the host genome (Naito *et al.*, 1995).
Genomic locations	Located in introns (group I, group II and Archaeal), protein introns, and intergenic regions of nuclear, mitochondrial and chloroplast genomes of all the three biological kingdoms.	Present as freestanding ORFs in the genome of archaea, eubacteria and some eukaryotic viruses, along with modification (methylase) gene.
Conserved domains	Classified into four families on the basis of conserved motifs: (a) LAGLIDADG, (b) GIY-YIG, (c) H-N-H and (d) His-Cys box.	Although PDX_{8-25}(E/D)XK motif is intrinsic to several restriction ENases, it is not a defined conserved motif.
Target sites	Recognize asymmetric, long (12-40 bp) homing site sequences in a sequence tolerant manner.	Highly sequence specific, recognize short (3-8 bp) sequences that are either symmetric or asymmetric.
Accessory molecules	Some require RNA component for activity (as in group II introns).	Some require modification and specificity subunits along with restriction subunit.
Invasiveness	Highly invasive and display super-Mendelian inheritance.	Non-invasive
Oligomeric status	Act as monomers or homodimers.	Act as monomers, dimers and sometimes as tetramers.

dency of Mg^{2+} to maintain outer hydration sphere and its ability to use the bonded water molecules for catalysis led to the choice of this metal ion by the enzymes of nucleic acid metabolism. Mg^{2+} follows the outer-sphere pathway, where it stabilizes the transition state either electrostatically, and/or via hydrogen bonding network with water (Cowan, 2002). Thiophilic Mn^{2+} also shows chemical properties somewhat similar to Mg^{2+} and has been shown to serve as a cofactor for few HEases. However, Mn^{2+} presents relaxed specificity for some REases and HEases, resulting in increased specific activity by compromising on fidelity. The differences in fidelity due to Mg^{2+} or Mn^{2+} could be explained by the disparity in their coordination sphere, which is involved in providing substrate specificity. Some HEases can maintain their activity when Mg^{2+} or Mn^{2+} is substituted by other divalent metal ions. Although Ca^{2+} can substitute Mg^{2+} or Mn^{2+} in binding of some HEases to their target DNA substrate, but fail to abet in cleavage function. Interestingly, I-*Nja*I, I-*Nan*I, and I-*Nit*I were able to cleave target DNA substrate inefficiently in the presence of Ca^{2+}, a feature distinct from the LAGLIDADG HEases (Elde *et al*., 1999, 2000). However, in contrast toºI-*Ppo*I, Zn^{2+} does not support catalysis of I-*Nja*I, I-*Nan*I, and I-*Nit*I.

C. Common Features of HEases and REases

Metal ions can be coordinated by enzymes through their carboxylate side chains of acidic amino acids, main chain carbonyl group, or by the polar amino acids. These residues are highly conserved among REases and HEases. Type II REases use their carboxylate side chains of the two acidic residues and the main chain carbonyl group of the X residue (PD…D/EXK) in binding the metal ion. For instance, LAGLIDADG HEases utilize two aspartate residues in coordinating the metal ion. Similarly, transposases and integrases make use of the DDE motif in coordinating the metal ion. The stoichiometry of metal-ions required by these enzymes for the display of optimal catalysis has been one of the most fascinating but poorly understood aspects: the numbers obtained from solution and structural analyses are not in accord in many instances. Regardless of the number of metal ions and the precise mechanism of phosphodiester bond hydrolysis, enzymes follow the simple acid-base catalysis as described earlier. For this type of catalysis, proton relay from solvent network and also from nearby chemical entities are important. Based on the number of metal ions involved in catalysis, enzymes follow one of the two mechanisms discussed below.

D. Single-Metal Mechanism

ENases that utilize the single-metal mechanism bind metal ion via their conserved acidic amino acids. The metal-ion-coordinated water molecule is positioned for in-line attack on the electrophilic phosphorous atom of the scissile phosphate (e.g., *Bgl* II, I-*Ppo*I). In addition, a second water molecule bound to the metal ion is also positioned to donate a proton for stabilizing the leaving group (e.g., *Eco*RI). Most REases that follow single-metal mechanism lack the general base that is required to abstract the proton from nucleophilic water molecule in an associative nucleophilic attack. However, they overcome this deficiency by substrate-assisted catalysis, or bulk solvent-supplied metal bound hydroxide ion, or a dissociative transition state. In substrate-assisted catalysis, the nonbridging oxygen atom of the phosphate acts as a general base. Having the general base oxygen on the 3' phosphate adjacent to the scissile phosphate provides added advantage by neutralizing the charge between phosphate and nucleophilic water after protonating the phosphate oxygen.

A member of the His-Cys box family of HEase, I-*Ppo*I, utilizes the single metal-ion mechanism for cleavage of a phosphodiester bond. Three different crystal structures of I-*Ppo*I showed that the enzyme bound metal-ion functions as Lewis acid in stabilizing the pentavalent transition state as well as induces bound water molecule to donate a proton to the leaving group (Flick *et al*., 1997; Galburt *et al*., 1999, 2000). Unlike most REases, I-*Ppo*I contains a general base (His^{98}) to activate the attacking water molecule. Interestingly, a nonspecific nuclease from *Serratia* and *Anabaena* share similar active sites, indicating the possibility that they might use single metal-ion mechanism during the cleavage of phosphodiester bond (Friedhoff *et al*., 1999).

E. Two-Metal Mechanism

ENases that follow a two metal-ion mechanism, the first metal-ion (Me1) is located at an identical position and plays a similar role as in the case of enzymes that follow single metal-ion mechanism. The second metal ion (Me2) is usually located on the other (rear) side of the scissile phosphate and is involved in protonating the leaving group using the metal-bound water molecule. It is also involved in stabilizing the negative charge that develops during the associative transition state. For enzymes that follow two-metal mechanism, the prerequisite is that the metal ions should be located in close proximity (~4 Å) and should be bridged by the substrate in order to display coherent cleavage of the duplex DNA strands (Cowan, 1998). Similar to single-metal mechanism, enzymes that follow two-metal mechanism also encounter the state of nonavailability of a general base for catalysis, and hence might follow substrate-assisted catalysis. Both I-*Cre*I and PI-*Sce*I follow two-metal mechanism, although they contain a different number of metal ions (Moure *et al.*, 2002).

Structural studies on I-*Cre*I revealed that it follows $1^1/_2$ metal ion mechanism in cleaving one phosphodiester bond (Chevalier *et al.*, 2001a). A unique feature of I-*Cre*I is that the second metal ion is shared between the two active sites (Figure 10). Overall, three metal ions are shared between two active sites in a small space. In substrate bound crystal structure, the metal-bound water is well positioned for nucleophilic attack, and there is no direct protein-nucleophilic water molecule contact. Rather, a well-ordered water network link the nucleophilic water to the leaving group or to the bulk solvent. In this regard, there is no difference between the substrate and product bound I-*Cre*I complex, except for the movement of the released free phosphate. This raises the interesting possibility of concerted transfer of hydrogen atoms via the water network to activate the nucleophile and protonate the leaving group. In water network as well as substrate-assisted catalysis, residues near the active sites stabilize the transition state. For example, three moderately conserved residues (Lys^{98}, Arg^{51}, Gln^{47}) in I-*Cre*I have been shown to be important for catalytic activity. However, other residues can also extend in and fill the pocket. Furthermore, diversity in the LAGLIDADG motif suggests that there are many possible ways by which water molecules can be positioned at the active site (Chevalier *et al.*, 2001a).

A critical issue in single- or two-metal mechanism is the pK_a of the general base. The pK_a of phosphodiester bond is approximately <2 and that of Mg^{2+}-water complex is between 11 and 12. Given the differences in the pK_a values, how then the enzyme catalyzes the cleavage reaction? There could be several different ways by which enzymes overcome the apparent differences in pK_a values. First, it may not need a general base if the transition state has a dissociative nature. Second, the nucleophile can be generated from the bulk solvent at the catalytic site. Alternatively, amino acids in the catalytic center might lower the pK_a of Mg^{2+} bound water molecule to generate the nucleophile.

VI. BIOLOGICAL AND EVOLUTIONARY ROLES OF HOMING ENDONUCLEASES

A. *In vivo* Assays to Evaluate HEase Activity

Characterization of HEase function involves mutational analyses of the target DNA substrate as well as the enzyme. It requires cloning, overexpression, and purification of variant enzymes; however, one limiting factor is that some of them induce host cell death (e.g., I-*Tev*I). Therefore, assays have been developed to circumvent these problems and at the same time enable one to monitor HEase activity *in vivo*. A rapid and sensitive selection system involves cotransformation of a plasmid containing toxic gene (Barnase — RNase from *Bacillus amyloliquefaciens*) containing the cleavage site for HEase under scrutiny and a second compatible plasmid with HEase gene (wild-type or variant) (Gruen *et al.*, 2002). In the absence of HEase, toxic gene expression is lethal to the host cells. However, after HEase expression, the toxic gene containing plasmid will be cleaved by HEase and rapidly degraded by the endogenous RecBCD nuclease, thereby allowing the cells to survive.

An alternate sensitive genetic assay has been developed in *E. coli* that involves ectopic expression of HEase to cleave the homing site within the kanamycin resistance gene (**F'*Cre*-Kan** assay) or in the *lac*O region of *lac* operon (**F'O-*Cre*** assay), resulting in either kanamycin sensitive or β-lactamase negative phenotype, respectively (Seligman *et al.*, 2002). The assay system was tested using the I-*Cre*I HEase. *In vitro*, I-*Cre*I was able to cleave variant homing sites containing up to 10-bp substitutions with equal efficiency, thereby implying that these are nonessential contacts (Durrenberger and Rochaix, 1993). However, in these genetic assays substitution of 1 or 2 bp at the homing site was resistant to cleavage by I-*Cre*I.

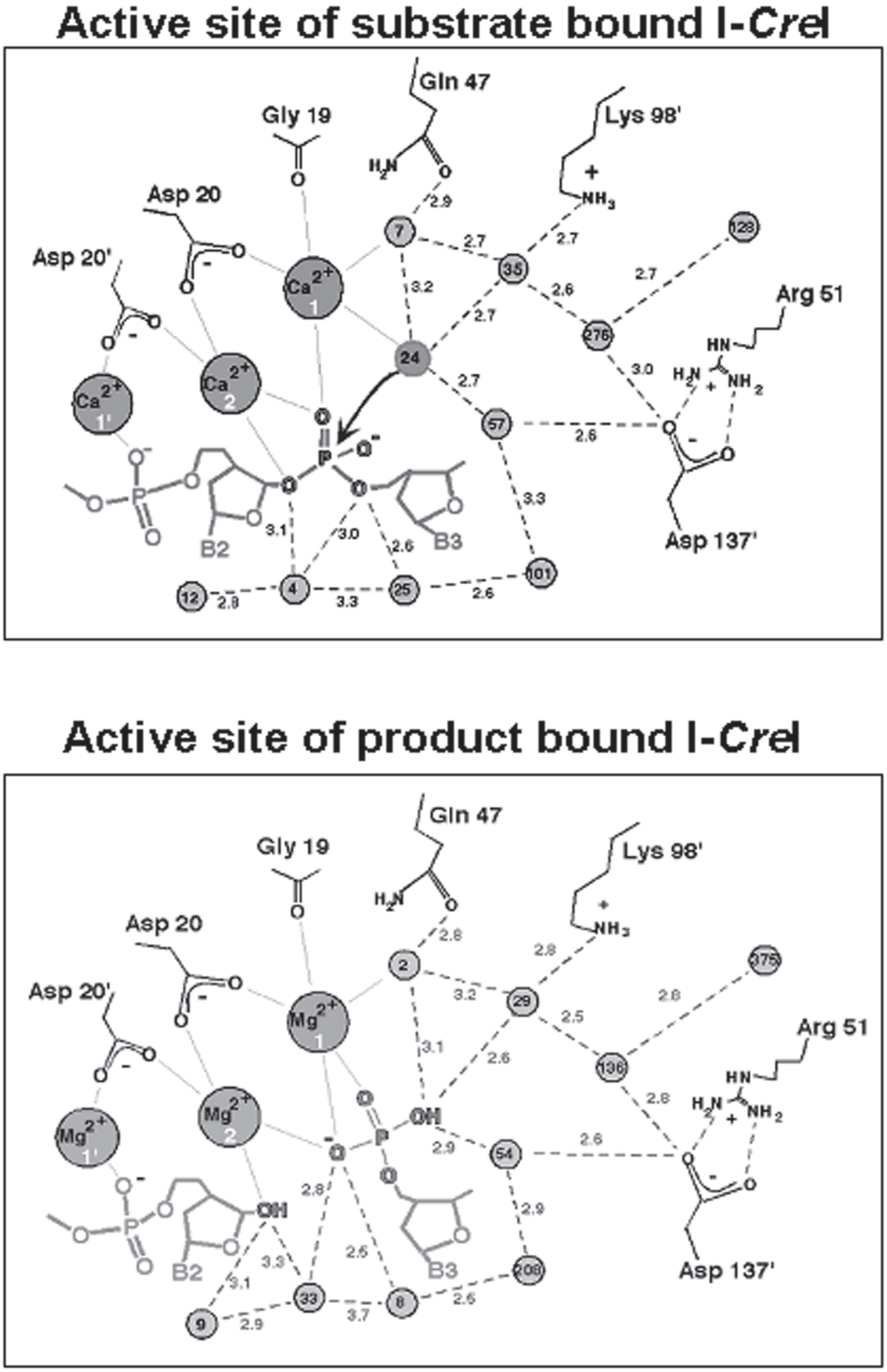

FIGURE 10. Active site of dimeric I-*Cre*I with its solvent network in the presence of substrate or product. Amino acid residues from each monomer in the dimeric I-*Cre*I is distinguished by normal or primed as Asp20 or Asp20′. The metal-ions calcium and magnesium are indicated and water molecules are represented as grey spheres. The dotted lines correspond to hydrogen bonds and the bonding distances are in Angstroms. The water molecules are numbered in accord with the corresponding PDB files. All the direct interactions between the side chains of amino acid residues, metal-ions, and water molecules are represented as thin lines, and contacts with the opposite DNA strand are omitted for clarity. In the product bound complex, the cleaved 5' phosphate had moved away. (Reproduced with permission from Chevalier *et al.*, 2001a and Nature publishing group.)

The power of F'O-*Cre* assay is that it is capable of identifying gain or loss of function mutants in a single step and could also create mutants with intermediate phenotypes by forming sectored colonies (due to cleavage of some sites in initially transformed cells followed by subsequent segregation and cleavage).

B. Evolution of HEases

While the evolutionary origin and function of introns is the subject of ongoing debate, it is believed that inteins have evolved from a state of molecular parasitism to commensalism to mutualism (Hickey, 1994). The comparison of integration sites among inteins unveiled 48 different positions in 34 different host proteins, suggesting that (1) inteins located at identical positions within extein homologs are very closely related and hence considered as "alleles", (2) multiple inteins in the same gene differ more from each other than inteins from different genes, and (3) diverse nature of inteins other than their allelic counterparts hinder phylogenetic analysis (Perler *et al*., 1997a, 1997b).

Similarity between allelic inteins could be because of vertical transmission of intein-coding genes during speciation or due to horizontal transmission. The latter could be invoked only if: (1) sequence similarity between allelic inteins is greater than their host protein sequences (e.g., *R. marinus* DnaB intein shows 54% sequence identity [74% sequence similarity] to *Synechocystis* sp. strain PCC6803 DnaB intein, a value markedly higher than extein sequence identity [37%]), and (2) when codon usage and G+C content of the intein alleles are different from their host proteins (Liu and Hu, 1997). Selfish genes with no useful function for the host are susceptible to degeneration unless they are fixed in a population, and regular horizontal transmission may be the only means of long-term persistence. Frequent horizontal transmissions are possible for mitochondrial HEases (some leakage of mitochondrial DNA during preliminary phase of interspecific matings), whereas nuclear horizontal transmissions are difficult because unlike transposable elements there is no extrachromosomal phase in the propagation of HEase genes. Self-reliance on horizontal transmission alone might plausibly explain the absence of HEase genes in animals (except in mitochondria of sea anemone, *Metridium senile*) because access to germline is more difficult. The route of horizontal transmission in yeasts is unclear because infectious viruses and plasmid vectors are absent. However, possibilities include interspecific hybridization, vectoring by predacious yeasts, and uptake of naked DNA from the environment (Koufopanou *et al*., 2002).

Most inteins are identified in DNA and RNA metabolizing enzymes. Liu (2000) suggested that there is a strong bias for inteins (70%) toward nucleic acid-metabolizing enzymes, and this is because (1) proteins that are involved in nucleic acid metabolism are present in viral and phage genomes and hence might help in the dispersal of inteins; (2) integration in them would ensure that inteins are expressed during replication and repair of DNA, an appropriate time for invasion; and (3) expression during replication and repair would reduce risk for the cell by allowing efficient repair of nonspecific endonuclease activity. Pietrokovski (2001) has argued that these are reasonable explanations; however, it does not address the question of why inteins are found in other types of proteins. From his perspective there is no bias for inteins toward their target proteins. Since inteins are located in conserved regions of host proteins, it is difficult to precisely eliminate them without affecting the host protein activity (Koufopanou *et al*., 2002). Hence, it reflects only the difficulty in eliminating inteins rather than some putative function attributed to them. There are several reasons why the host organism tolerates intein sequences. They are (1) inteins are retained at conserved sites on host proteins and hence cause negligible effect on its function (Gimble, 2000), (2) they might benefit hosts by regulating host protein splicing (e.g., *Synechocystis sp*. PCC6803 split DnaE intein; Wu *et al*., 1998a, 1998b). This latter point suggests that probably inteins regulate the genes in which they reside and might offer selective advantage to them in the organism under different environmental conditions. They may either facilitate their entry into the host genome or they are maintained because deletion of intein insertion site in the host gene would affect host protein function.

Several lines of evidence argue that primeval inteins had only the protein-splicing domain. First, ENase and DNA-binding domains are not essential for protein splicing (Telenti *et al*., 1997; Derbyshire *et al*., 1997; Chong and Xu, 1997; Shingledecker *et al*., 1998). Second, the presence of different classes of ENases with different specificities suggests that the ENase domains were acquired independently in more than one way. The basis for this proposal is that the ENase domains similar to inteins are also present in group I and group II intron encoded ORFs. Finally, the protein-splicing domain bears similarity to the C-terminal auto-processing domain of hedgehog proteins (Porter *et al*., 1996; Hall *et al*., 1997; Klabunde *et al*., 1998; Perler, 1998; Pietrokovski, 1998a). In addition, the natural occurrence of mini-inteins led to the speculation that mobile elements are composite genes arising from the invasion of an ENase ORF into a preexisting gene harboring a protein-splicing element. This view is further supported by phylogenetic analysis of all the identified intein sequences, where mini-inteins do not form a coherent group and inteins without ENase domains were found in several different alleles (Figure 11). The occurrence of an auto-processing domain in hedgehog proteins of invertebrates, vertebrates, and also in nematodes suggests that inteins were present in the last common ancestor of bacteria, archaea, eukaryotes, and possibly before the appearance of metazoans. Also, they might have been involved in domain shuffling at the protein level and/or by recombination at the DNA level (Perler, 1999). It is unlikely that proteins with such a complex biochemical function would have evolved by chance without a positive selection. There must have existed a beneficial role for them in the primeval inteins or their unknown progenitors that are absent now. The extinction of inteins could be due to the loss of their beneficial role or emergence of host-specific defensive processes to expel them from the genome, or perhaps both (Wu *et al*., 1998a, 1998b).

Some phage genes might be acting as sinks for these mobile sequences because they reside in proteins, which have homologs in bacteria and might act as vehicles for their dispersal (Derbyshire and Belfort, 1998). For example, intein sequence identified in the ribonucleotide reductase gene of *Chiloiridescent* double-stranded DNA virus might disperse intein sequence when they infect invertebrates, amphibia, and fish (Pietrokovski, 1998b). Similarly, the presence of inteins in human pathogens (*M. tuberculosis, M. leprae, C. tropicalis*) might also provide an opportunity for inteins to invade metazoan genomes, although it has not been identified to date.

C. Examples for Evolution of HEases

The most interesting question, which is poorly understood, is how an RNA splicing activity intrinsic to several intron-encoded HEases evolved into a DNA-binding enzyme? Since the LAGLIDADG HEase and maturases have analogous structures, it is possible that residues present in the saddle-shaped DNA-binding surface could also interact with RNA (Gimble, 2000). An interesting example is the tRNA splicing endoribonuclease (EndA) of *Methanococcus jannaschii,* which contains RNaseA-like active site similar to PD…(D/E)XK family in the C-terminal domain. The N-terminal domain of EndA is similar to LAGLIDADG HEase, suggesting the fusion of two distinct ENase activities during evolution (Bujnicki and Rychlewski, 2001b). Intriguingly, HEases are highly invasive and are not biased toward any particular host protein. As such, there are instances where they have invaded another intron (twintrons). For example, I-*Pcu*I, a mitochondrial group I intron (*cyt*b-i2) encoded GIY-YIG enzyme of *Podospora curvicola* had invaded the mobile LAGLIDADG intron thereby abolishing its activity, and hence mobility. This is the first example of a HEase that invades another active HEase emphasizing the evolutionary independence of HEase genes from intervening sequences (Saguez *et al*., 2000). *M. xenopii gyr*A intein is an instructive example to illustrate how inteins had undergone a complex series of evolutionary events such as serial deletions (HEase domain (block C, D, E, and H) and insertions (24 amino acid linker). The insertion of a spacer represents a structural requirement, as its deletion results in failure of its splicing activity (Telenti *et al*., 1997).

Desulfurococcus mobilis group I intron provides yet another interesting example because two site-specific LAGLIDADG HEase are encoded from a single ORF in the same reading frame with similar biochemical properties (Lykke-Andersen *et al*., 1996; Agaard *et al*., 1997). The most abundant form, I-*Dmo*I$_{c}$ (circular) contains six additional amino acid residues (RAGGYT) at the carboxyl end (Dalgaard *et al*., 1994; Silva *et al*., 1999) when compared with its linear counterpart I-*Dmo*II. The comparison of I-DmoI with I-*Cre*I and PI-*Sce*I revealed that the amino-terminal domain (domain A) of I-*Dmo*I was homologous to carboxyl terminus of PI-*Sce*I, and the C-terminal domain of I-*Dmo*I was similar to N-terminal domain of PI-*Sce*I. This apparent reversal in the order of domains might have implications in the evolution of two domain LAGLIDADG ENases (Silva *et al*., 1999). The relevance of expression of a circular form of the intron ENase is still unknown. Further, H-N-H enzymes display limited active site similarity to

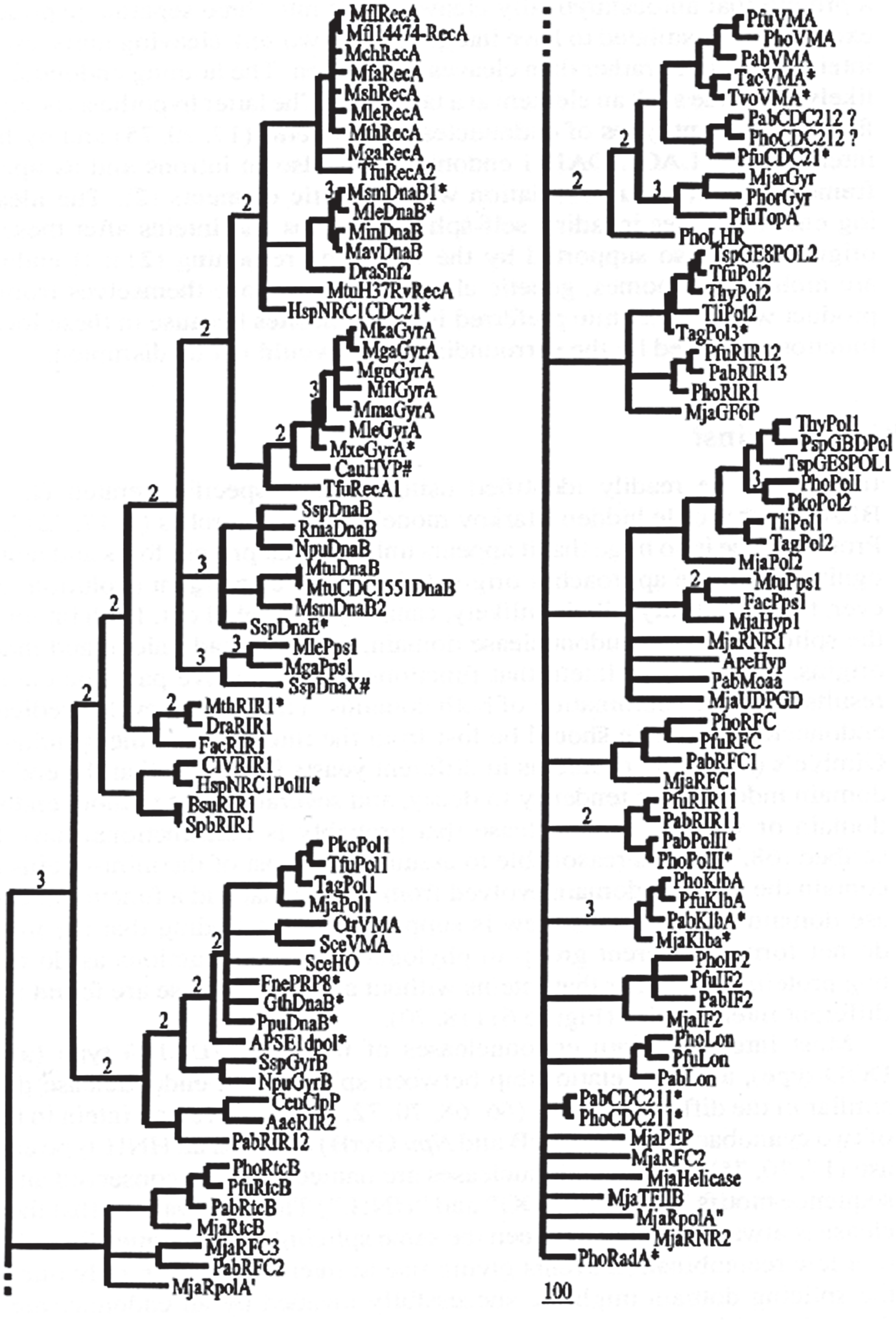

FIGURE 11. A phylogenetic tree of 128 known inteins and HO endonuclease of *S. cerevisiae*. Inteins are named according to the current nomenclature, and their amino acid sequence were aligned using SAM by Gogarten *et al.* (2002). Subsequently, the phylogenetic tree was constructed using parsimony as instigated in PAUP Ver.4.0 beta8, and the numbers denote branches with Bremer decay indices smaller than 4. Branches without labels do not decay even after three additional steps. 'Asterisk' indicates inteins without endonuclease domains, 'hash' indicates inteins with endonuclease domain in an alternate reading frame, and 'question' mark denotes inteins, whose endonuclease domain is questionable. (Reproduced with permission from Gogarten *et al.*, 2002 and *Annual Reviews of Microbiology,* Volume 56 ©2002 by *Annual Reviews* www.annualreviews.org.)

eukaryotic CAD (Caspase Activated DNase) enzymes that are involved in chromatin degradation during apoptosis. In addition, CAD enzymes display biochemical properties analogous to colicin E9. Accordingly, it has been speculated that both CAD and H-N-H enzymes might function through a similar mechanism, where in one case the activity is used against the competing cells, and in the other for self-destruction (Walker *et al*., 2002). Zinc-finger motif in GIY-YIG family of HEases would have arisen by a domain fusion event that was futile, but might have evolved as a distance determinant. The zinc-finger motif not only promotes dissemination of host intron by enhancing the activity at its natural cleavage site, but also by promoting cleavage at a favored distance in the absence of cognate site (Dean *et al*., 2002).

D. Life-Cycle of HEases

HEases are nonessential genes at least in eukaryotes and display super-Mendelian inheritance. The most common state seems to be nonfunctional (*P. abyssi* Lon and RIR1-1 [Saves *et al*., 2002b], *C. tropicalis* VMA1 [Gimble, 2001]) followed by the next common no element state (*S. paradoxus, S. castellii, K. lactis*) to the least common putative functional element state (*S. cerevisiae VMA1*, *K. thermotolerans*). Cyclical evolutionary model proposed by Goddard and Burt (1999) suggests that there are three characteristic states (functional, nonfunctional, and no element), and only three of the six potential transitions are evolutionarily possible: empty→functional (horizontal transmission) → nonfunctional (degeneration by mutation pressure and/or costs of enzyme production) → empty (Figure 12). The final step requires precise excision of the element to reconstitute the recognition sequence. The cycle may reinitiate by another horizontal transfer, and such frequent transfers may allow them to persist over evolutionary time scales. The results of Goddard *et al*. (2001) support how a host mating system could play a key role in determining population dynamics of a selfish gene. The determination of the driving force for HEase transmission and selection would reveal whether HEases are spreading or becoming extinct as well as their habitat and also their functions (Pietrokovskii, 2001).

E. Beneficial Effects of HEase

Some of the beneficial roles that have been identified for HEases are indicated below.

1. *S. cerevisiae* haploid cells use HO endonuclease (F-*Sce*II) to initiate mating-type switching through programmed DNA rearrangements. In the absence of HO endonuclease, mating between two haploid cells is not possible, and therefore diploid cells will not be generated. However, HO gene, unlike HEases, does not propagate itself, rather it gets inherited to the progeny.
2. Some of the $\beta\beta\alpha$-Me family members in bacteria act as colicins and help their hosts under competing growth conditions.
3. *B. subtilis* phage SP82 cleaves SP01 phage genome during mixed infections to promote SP82 propagation (Goodrich-Blair and Shub, 1996). This phenomenon of phage-exclusion is analogous to that of the R-M system that the eubacteria enforce to avoid phage infection.

However, whether all HEases assist their host cells require additional investigations.

F. Role of HEase in Senescence

The instability of mitochondria in *Podospora anserina* has been shown to be responsible for the degenerative premature cell-death syndrome or senescence. A covalently closed circular DNA (αsenDNA) has been shown to accumulate in the mitochondria of senescent cultures and later identified it to be the first intron of the *COX1* gene. Intron transposition into ectopic sites followed by recombination leads to genome rearrangements (site-specific deletion) in this organism. Subsequently, PCR analysis indicated that the nuclear genes controlling premature death syndrome might be activated by one of the steps involved in deletion or positive selection for this molecule (Sellem *et al*., 1993; reviewed by Osiewacz, 2002).

G. Applications of HEases

With the emergence of genome mapping and gene therapy, considerable interest has been

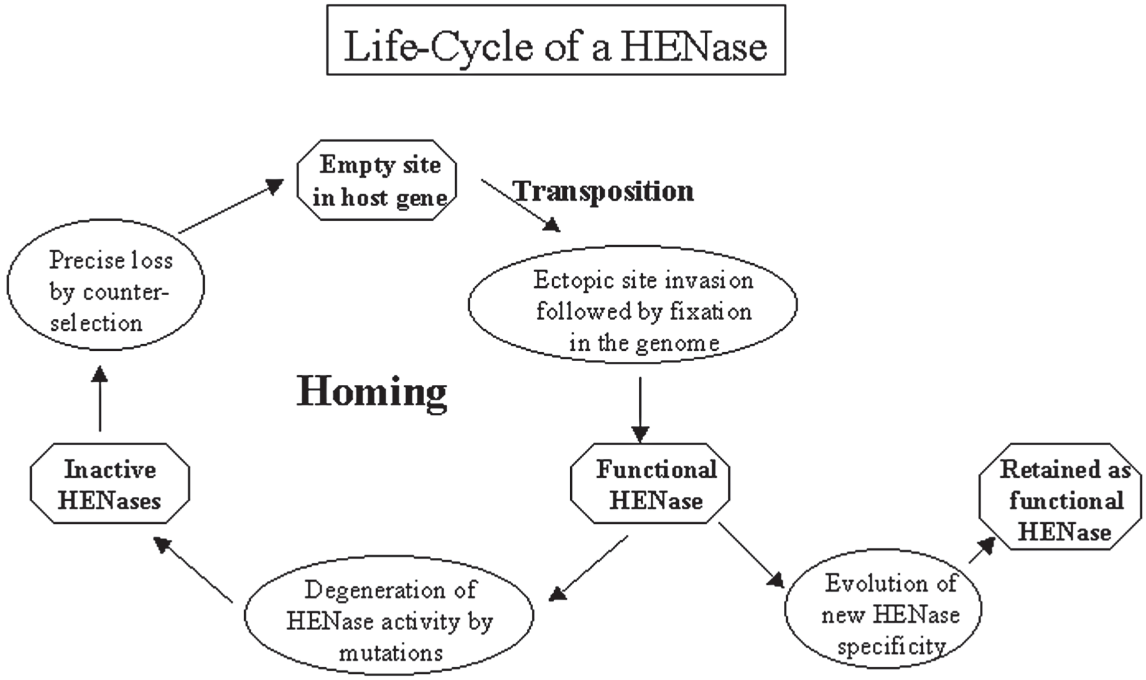
Life-Cycle of a HENase
Empty site in host gene
Transposition
Ectopic site invasion followed by fixation in the genome
Functional HENase
Retained as functional HENase
Evolution of new HENase specificity
Degeneration of HENase activity by mutations
Inactive HENases
Precise loss by counter-selection
Homing

FIGURE 12. The life-cycle of a homing endonuclease. The life-cycle of homing and transposition of a HEase. For details see the text. (Modified from Goddard and Burt, 1999; Gogarten *et al.*, 2002 and reproduced with permission from High Wire Press.)

developed in identifying new enzymes with unique specificities. Since restriction enzymes contact bases due to redundant interactions, engineering them is extremely challenging. However, HEases do not pose such a problem, and natural existence of LAGLIDADG HEase variants suggests that these enzymes are malleable. Modified HEases that target the sequence of interest could be developed to enable new gene replacement and inactivation strategies in a wide variety of organisms (Seligman *et al.*, 2002). Stoddard and his colleagues (2002) have engineered an artificial HEase (H-*Dre*I, fusion of I-*Cre*I and I-*Dmo*I), which recognizes and cleaves substrates with recognition half-sites from the parent enzymes. In addition, by altering the RNA sequence in the ribonucleoparticle of group II HEase, the target site (group II introns inserted into HIV-1 proviral DNA and human CCR5 gene target sites) within human cells was modified (Guo *et al.*, 2000). These results suggest that HEases are potentially useful in genetic engineering, functional genomics, gene therapy, gene cloning and targeting, and for elucidating the mechanism of DSB repair in diverse biological systems (reviewed by Belfort and Roberts, 1997; Jurica and Stoddard, 1999).

ACKNOWLEDGMENTS

We thank Drs. P. VanRoey, B. Stoddard, and P. Gogarten for kindly providing X-ray crystal structure figures, and the publishers of various journals for the permission to reproduce figures for this review. This research was supported by a grant from the Indian Council of Medical Research, New Delhi, to K.M., and a fellowship to N. Guhan from the Council of Scientific and Industrial Research, New Delhi.

REFERENCES

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.

Aggarwal, A. K. 1995. Structure and function of restriction endonucleases. *Curr. Opin. Struct. Biol.* **5**: 11–19.

Aggarwal, A. K. and Wah, D. A. 1998. Novel site-specific DNA endonucleases. *Curr. Opin. Struct. Biol.* **8**: 19–25.

Amitai, G., Belenkiy, O., Dassa, B., Shainskaya, A., and Pietrokovski, S. 2003. Distribution and function of new bacterial intein-like protein domains. *Mol. Microbiol.* **47**: 61–73.

Anraku, Y. 1997. Protein splicing: its chemistry and biology. *Genes Cells* **2**: 359–367.

Argast, G. M., Stephens, K. M., Emond, M. J., and Monnat, R. J. Jr. 1998. I-*Ppo*I and I-*Cre*I homing site degeneracy determined by random mutagenesis and sequential *in vitro* enrichment. *J. Mol. Biol.* **280**: 345–353.

Belfort, M. 1990. Phage T4 introns: self-splicing and mobility. *Annu. Rev. Genet.* **24**: 363–385.

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.

Belfort, M., and Roberts, R. J. 1997. Homing endonucleases: keeping the house in order. *Nucleic Acids Res.* **25**: 3379–3388.

Bell-Pedersen, D., Quirk, S., Clyman, J., and Belfort, M. 1990. Intron mobility in phage T4 is dependent upon a distinctive class of endonucleases and independent of DNA sequences encoding the intron core: mechanistic and evolutionary implications. *Nucleic Acids Res.* **18**: 3763–3770.

Beylot, B., and Spassky, A. 2001. Chemical probing shows that the intron-encoded endonuclease I-*Sce*I distorts DNA through binding in monomeric form to its homing site. *J. Biol. Chem.* **276**: 25243–25253.

Blackwood, K. S., He, C., Gunton, J., Turenne, C. Y., Wolfe, J., and Kabani, A. M. 2000. Evaluation of *recA* sequences for identification of *Mycobacterium* species. *J. Clin. Microbiol.* **38**: 2846–2852.

Bobola, N., Jansen, R. P., Shin, T. H., and Nasmyth, K. 1996. Asymmetric accumulation of Ash1p in post anaphase nuclei depends on a myosin and restricts yeast mating-type switching to mother cells. *Cell* **84**: 699–709.

Bolotin, M., Coen, D., Deutsch, J., Dujon, B., Netter, P., Petrochilo, E., and Slonimski, P. P. 1971. La recombinaison des mitochondries chez *Saccharomyces cerevisiae*. *Bull. Inst. Pasteur* **69**: 215–239.

Bonen, L., and Vogel, J. 2001. The ins and outs of group II introns. *Trends Genet.* **17**: 322–331.

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.

Bremer, M. C. D., Gimble, F. S., Thorner, J., and Smith, C. L. 1992. VDE endonuclease cleaves *Saccharomyces cerevisiae* genomic DNA at a single site: physical mapping of the *VMA1* gene. *Nucleic Acids Res.* **20**: 5484.

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. *The EMBO J.* **12**: 2141–2149.

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.

Bujnicki, J. M., Rotkiewicz, P., Kolinski, A., and Rychlewski, L. 2001a. Three-dimensional modeling of the I-*Tev*I homing endonuclease catalytic domain, a GIY-YIG superfamily member, using NMR restraints and Monte Carlo dynamics. *Protein Eng.* **14**: 717–721.

Bujnicki, J. M. and Rychlewski, L. 2001b. Unusual evolutionary history of the tRNA splicing endonuclease EndA: relationship to the LAGLIDADG and PD-(D/E)XK deoxyribonucleases. *Protein Sci.* **10**: 656–660.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S., and Venter, J. C. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**: 1058–1073.

Burggraf, S., Larsen, N., Woese, C. R., and Stetter, K. O. 1993. An intron within the 16S ribosomal RNA gene of the archaeon *Pyrobaculum aerophilum*. *Proc. Natl. Acad. Sci. USA* **90**: 2547–2550.

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

Cech, T. R. 1990. Self-splicing of group I introns. *Annu. Rev. Biochem.* **59**: 543–568.

Chen, X., Xu, M.–Q., Ding, Y., Ferrandon, S., and Rao, Z. 2002a. Purification and initial crystallization studies of a DnaB intein from *Synechocystis sp.* PCC 6803. *Acta. Crystallog. D. Biol. Crystallogr.* 58 (Pt 7): 1201–1203.

Chen, X., Xu, M.–Q., Ding, Y., Ferrandon, S., and Rao, Z. 2002b. Crystallographic study of a naturally occurring trans-splicing intein from *Synechocystis* sp. PCC 6803. *Acta. Crystallog. D. Biol. Crystallogr.* 58 (Pt 7): 1204–1206.

Cheng, Y.–S., Hsia, K.–C., Doudeva, L. G., Chak, K.–F., and Yuan, H. S. 2002. The crystal structure of the nuclease domain of colicin E7 suggests a mechanism for binding to double-stranded DNA by the H-N-H endonucleases. *J. Mol. Biol.* **324**: 227–236.

Chevalier, B. S., Monnat, R. J., Jr., and Stoddard, B. L. 2001a. The homing endonuclease I-*Cre*I uses three metals, one of which is shared between the two active sites. *Nat. Struct. Biol.* **8**: 312–316.

Chevalier, B. S. and Stoddard, B. L. 2001b. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res.* **29**: 3757–3774.

Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat, R. J., Jr., and Stoddard, B. L. 2002. Design, activity, and structure of a highly specific artificial endonuclease. *Mol. Cell.* **10**: 895–905.

Chong, S. and Xu, M.–Q. 1997. Protein splicing of the *Saccharomyces cerevisiae VMA* intein without the endonuclease motifs. *J. Biol. Chem.* **272**: 15587–15590.

Christ, F., Schoettler, S., Wende, W., Steuer, S., Pingoud, A., and Pingoud, V. 1999. The monomeric homing endonuclease PI-*Sce*I has two catalytic centers for cleavage of the two strands of its DNA substrate. *EMBO J.* **18**: 6908–6916.

Christ, F., Steuer, S., Thole, H., Wende, W., Pingoud, A., and Pingoud, V. 2000. A model for the PI-*Sce*IxDNA complex based on multiple base and phosphate backbone-specific photocross-links. *J. Mol. Biol.* **300**: 867–875.

Chu, F. K., Maley, G. F., Maley, F., and Belfort, M. 1984. Intervening sequence in the thymidylate synthase gene of bacteriophage T4. *Proc. Natl. Acad. Sci. USA* **81**: 3049–3053.

Coen, D., Deutsch, J., Netter, P., Petrochilo, E., and Slonimski, P. P. (1971) Mitochondrial genetics I. Methodology and phenomenology. In Miller (Ed.), Control of Organelle Development, Symp. Soc. Exp. Biol. Vol. 24, Cambridge University Press, Cambridge, U.K., pp. 449–496.

Coffey, T. J., Enright, M. C., Daniels, M., Morona, J. K., Morona, R., Hryniewicz, W. Paton, J. C., and Spratt, B. G. 1998. Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae*. *Mol. Microbiol.* **27**: 73–83.

Cole, S. T., Brosch., R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry III, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M.–A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., and Barrell, B. G. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537–544.

Cole, S.T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R. M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamline, N., Holroyd, S., Hornsby, T., Jagels, K., Lacrolz, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quall, M. A., Ragandream, M.–A., Rutherford, K. M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J. R., and Barrell, B. G. 2001. Massive gene decay in the leprosy bacillus. *Nature* **409**: 1007–1011.

Colleaux, L., D'Auriol, L., Betermier, M., Cottarel, G., Jacquier, A., Galibert, F., and Dujon, B. 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.

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.

Cooper, A. A., Chen, Y.-J., Lindorfer, M. A., and Stevens, T. H. 1993. Protein splicing of the yeast *TFP1* intervening protein sequence: a model for self-excision. *EMBO J.* **12**: 2575–2583.

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.

Cousineau, B., Lawrence, S., Smith, D., and Belfort, M. 2000. Retrotransposition of a bacterial group II intron. *Nature* **404**: 1018–1021

Cowan, J. A. 1998. Metal Activation of Enzymes in Nucleic Acid Biochemistry. *Chem. Rev.* **98**: 1067–1087.

Cowan, J. A. 2002. Structural and catalytic chemistry of magnesium-dependent enzymes. *Biometals* **15**: 225–235.

Crutz-Le Coq, A.–M., Cesselin, B., Commissaire, J., Anba, J. 2002. Sequence analysis of the lactococcal bacteriophage bIL170: insights into structural proteins and HNH endonucleases in dairy phages. *Microbiology* **148**: 985–1001.

Curcio, M. J. and Belfort, M. 1996. Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. *Cell* **84**: 9–12.

Dalgaard, J. Z. and Garrett, R. A. 1992. Protein-coding introns from the 23S rRNA-encoding gene form stable circles in the hyperthermophilic archaeon *Pyrobaculum organotrophum*. *Gene* **121**: 103–110.

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.

Dalgaard, J. Z., Klar, A. J., Moser, M. J., Holley, W. R., Chatterjee, A., and Mian, I. S. 1997a. Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucleic Acids Res.* **25**: 4626–4638.

Dalgaard, J. Z., Moser, M. J., Hughey, R., and Mian, I. S. 1997b. Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J. Comput. Biol.* **4**: 193–214.

Davis, E. O., Sedgwick, S. G., and Colston, M. J. 1991. Novel structure of the *recA* locus of *Mycobacterium tuberculosis* implies processing of the gene product. *J. Bacteriol.* **173**: 5653–5662.

Davis, E. O., Jenner, P. J., Brooks, P. C., Colston, M. J., and Sedgwick, S. G. 1992. Protein splicing in the maturation of *M. tuberculosis recA* protein: a mechanism for tolerating a novel class of intervening sequence. *Cell* **71**: 201–210.

Davis, E. O., Thangaraj, H. S., Brooks, P. C., and Colston, M. J. 1994. Evidence of selection for protein introns in the *recA*s of pathogenic mycobacteria. *EMBO J.* **13**: 699–703.

Dean, A. B., Stanger, M. J., Dansereau, J. T., Roey, P. V., Derbyshire, V., and Belfort, M. 2002. Zinc finger as distance determinant in the flexible linker of intron endonuclease I-*Tev*I. *Proc. Natl. Acad. Sci. USA* **99**: 8554–8561.

Decatur, W. A., Einvik, C., Johansen, S., and Vogt, V. M. 1995. Two group I ribozymes with different functions in a nuclear rDNA intron. *EMBO J.* **14**: 4558–4568.

Delahodde, A., Goquel, V., Becam, A. M., Creusot, F., Perea, J., Banroques, J., and Jacq, C. 1989. Site-specific DNA endonuclease and RNA maturase activities of two homologous intron-encoded proteins from yeast mitochondria. *Cell* **56**: 431–441.

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.

Derbyshire, V. and Belfort, M. 1998. Lightning strikes twice: intron-intein coincidence. *Proc. Natl. Acad. Sci. USA* **95**: 1356–1357.

Dickson, L., Huang, H. -R., Liu, L., Matsurra, M., Lambowitz, A. M., and Perlman, P. S. 2001. Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. *Proc. Natl. Acad. Sci. USA* **98**: 13207–13212.

Drouin, M., Lucas, P., Otis, C., Lemieux, C., and Turmel, M. 2000. Biochemical characterization of I-*Cmoe*I reveals that this H-N-H homing endonuclease shares functional similarities with H-N-H colicins. *Nucleic Acids Res.* **28**: 4566–4572.

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.

Dujon, B., Bolotin-Fukuhara, M., Coen, D., Deutsch, J., Netter, P., Slonimski, P. P., and Weil, L. 1976. Mitochondrial genetics. XI. Mutations at the mitochondrial locus omega affecting the recombination of mitochondrial genes in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.* **143**: 131–165.

Dujon, B. 1989. Group I introns as mobile genetic elements: facts and mechanistic speculations—a review. *Gene* **82**: 91–114.

Durrenberger, F. and Rochaix, J. -D. 1993. Characterization of the cleavage site and the recognition sequence of the I-*Cre*I DNA endonuclease encoded by the chloroplast ribosomal intron of *Chlamydomonas reinhardtii*. *Mol. Gen. Genet.* **236**: 409–414.

Durrenberger, F., Thompson, A. J., Herrin, D. L., and Rochaix, J.–D. 1996. Double strand break-induced recombination in *Chlamydomonas reinhardtii* chloroplasts. *Nucleic Acids Res.* **24**: 3323–3331.

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.

Edgell, D. R., Belfort, M., and Shub, D. A. 2000. Barriers to intron promiscuity in bacteria. *J. Bacteriol.* **182**: 5281–5289.

Edgell, D. R. and Shub, D. A. 2001. Related homing endonucleases I-*Bmo*I and I-*Tev*I use different strategies to cleave homologous recognition sites. *Proc. Natl. Acad. Sci. USA* **98**: 7898–7903.

Elde, M., Haugen, P., Willassen, N. P., and Johansen, S. 1999. I-*Nja*I, a nuclear intron-encoded homing endonuclease from *Naegleria*, generates a pentanucleotide 3' cleavage-overhang within a 19 base-pair partially symmetric DNA recognition site. *Eur. J. Biochem.* **259**: 281–288.

Elde, M., Willassen, N. P., and Johansen, S. 2000. Functional characterization of isoschizomeric His-

Cys box homing endonucleases from *Naegleria*. *Eur. J. Biochem.* **267**: 7257–7265.

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.

Evans, T. C. Jr., Martin, D., Kolly, R., Panne, D., Sun, L., Ghosh, I., Chen, L., Benner, J., Liu, X.-Q., and Xu, M.-Q 2000. Protein trans-splicing and cyclization by a naturally split intein from the *dnaE* gene of *Synechocystis species* PCC6803. *J. Biol. Chem.* **275**: 9091–9094.

Evans, T. C., Jr., and Xu, M.–Q. 2002. Mechanistic and kinetic considerations of protein splicing. *Chem. Rev.* **102**: 4869–4884.

Everett, K. D., Kahane, S., Bush, R. M., and Friedman, M. G. 1999. An unspliced group I intron in 23S rRNA links *Chlamydiales*, chloroplasts, and mitochondria. *J. Bacteriol.* **181**: 4734–4740.

Flick, K. E., McHugh, D., Heath, J. D., Stephens, K. M., Monnat, R. J., Jr., 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.

Flick, K. E., Jurica, M. S., Monnat, R. R. Jr., and Stoddard, B. L. 1998. DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-*Ppo*I. *Nature* **394**: 96–101.

Friedhoff, P., Franke, I., Meiss, G., Wende, W., Krause, K. L., and Pingoud, A. 1999. A similar active site for non-specific and specific endonucleases. *Nat. Struct. Biol.* **6**: 112–113.

Galburt, E. A., Chevalier, B., Tang, W., Jurica, M. S., Flick, K. E., Monnat, R. J., Jr., and Stoddard, B. L. 1999. A novel endonuclease mechanism directly visualized for I-*Ppo*I. *Nature Struct. Biol.* **6**: 1096–1099.

Galburt, E. A. Chadsey, M. S., Jurica, M. S., Chevalier, B. S., Erho, D., Tang, W., Monnat, R. J., Jr., and Stoddard, B. L. 2000. Conformational changes and cleavage by the homing endonuclease I-*Ppo*I: a critical role for a leucine residue in the active site. *J. Mol. Biol.* **300**: 877–887.

Galburt, E. A. and Stoddard, B. L. 2002. Catalytic mechanisms of restriction and homing endonucleases. *Biochemistry* **41**: 13851–13860.

Garrett, R. A., Dalgaard, J., Larsen, N., Kjems, J., and Mankin, A. S. 1991. Archaeal rRNA operons. *Trends Biochem. Sci.* **16**, 22–26.

Gimble, F. S. and Thorner, J. 1992. Homing of a DNA endonuclease gene by meiotic gene conversion in Saccharomyces cerevisiae. *Nature* **357**: 301–306.

Gimble, F. S. and Thorner, J. 1993. Purification and characterization of VDE, a site-specific endonuclease from the yeast Saccharomyces cerevisiae. *J. Biol. Chem.* **268**: 21844–21853.

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.

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.

Gimble, F. S. 1998. Putting protein splicing to work. *Chem. Biol.* **5**: R251–R256.

Gimble, F. S. 2000. Invasion of a multitude of genetic niches by mobile endonuclease genes. *FEMS Microbiol. Lett.* **185**: 99–107.

Gimble, F. S. 2001. Degeneration of a homing endonuclease and its target sequence in a wild yeast strain. *Nucleic Acids Res.* **29**: 4215–4223.

Giriat, I., Muir, T. W., and Perler, F. B. 2001. Protein splicing and its applications. *Genet. Eng.* **23**: 171–199.

Goddard, M. R. and Burt, A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci. USA* **96**: 13880–13885.

Goddard, M. R., Greig, D., and Burt, A. 2001. Outcrossed sex allows a selfish gene to invade yeast populations. *Proc. R. Soc. Lond. (Biol.)* **268**: 2537–2542.

Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L., and Hilario, E. 2002. Inteins: structure, function, and evolution. *Annu. Rev. Microbiol.* **56**: 263–287.

Goodrich-Blair, H., Scarlato, V., Gott, J. M., Xu, M.–Q., and Shub, D. A. 1990. A self-splicing group I intron in the DNA polymerase gene of *Bacillus subtilis* bacteriophage SPO1. *Cell* **63**: 417–424.

Goodrich-Blair, H. and Shub, D. A. 1994. The DNA polymerase genes of several HMU-bacteriophages have similar group I introns with highly divergent open reading frames. *Nucleic Acids Res.* **22**: 3715–3721.

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.

Gorbalenya, A. E. 1994. Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. *Prot. Sci.* **3**: 1117–1120.

Gorbalenya, A. E. 1998. Non-canonical inteins. *Nucleic Acids Res*. **26**: 1741–1748.

Gott, J. M., Shub, D. A., and Belfort, M. 1986. Multiple self-splicing introns in bacteriophage T4: evidence from autocatalytic GTP labeling of RNA *in vitro*. *Cell* **47**: 81–87.

Gott, G. M., Zeeh, A., Bell-Pedersen, D., Ehrenman, K., Belfort, M., and Shub, D. A. 1988. Genes within genes: independent expression of phage T4 intron open reading frames and the genes in which they reside. *Genes Dev*. **2**: 1791–1799.

Gray, M. W., Lang, B. F., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M., Brossard, N., Delage, E., Littlejohn, T. G., Plante, I., Rioux, P., Saint-Louis, D., Zhu, Y., and Burger, G. 1998. Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res*. **26**: 865–878.

Grindl, W., Wende, W., Pingoud, V., and Pingoud, A. 1998. The protein splicing domain of the homing endonuclease PI-*Sce*I is responsible for specific DNA binding. *Nucleic Acids Res*. **26**: 1857–1862.

Gruen, M., Chang, K., Serbanescu, I., and Liu, D. R. 2002. An *in vivo* selection system for homing endonuclease activity. *Nucleic Acids Res*. **30**: e29.

Guhan, N., and Muniyappa, K. 2002a. *Mycobacterium tuberculosis* RecA intein possesses a novel ATP-dependent site-specific double-stranded DNA endonuclease activity. *J. Biol. Chem*. **277**: 16257–16264.

Guhan, N. and Muniyappa, K. 2002b. The RecA intein of *Mycobacterium tuberculosis* promotes cleavage of ectopic DNA sites. Implications for the dispersal of inteins in natural populations. *J. Biol. Chem.* **277**: 40352–40361.

Guhan, N. and Muniyappa, K. (2003) Mycobacterium tuberculosis RecA intein, a LAGLIDADG homing endonuclease, displays Mn2+ and DNA-dependent ATPase activity. *Nuceic Acids Res*. In press.

Guo, H., Karberg, M., Long, M., Jones, J. P., III, Sullenger, B., and Lambowitz, A. M. 2000. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* **289**: 452–457.

Hall, T. M., Porter, J. A., Young, K. E., Koonin, E. V., Beachy, P. A., and Leahy, D. J. 1997. Crystal structure of a Hedgehog auto-processing domain: homology between Hedgehog and self-splicing proteins. *Cell* **91**: 85–97.

Hallick, R. B., Hong, L., Drager, R. G., Favreau, M. R., Monfort, A., Orsat, B., Spielmann, A., and Stutuz, E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res*. **21**: 3537–3544.

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.

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. *Nat. Struct. Biol*. **4**: 468–476.

Hensgens, L. A., Bonen, L., de Haan, M., van der Horst, G., and Grivell, L. A. 1983. Two intron sequences in yeast mitochondrial *COX1* gene: homology among URF-containing introns and strain-dependent variation in flanking exons. *Cell* **32**: 379–389.

Heyting, C. and Menke, H. H. 1979. Fine structure of the 21S ribosomal RNA region on yeast mitochondrial DNA. III. Physical location of mitochondrial genetic markers and the molecular nature of omega. *Mol. Gen. Genet*. **168**: 279–291.

Hickey, D. A., and Benkel, B. 1986. Introns as relict retrotransposons: implications for the evolutionary origin of eukaryotic mRNA splicing mechanisms *J. Theor. Biol*. **121**: 283–291.

Hickey, D. A. 1994. Protein introns: optional or essential? *Trends Genet*. **10**: 147–149.

Hirata, R., Ohsumi, Y., Nakano, A., Kawasaki, H., Suzuki, K., and Anraku, Y. 1990. Molecular structure of a gene, *VMA1*, encoding the catalytic subunit of H+-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. *J. Biol. Chem*. **265**: 6726–6733.

Hu, D., Crist, M., Duan, X., Quiocho, F. A., and Gimble, F. S. 2000. Probing the structure of the PI-SceI-DNA complex by affinity cleavage and affinity photocross-linking. *J. Biol. Chem*. **275**: 2705–2712.

Hurst, G. D. D. and Werren, J. H. 2001. The role of selfish genetic elements in eukaryotic evolution. *Nature Rev. Genet*. **2**: 597–606.

Ichiyanagi, K., Ishino, Y., Ariyoshi, M., Komori, K., and Morikawa, K. 2000. Crystal structure of an archaeal intein-encoded homing endonuclease PI-*Pfu*I. *J. Mol. Biol*. **300**: 889–901.

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.

Johansen, S., Embley, T. M., and Willassen, N. P. 1993. A family of nuclear homing endonucleases. *Nucleic Acids Res*. **21**: 4405.

Jurica, M. S., Monnat, R. J., Jr., and Stoddard, B. L. 1998. DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-*Cre*I. *Mol. Cell* **2**: 469–476.

Jurica, M. S. and Stoddard, B. L. 1999. Homing endonucleases: structure, function and evolution. *Cell Mol. Life Sci*. **55**: 1304–1326.

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 to the 69–kDα subunit of the vacuolar H+-adenosine triphosphatase. *Science* **250**: 651–657.

Kazazian, H. H., Jr., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., and Antonarakis, S. E. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**: 164–166.

Kjems, J. and Garrett, R. A. 1988. Novel splicing mechanism for the ribosomal RNA intron in the archaebacterium *Desulfurococcus mobilis*. *Cell* **54**: 693–703.

Klabunde, T., Sharma, S., Telenti, A., Jacobs, W. R., Jr., and Sacchettini, J. C. 1998. Crystal structure of GyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing. *Nat. Struct. Biol*. **5**: 31–36.

Klar, A. J. 1987. The mother-daughter mating type switching asymmetry of budding yeast is not conferred by the segregation of parental HO gene DNA strands. *Genes Dev*. **1**: 1059–1064.

Kleanthous, C. Kuhlmann, U. C., Pommer, A. J., Ferguson, N., Radford, S. E., Moore, G. R., James, R., and Hemmings, A. M. 1999. Structural and mechanistic basis of immunity toward endonuclease colicins. *Nat. Struct. Biol*. **6**: 243–252.

Ko, T.–P., Liao, C.–C., Ku, W.–Y., Chak, K.–F., and Yuan, H. S. 1999. The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. *Structure Fold Des*. **7**: 91–102.

Ko, M., Choi, H., and Park, C. 2002. Group I self-splicing intron in the *recA* gene of *Bacillus anthracis*. *J. Bacteriol*. **184**: 3917–3922.

Kobayashi, K., Nakahori, Y., Miyake, M., Matsumura, K., Kondo-Iida, E., Nomura, Y., Segawa, M., Yoshioka, M., Saito, K., Osawa, M., Hamano, K., Sakakihara, Y., Nonaka, I., Nakagome, Y., Kanazawa, I., Nakamura, Y., Tokunaga, K., and Toda, T. 1998. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature* **394**: 388–392.

Komori, K., Fujita, N., Ichiyanagi, K., Shinagawa, H., Morikawa, K., and Ishino, Y. 1999a. PI-*Pfu*I and PI-*Pfu*II, intein-coded homing endonucleases from *Pyrococcus furiosus*. I. Purification and identification of the homing-type endonuclease activities. *Nucleic Acids Res*. **27**: 4167–4174.

Komori, K., Ichiyanagi, K., Morikawa, K., and Ishino, Y. 1999b. PI-*Pfu*I and PI-*Pfu*II, intein-coded homing endonucleases from *Pyrococcus furiosus*. II. Characterization of the binding and cleavage abilities by site-directed mutagenesis. *Nucleic Acids Res*. **27**: 4175–4182.

Kostriken, R., Strathern, J. N., Klar, A. J., Hicks, J. B., and Heffron, F. 1983. A site-specific endonuclease essential for mating-type switching in *Saccharomyces cerevisiae*. *Cell* **35**: 167–174.

Koufopanou, V., Goddard, M. R., and Burt, A. 2002. Adaptation for horizontal transfer in a homing endonuclease. *Mol. Biol. Evol*. **19**: 239–246.

Kovall, R. A., and Matthews, B. W. 1998. Structural, functional, and evolutionary relationships between lambda-exonuclease and the type II restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **95**: 7893–7897.

Kovall, R. A. and Matthews, B. W. 1999. Type II restriction endonucleases: structural, functional and evolutionary relationships. *Curr. Opin. Chem. Biol*. **3**: 578–583.

Kowalski, J. C., Belfort, M., Stapleton, M. A., Holpert, M., Dansereau, J. T., Pietrokovski, S., Baxter, S. M., and Derbyshire, V. 1999. Configuration of the catalytic GIY-YIG domain of intron endonuclease I-*Tev*I: coincidence of computational and molecular findings. *Nucleic Acids Res*. **27**: 2115–2125.

Ku, W.-Y., Liu, Y.-W., Hsu, Y.-C., Liao, C.-C., Liang, P. -H., Yuan, H. S., and Chak, K.-F. 2002. The zinc ion in the HNH motif of the endonuclease domain of colicin E7 is not required for DNA binding but is essential for DNA hydrolysis. *Nucleic Acids Res*. **30**: 1670–1678.

Kuhlmann, U. C., Moore, G. R., James, R., Kleanthous, C., and Hemmings, A. M. 1999. Structural parsimony in endonuclease active sites: should the number of homing endonuclease families be redefined? *FEBS Lett*. **463**: 1–2.

Kulaeva, O. I., Koonin, E. V., Wootton, J. C., Levine, A. S., Woodgate, R. 1998. Unusual insertion element polymorphisms in the promoter and terminator regions of the mucAB-like genes of R471a and R446b. *Mutat. Res*. **397**: 247–262.

Lambowitz, A. M. 1989. Infectious introns. *Cell* **56**: 323–326.

Lambowitz, A. M. and Belfort, M. 1993. Introns as mobile genetic elements. *Annu. Rev. Biochem*. **62**: 587–622.

Li, W., Dennis, C. A., Moore, G. R., James, R., and Kleanthous, C. 1997. Protein-protein interaction specificity of Im9 for the endonuclease toxin colicin E9 defined by homologue-scanning mutagenesis. *J. Biol. Chem*. **272**: 22253–22258.

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.

Liu, X.-Q. 2000. Protein-splicing intein: genetic mobility, origin, and evolution. *Annu. Rev. Genet*. **34**: 61–76.

Long, R. M., Singer, R. H., Meng, X., Gonzalez, I, Nasmyth, K., and Jansen, R. P. 1997. Mating type switching in yeast controlled by asymmetric localization of *ASH1* mRNA. *Science* **277**: 383–387.

Loizos, N., Silva, G. H., and Belfort, M. 1996. Intron-encoded endonuclease I-*Tev*II binds across the minor groove and induces two distinct conformational changes in its DNA substrate. *J. Mol. Biol*. **255**: 412–424.

Lucas, P., Otis, C., Mercier, J.-O., Turmel, M., and Lemieux, C. 2001. Rapid evolution of the DNA-binding site in LAGLIDADG homing endonucleases. *Nucleic Acids Res*. **29**: 960–969.

Lykke-Andersen, J., Garret, R. A., and Kjems, J. 1996. Protein footprinting approach to mapping DNA binding sites of two archaeal homing enzymes: evidence for a two-domain protein structure. *Nucleic Acids Res*. **24**: 3982–3989.

Lykke-Andersen, J., Aagaard, C., Semionenkov, M., and Garrett, R. A. 1997a. Archaeal introns: splicing, intercellular mobility and evolution. *Trends Biochem. Sci*. **22**: 326–331.

Lykke-Andersen, J., Garrett, R. A., and Kjems, J. 1997b. Mapping metal ions at the catalytic centers of two intron-encoded endonucleases. *EMBO. J*. **16**: 3272–3281.

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.

Mannino, S. J., Jenkins, C. L., and Raines, R. T. 1999. Chemical mechanism of DNA cleavage by the homing endonuclease I-*Ppo*I. *Biochemistry* **38**: 16178–16186.

Martin, D. D., Xu, M.-Q., and Evans, T.C. Jr. 2001. Characterization of a naturally occurring trans-splicing intein from *Synechocystis* sp. PCC6803. *Biochemistry* **40**: 1393–1402.

Martinez-Abarca, F., Zekri, S., and Toro, N. 1998. Characterization and splicing *in vivo* of a *Sinorhizobium meliloti* group II intron associated with particular insertion sequences of the IS630–Tc1/IS3 retroposon superfamily. *Mol. Microbiol*. **28**: 1295–1306.

Martinez-Abarca, F., Garcia-Rodriguez, F. M., and Toro, N. 2000. Homing of a bacterial group II intron with an intron-encoded protein lacking a recognizable endonuclease domain. *Mol. Microbiol*. **35**: 1405–1412.

Michel, F. and Lang, B. F. 1985. Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses. *Nature* **316**: 641–643.

Michel, F. and Cummings, D. J. 1985. Analysis of class I introns in a mitochondrial plasmid associated with senescence of *Podospora anserina* reveals extraordinary resemblance to the *Tetrahymena* ribosomal intron. *Curr. Genet*. **10**: 69–79.

Michel, F., Umesono, K., and Ozeki, H. 1989. Comparative and functional anatomy of group II catalytic introns—a review. *Gene* **82**: 5–30.

Michel, F. and Ferat, J. L. 1995. Structure and activities of group II introns. *Annu. Rev. Biochem*. **64**: 435–461.

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.

Mills, D. A., Manias, A. A., McKay, L. L., and Dunny, G. M. 1997. Homing of a group II intron from *Lactococcus lactis* subsp. lactis ML3. *J. Bacteriol*. **179**: 6107–6111.

Monteilhet, C., Perrin, A., Thierry, A., Colleaux, L., and Dujon, B. 1990. Purification and characterization of the *in vitro* activity of I-*Sce*I, a novel and highly specific endonuclease encoded by a group I intron. *Nucleic Acids Res*. **18**: 1407–1413.

Monteilhet, C., Dziadkowiec, D., Szczepanek, T., and Lazowska, J. 2000. Purification and characterization of the DNA cleavage and recognition site of I-*Sca*I mitochondrial group I intron encoded endonuclease produced in *Escherichia coli*. *Nucleic Acids Res*. **28**: 1245–1251.

Moran, J. V., Zimmerly, S., Eskes, R., Kennel, J. C., Lambowitz, A. M., Butow, R. A., and Perlman, P. S. 1995. Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. *Mol. Cell. Biol*. **15**: 2828–2838.

Morozova, T., Seo, W., and Zimmerly, S. 2002. Non-cognate template usage and alternative priming by a group II intron-encoded reverse transcriptase. *J. Mol. Biol*. **315**: 951–963.

Moure, C. M., Gimble, F. S., and Quiocho, F. A. 2002. Crystal structure of the intein homing endonuclease PI-*Sce*I bound to its recognition sequence. *Nat. Struct. Biol*. **9**: 764–770.

Mueller, J. E., Smith, D., Bryk, 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.

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.

Mueller, M. W., Allmaier, M., Eskes, R., and Schweyen, R. J. 1993. Transposition of group II intron aI1 in yeast and invasion of mitochondrial genes at new locations. *Nature* **366**: 174–176.

Mullany, P., Pallen, M., Wilks, M., Stephen, J. R., and Tabaqchali, S. 1996. A group II intron in a conjugative transposon from the Gram-positive bacterium, *Clostridium difficile*. *Gene* **174**: 145–150.

Munoz, E., Villadas, P. J., and Toro, N. 2001. Ectopic transposition of a group II intron in natural bacterial populations. *Mol. Microbiol*. **41**: 645–652.

Muscarella, D. E. and Vogt, V. M. 1989. A mobile group I intron in the nuclear rDNA of *Physarum polycephalum*. *Cell* **56**: 443–454.

Muscarella, D. E., Ellison, E. L., Ruoff, B. M., and Vogt, V. M. 1990. Characterization of I-*Ppo*I, 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.

Nagai, Y., Nogami, S., Kumagia-Sano, F., and Ohya, Y. 2003. Karyopherin-mediated nuclear import of the homing endonuclease *VMA*1–derived endonuclease is required for self-propagation of the coding region. *Mol. Cell. Biol*. **23**: 1726–1736.

Naito, T., Kusano, K., and Kobayashi, I. 1995. Selfish behavior of restriction-modification systems. *Science* **267**: 897–899.

Nishioka, M., Fujiwara, S., Takagi, M., and Imanaka, T. 1998. Characterization of two intein homing endonucleases encoded in the DNA polymerase gene of *Pyrococcus kodakaraensis* strain KOD1. *Nucleic Acids Res*. **26**: 4409–4412.

Nogami, S., Fukuda, T., Nagai, Y., Yabe, S., Sugiura, M., Mizutani, R., Satow, Y., Anraku, Y., and Ohya, Y. 2002. Homing at an extragenic locus mediated by VDE (PI-*Sce*I) in *Saccharomyces cerevisiae*. *Yeast* **19**: 773–782.

Osiewacz, H. D. 2002. Mitochondrial functions and aging. *Gene* **286**: 65–71.

Paulus, H. 2000. Protein splicing and related forms of protein autoprocessing. *Annu. Rev. Biochem*. **69**: 447–496.

Paulus, H. 2001. Inteins as enzymes. *Bioorg. Chem*. **29**: 119–129.

Perler, F. B., Davis, E. O., Dean, G. E., Gimble, F. S., Jack, W. E., Neff, N., Noren, C. J., Thorner, J., and Belfort, M. 1994. Protein splicing elements: inteins and exteins—a definition of terms and recommended nomenclature. *Nucleic Acids Res*. **22**: 1125–1127.

Perler, F. B., Olsen, G. J., and Adam, E. 1997a. Compilation and analysis of intein sequences. *Nucleic Acids Res*. **25**: 1087–1093.

Perler, F.B., Xu, M.-Q., and Paulus, H. 1997b. Protein splicing and autoproteolysis mechanisms. *Curr. Opin. Chem. Biol*. **1**: 292–299.

Perler, F. B. 1998. Protein splicing of inteins and hedgehog autoproteolysis: structure, function, and evolution. *Cell* **92**: 1–4.

Perler, F. B. 1999. A natural example of protein trans-splicing. *Trends Biochem. Sci*. **24**: 209–211.

Perler, F. B., and Adam, A. 2000. Protein splicing and its applications. *Curr. Opin. Biotechnol*. **11**: 377–383.

Perler, F. B. 2002. InBase: the Intein Database. *Nucleic Acids Res*. **30**: 383–384.

Perlman, P. S., and Butow, R. A. 1989. Mobile introns and intron-encoded proteins. *Science* **246**: 1106–1109.

Philips, S. E. V. 1994. The β-ribbon of DNA recognition motif. *Annu. Rev. Biophys. Biomol. Struct*. **23**: 671–701.

Pingoud, A. and Jeltsch, A. 2001. Structure and function of type II restriction endonucleases. *Nucleic Acids Res*. **29**: 3705–3727.

Pingoud, V., Thole, H., Christ, F., Grindl, W., Wende, S., and Pingoud, A. 1999. Photocross-linking of the homing endonuclease PI-*Sce*I to its recognition sequence. *J. Biol. Chem*. **274**: 10235–10243.

Pingoud, V., Grindl, W., Wende, W., Thole, H., and Pingoud, A.1998. Structural and functional analysis of the homing endonuclease PI-*Sce*I by limited proteolytic cleavage and molecular cloning of partial digestion products. *Biochemistry* **37**: 8233–8243.

Pietrokovski, S. 1994. Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein. Sci*. **3**: 2340–2350.

Pietrokovski, S. 1996. A new intein in cyanobacteria and its significance for the spread of inteins. *Trends Genet*. **12**: 287–288.

Pietrokovski, S. 1998a. Modular organization of inteins and C-terminal autocatalytic domains. *Protein. Sci*. **7**: 64–71.

Pietrokovski, S. 1998b. Identification of a virus intein and a possible variation in the protein-splicing reaction. *Curr. Biol*. **8**: R634–R635.

Pietrokovski, S. 2001. Intein spread and extinction in evolution. *Trends Genet*. **17**: 465–472.

Porter, J. A., Young, K. E., and Beachy, P. A. 1996. Cholesterol modification of hedgehog signaling proteins in animal development. *Science* **274**: 255–259.

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.

Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J. et al., (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res*. **31**: 1805–1812.

Romine, M. F., Stillwell, L. C., Wong, K.–K., Thurston, S. J., Sisk, E. C., Sensen, C., Gaasterland, T., Fredrickson, J. K., and Saffer, J. D. 1999. Complete sequence of a 184–kilobase catabolic plasmid from *Sphingomonas aromaticivorans* F199. *J. Bacteriol*. **181**: 1585–1602.

Saldanha, R., Mohr, G., Belfort, M., and Lambowitz, A. M. 1993. Group I and group II introns. *FASEB J*. **7**: 15–24.

Sargueil, B., Delahodde, A., Hatat, D., Tian, G. L., Lazowska, J., and Jacq, C. 1991. A new specific DNA endonuclease activity in yeast mitochondria. *Mol. Gen. Genet*. **225**: 340–341.

Saguez, C., Lecellier, G., and Koll, F. 2000. Intronic GIY-YIG endonuclease gene in the mitochondrial genome of *Podospora curvicolla*: evidence for mobility. *Nucleic Acids Res*. **28**: 1299–1306.

Saves, I., Ozanne, V., Dietrich, J., and Masson, J.-M. 2000a. Inteins of *Thermococcus fumicolans* DNA polymerase are endonucleases with distinct enzymatic behaviors. *J. Biol. Chem*. **275**: 2335–2341.

Saves, I., Laneelle, M. -A., Daffe, M., and Masson, J.-M. 2000b. Inteins invading mycobacterial RecA proteins. *FEBS Lett*. **480**: 221–225.

Saves, I., Westrelin, F., Daffe, M., and Masson, J.-M. 2001b. Identification of the first eubacterial endonuclease coded by an intein allele in the *pps1* gene of mycobacteria. *Nucleic Acids Res*. **29**: 4310–4318.

Saves, I., Lewis, L.-A., Westrelin, F., Warren, R., Daffe, M., and Masson, J.-M. 2002a. Specificities and functions of the *recA* and *pps1* intein genes of *Mycobacterium tuberculosis* and application for diagnosis of tuberculosis. *J. Clin. Microbiol*. **40**: 943–950.

Saves, I., Morlot, C., Thion, L., Rolland, J.-L., Dietrich, J., and Masson, J. -M. 2002b. Investigating the endonuclease activity of four *Pyrococcus abyssi* inteins. *Nucleic Acids Res*. **30**: 4158–4165.

Schottler, S., Wende, W., Pingoud, V., and Pingoud, A. 2000. Identification of Asp218 and Asp326 as the principal Mg^{2+} binding ligands of the homing endonuclease PI-*Sce*I. *Biochemistry* **39**: 15895–15900.

Seligman, L. M., Chisholm, K. M., Chevalier, B. S., Chadsey, M. S., Edwards, S. T., Savage, J. H., and Veillet, A. L. 2002. Mutations altering the cleavage specificity of a homing endonuclease. *Nucleic Acids Res*. **30**: 3870–3879.

Sellem, C. H., Lecellier, G., and Belcour, L. 1993. Transposition of a group II intron. *Nature* **366**: 176–178.

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.

Sharp, P. A. 1985. On the origin of RNA splicing and introns. *Cell* **41**: 397–400.

Shao, Y. and Kent, S. B. 1997. Protein splicing: occurrence, mechanisms and related phenomena. *Chem. Biol*. **4**: 187–194.

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.

Shih, C. K., Wagner, R., Feinstein, S., KannikEnnulat, C., and Neff, N. 1988. A dominant trifluoperazine resistance gene from *Saccharomyces cerevisiae* has homology with F_0F_1 ATP synthase and confers calcium-sensitive growth. *Mol. Cell. Biol*. **8**: 3094–3103.

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.

Shub, D. A., Gott, J. M., Xu, M.-Q., Lang, B. F., Michel, F., Tomaschewski, J., Pederson-Lane, J., and Belfort, M. 1988. Structural conservation among three homologous introns of bacteriophage T4 and the group I introns of eukaryotes. *Proc. Natl. Acad. Sci. USA* **85:** 1151–1155.

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.

Strathern, J. N., Klar, A. J., Hicks, J. B., Abraham, J. A., Ivy, J. M., Nasmyth, K. A., and McGill, C. 1982. Homothallic switching of yeast mating type cassettes is initiated by a double-stranded cut in the *MAT* locus. *Cell* **31**: 183–192.

Sui, M. J., Tsai, L. C., Hsia, K. C., Doudeva, L. G., Ku, W. Y., Han, G. W., and Yuan, H. S. 2002. Metal ions and phosphate binding in the H-N-H motif: crystal structures of the nuclease domain of ColE7/Im7 in complex with a phosphate ion and different divalent metal ions. *Protein Sci*. **11**: 2947–2957.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., Stahl, F. W. 1983. The double-strand-break repair model for recombination. *Cell* **33**: 25–35.

Telenti, A., Southworth, M., Alcaide, F., Daugelat, S., Jacobs, W. R. Jr., and Perler, F. B. 1997. The *Mycobacterium xenopi* GyrA protein splicing element: characterization of a minimal intein. *J. Bacteriol*. **179**: 6378–6382.

Thompson, L. D., and Daniels, C. J. 1988. A tRNA(Trp) intron endonuclease from *Halobacterium volcanii*. Unique substrate recognition properties. *J. Biol. Chem*. **263**: 17951–17959.

Thompson, L. D. and Daniels, C. J. 1990. Recognition of exon-intron boundaries by the *Halobacterium volcanii* tRNA intron endonuclease. *J. Biol. Chem*. **265**: 18104–18111.

Toor, N., Hausner, G., and Zimmerly, S. 2001. Coevolution of group II intron RNA

structures with their intron-encoded reverse transcriptases. *RNA* **7**: 1142–1152.

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.

Vader, A., Nielsen, H., and Johansen, S. 1999. *In vivo* expression of the nucleolar groupI intron-encoded I-*Dir*I homing endonuclease involves the removal of a spliceosomal intron. *EMBO J*. **18**: 1003–1013.

VanRoey, P. V., Waddling, C. A., Fox, K. M., Belfort, M., and Derbyshire, V. 2001. Intertwined structure of the DNA-binding domain of intron endonuclease I-*Tev*I with its substrate. *EMBO J*. **20**: 3631–3637.

VanRoey, P. V., Meehan, L., Kowalski, J. C., Belfort, M., and Derbyshire, V. 2002. Catalytic domain structure and hypothesis for function of GIY-YIG intron endonuclease I-*Tev*I. *Nature Struct. Biol*. **9**: 806–811.

Walker, D. C., Georgiou, T., Pommer, A. J., Walker, D., Moore, G. R., Kleanthous, C., and James, R. 2002. Mutagenic scan of the H-N-H motif of colicin E9: implications for the mechanistic enzymology of colicins, homing enzymes and apoptotic endonucleases. *Nucleic Acids Res*. **30**: 3225–3234.

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.

Wende, W., Grindl, W., Christ, F., Pingoud, A., and Pingoud, W. 1996. Binding, bending and cleavage of DNA substrates by the homing endonuclease PI-*Sce*I. *Nucleic Acids Res*. **24**: 4123–4132.

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.

Werner, E., Wende, W., Pingoud, A., and Heinemann, U. 2002. High resolution crystal structure of domain I of the *Saccharomyces cerevisiae* homing endonuclease PI- *Sce*I. *Nucleic Acids Res*. **30**: 3962–3971.

Wittmayer, P. K., and Raines, R. T. 1996. Substrate binding and turnover by the highly specific I-*Ppo*I endonuclease. *Biochemistry* **35**: 1076–1083.

Wu, H., Hu, Z., and Liu, X. -Q. 1998a. Protein trans-splicing by a split intein encoded in a split *dnaE* gene of *Synechocystis sp*. PCC6803. *Proc. Natl. Acad. Sci. USA* **95**: 9226–9231.

Wu. H., Xu, M.-Q., and Liu, X.-Q. 1998b. Protein trans-splicing and functional mini-inteins of a cyanobacterial *dnaB* intein. *Biochim. Biophys. Acta* **1387**: 422–432.

Wu, W., Wood, D. W., Belfort, G., Derbyshire, V., and Belfort, M. 2002. Intein-mediated purification of cytotoxic endonuclease I-*Tev*I by insertional inactivation and pH-controllable splicing. *Nucleic Acids Res*. **30**: 4864–4871.

Xiong, Y. and Eickbush, T. H. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J*. **9**: 3353–3362.

Xu, M. -Q., Southworth, M. W., Mersha, F. B., Hornstra, L. J., and Perler, F. B. 1993. *In vitro* protein splicing of purified precursor and the identification of a branched intermediate. *Cell* **75**, 1371–1377.

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. *Nature* **381**: 332–335.

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 aI1 and reverse splicing into DNA transposition sites in vitro. *J. Mol. Biol*. **282**: 505–523.

Yeo, C. C., Tham, J. M., Yap, M. W., and Poh, C. L. 1997. Group II intron from *Pseudomonas alcaligenes* NCIB 9867 (P25X): entrapment in plasmid RP4 and sequence analysis. *Microbiology* **143** (Pt 8): 2833–2840.

Young, P., Ohman, M., Xu, M.–Q., Shub, D. A., and Sjoberg, B. M. 1994. Intron-containing T4 bacteriophage gene *sunY* encodes an anaerobic ribonucleotide reductase. *J. Biol. Chem*. **269**: 20229–20232.

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.

Zimmerly, S., Guo, H., Eskes, R., Yang, J., Perlman, P. S., and Lambowitz, A. M. 1995a. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell* **83**: 529–538.

Zimmerly, S., Guo, H., Perlman, S., and Lambowitz, A. M. 1995b. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**: 545–554.