Published online 9 March 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 12* ***5189–5200***
doi:10.1093/nar/gks226

SURVEY AND SUMMARY

Structural, functional and evolutionary relationships between homing endonucleases and proteins from their host organisms

Gregory K. Taylor[1,2] and Barry L. Stoddard[2,*]

[1]Graduate Program in Molecular and Cellular Biology, University of Washington and [2]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. A3-025, Seattle, WA 90109, USA

Received December 12, 2011; Revised February 12, 2012; Accepted February 22, 2012

ABSTRACT

Homing endonucleases (HEs) are highly specific DNA-cleaving enzymes that are encoded by invasive DNA elements (usually mobile introns or inteins) within the genomes of phage, bacteria, archea, protista and eukaryotic organelles. Six unique structural HE families, that collectively span four distinct nuclease catalytic motifs, have been characterized to date. Members of each family display structural homology and functional relationships to a wide variety of proteins from various organisms. The biological functions of those proteins are highly disparate and include non-specific DNA-degradation enzymes, restriction endonucleases, DNA-repair enzymes, resolvases, intron splicing factors and transcription factors. These relationships suggest that modern day HEs share common ancestors with proteins involved in genome fidelity, maintenance and gene expression. This review summarizes the results of structural studies of HEs and corresponding proteins from host organisms that have illustrated the manner in which these factors are related.

Homing endonucleases (HEs) are mobile genetic elements that selfishly propagate their own reading frames in a dominant non-Mendelian fashion (1). These proteins generally display no obvious biological role other than to perpetuate themselves through a mechanism that is initiated by cleavage of a specific-genomic target, which then is forced to act as a recipient for the HE gene. Insertion of these mobile elements occurs because DNA cleavage by the HE stimulates double-strand break repair via homologous recombination, which results in precise insertion of the HE reading frame (often in concert with an associated intron or intein sequence) into the DNA-target site.

At least six distinct structural families of HEs (the 'LAGLIDADG', 'HNH', 'His-Cys box', 'GIY-YIG', 'PD-(D/E)xK' and the most recently described 'EDxHD' proteins) have been identified (2,3). Each family is classified and named according to the presence of a conserved sequence motif that corresponds to critical structural and catalytic residues. These six HE structural families span at least four distinct active site catalytic motifs that are each found broadly throughout all kingdoms of life and are associated with a wide variety of additional nuclease and/or DNA-binding activities. This includes the 'GIY-YIG' nuclease motif (4) (Figure 1); the 'LAGLIDADG' motif (5,6) (Figure 2); the 'ββα-metal' motif (7) (Figure 3) and the 'PD-(D/E)xK' motif (8) (Figure 4). The latter two catalytic motifs are each distributed across two separate HE lineages. The phage-derived HNH endonucleases (9) and the His-Cys box HEs from protista (10) contain closely related ββα-metal active sites, while the 'EDxHD' HEs in phage (3) and PD-(D/E)xK HEs from bacteria (11,12) also contain related (but more significantly diverged) catalytic core motifs.

Despite a wide variation in their structures, mechanisms and catalytic core motifs, all HEs must successfully meet similar functional requirements (2). They are usually encoded by short reading frames (<1 kB), presumably to minimize their impact upon the folding and function of their surrounding mobile elements (which often correspond to self-splicing introns or inteins). Their biological function requires the readout of long DNA targets (that range from about 14 to over 30 bp in length) and the simultaneous accommodation of sequence polymorphisms that correspond to poorly conserved bases in their host target sites (such as wobble positions in protein coding sequences). This combination of properties allows an

*To whom correspondence should be addressed. Tel: +1 206 667 4031; Fax: +1 206 667 3331; Email: bstoddar@fhcrc.org

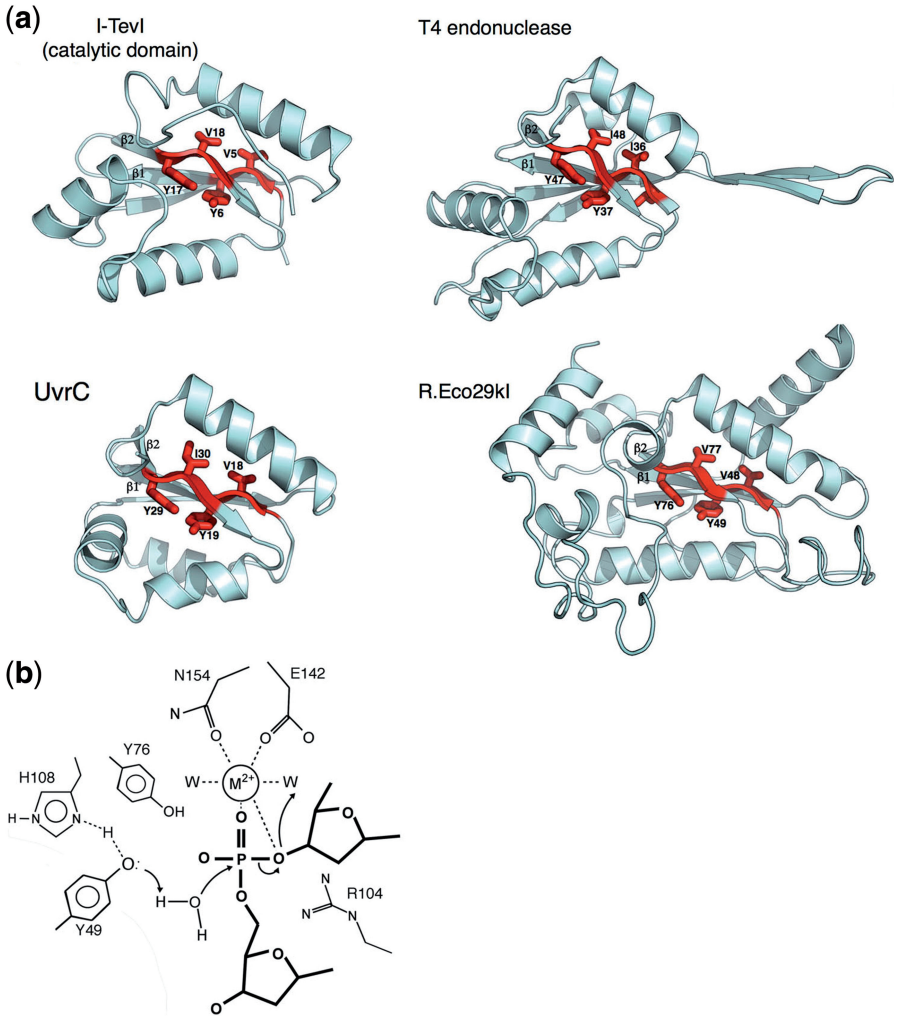

Figure 1. The GIY-YIG motif. (**a**) Enzymes involved in homing (I-TevI), DNA degradation (T4 endonuclease), restriction (R.Eco29kI) and DNA repair (UvrC) contain similar GIY-YIG catalytic cores, in which two short antiparallel β-strands contain the conserved signature residues for the enzyme family (shown in red). Based on crystal structures of the R.Eco29kI and R.Hpy188I REases bound to their DNA targets, the two strongly conserved tyrosine residues in the catalytic motif (Y6 and Y17 in I-TevI) are believed to be involved in general acid–base catalysis and activation of a water nucleophile (13,14). (**b**) The putative catalytic mechanism of GIY-YIG endonucleases involves activation of incoming nucleophilic water by an active-site tyrosine, which is itself activated through interactions with surrounding basic side chains. A single-bound divalent metal ion is coordinated by the scissile phosphate and neighboring active site side chains. This mechanism and active site bears a strong resemblance to the HNH nuclease motif (Figure 3), but has evolved using a completely different surrounding protein-fold topology. The side chain labels and features shown are based on the R.Eco29kI/DNA crystal structure (13).

HE to display sufficient specificity to avoid significant toxicity to its host, while facilitating its continued vertical inheritance and persistence within potential future generations of organisms.

The evolutionary origin of the first HE is unknown, and the precise evolutionary route by which any of the modern HEs families were generated is not understood. However, bioinformatic and structural studies of representatives from each unique HE lineage have repeatedly demonstrated that they share common structural folds with a wide variety of proteins that are involved in many biological functions and pathways.

In this review, we summarize the results of structural studies, now spanning the past 15 years, which have collectively illustrated the various manners in which individual HE families are related to proteins of different biological and molecular functions. Implicit in this summary is a view that there are at least three evolutionary scenarios by which such relationships might have been established. In the first, a modern HE family and one or more proteins from the host organism (referred to through this review as 'host proteins') represent the products of divergence from a common ancestor. In the second, an established HE might have acquired a secondary biological function (e.g. the ability to act as a 'maturase' and thereby facilitate intron splicing). This may involve the acquisition of additional functional domains as has been seen in the evolution of host proteins related to the HNH, GIY-YIG and LAGLIDADG HE families. This form of functional moonlighting can result in the loss of the original HE function and subsequent specialization in the protein's newly acquired function, presumably because that host-specific biological role then became the primary target of selective pressure to maintain the

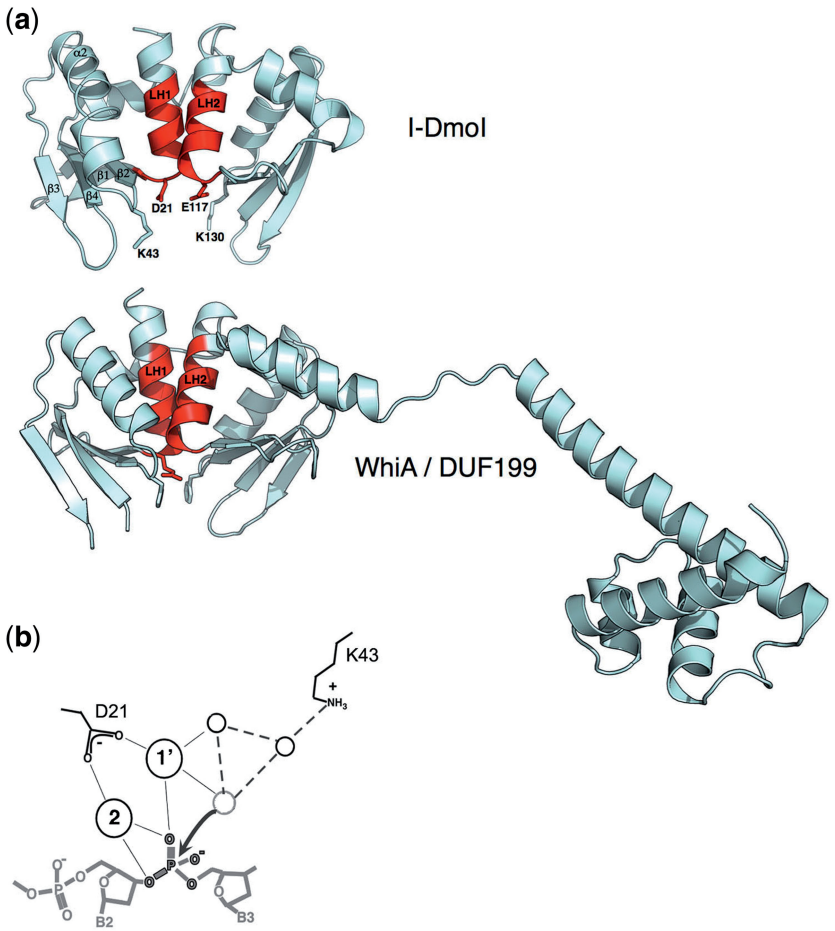

Figure 2. The LAGLIDADG motif. (**a**) HEs such as I-DmoI display a core fold consisting of two copies of an 'αββαββ' topology in which the first helix in each fold (colored red and labeled 'LH1' and 'LH2') contain the consensus sequence motif, and pack against one another to comprise both a domain interface and the center of the endonuclease active site. An acidic residue (usually an aspartate, but in many cases, a glutamate) extends from the bottom of each helix (D21 and E117 in I-DmoI); together they coordinate multiple divalent metal ions in conjunction with the scissile phosphate oxygens. Strongly conserved basic residues (K43 and K130 in I-DmoI) extend from the β1−β2 loop in the active sites and are believed to play a role in stabilizing the phosphoanion transition state and/or assisting in general acid/base catalysis. In contrast, the WhiA/DUF199 family of bacterial gene regulators contains a LAGLIDADG protein domain that closely resembles the HE structural family, tethered to a C-terminal helix-turn-helix domain. However, the catalytic acidic and basic residues described above are not conserved (in the case of the WhiA protein from Thermatogus maritima, the positions of the LAGLIDADG acidic residues are instead an arginine and glycine; the positions of the neighboring basic residues are two phenylalanines. As well, the overall positively charged surface of the HE that is formed by its β-sheets is instead considerably more varied in its charge composition, indicating that the DNA-binding properties of the LAGLIDADG fold have been replaced with alternative roles. (**b**) The putative mechanism of the LAGLIDADG endonucleases involves activation of incoming metal-bound nucleophilic water by a network of surrounding basic side chains and additional solvent molecules. The two most conserved residues in the active site (indicated and labeled based on the structure of the I-DmoI endonuclease) are an acidic metal-binding residue contributed by the penultimate residue of each LAGLIDADG motif and a neighboring basic residue (usually a lysine) bound on an adjacent DNA-binding loop. Cleavage of the DNA appears to follow a mechanism that involves two bound metals for each DNA-strand scission event. Many LAGLIDADG endonucleases display considerable disparity in the kinetics of individual strand cleavage events, such that significant fraction of nicked intermediate accumulates prior to final double-strand break formation.

protein's form and function. Finally, it is formally possible that some of these relationships are the result of convergent evolution and that HEs and proteins from their biological hosts appear structurally similar by chance (i.e. via convergent evolution) rather than as a result of divergence from a common ancestor. This final scenario is generally considered most likely for proteins that share relatively simple structural motifs, and less likely where extensive topologies are found in common between two proteins.

While the introduction above and the corresponding figures throughout this review are arranged according to the divisions between established structural families of HEs and their corresponding catalytic motifs, the following sections present a series of biological functions (ranging from genomic modification and repair to transcriptional regulation) that offer proteins with a diverse set of biochemical and biological functions that harbor clear relationships with HEs.

COMPETITIVE CYTOTOXICITY

Escherichia coli and many other bacterial species can produce and release a family of cytotoxic proteins termed colicins, often under various stress conditions (15). Colicins

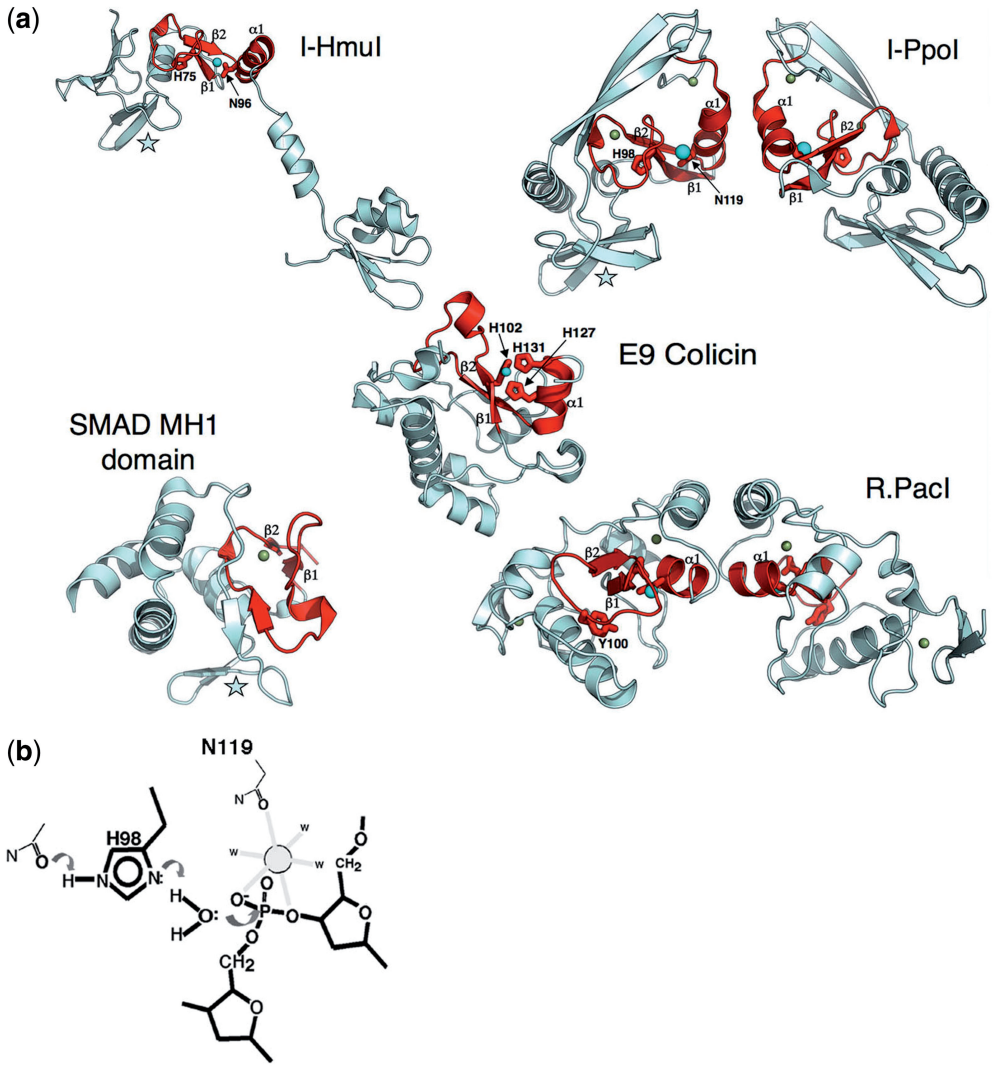

Figure 3. The ββα−metal motif. (**a**) HE families found in either phage (such as I-HmuI) or in protists (such as I-PpoI) contain quite similar ββα-metal catalytic core folds and active-site residues (colored red). In both enzymes, an active-site histidine residue (H75 in I-HmuI; H98 in I-PpoI) acts as a general base to assist in deprotonation and activation of an active-site water molecule. A neighboring asparagine located at the N-terminal end of the motif's α-helix is involved in coordination of a single-bound divalent metal ion, which participates in transition state stabilization. Both active-site core folds are embellished by a small antiparallel β-sheet (denoted by a star), which is involved in sequence-specific DNA-site recognition, by forming base pair interactions in the target site major groove. Beyond these common elements, the two HEs differ significantly. I-HmuI displays an extended monomeric structure in which a pair of α-helices forms additional contacts to the 3′ end of the DNA-target site. In contrast, I-PpoI forms a homodimeric structure in which two identical copies of the enzyme DNA-binding surface interact with distal ends of a symmetric DNA sequence. The ββα-metal core fold is found in proteins with substantially different biochemical functions and biological roles, ranging from competitive DNA degradation and toxicity (the bacterial colicins), phage restriction (R.PacI), and even eukaryotic transcriptional regulation (the SMAD proteins). The divergent biological roles of these proteins is reflected in additional sequence and structural variation of the ββα-metal core: the colicins display considerably different metal coordination schemes from the HEs (E9 colicin employs three histidine residues); R.PacI appears to have replaced the active-site histidine base with a tyrosine, and the DNA-binding MH1 domain of the SMAD proteins have completely devolved catalytic regions, but have maintained the same overall architecture of the DNA-binding surface (also indicated with a star) as is observed for I-HmuI and I-PpoI. (**b**)The putative mechanisms of DNA cleavage by the ββα-metal nucleases involves activation of an incoming nucleophilic water by an active-site histidine, which is itself activated through interactions with surrounding basic side chains. A single-bound divalent metal ion is coordinated by the scissile phosphate and neighboring active site side chains. This mechanism and active site bears a strong resemblance to the HNH nuclease motif (Figure 1), but has evolved using a completely different surrounding protein-fold topology. The side chain labels and features shown are based on the I-PpoI/DNA crystal structure (10).

are believed to confer an advantage to their hosts in the presence of competing bacterial organisms, particularly when nutrients are limited or the cell is otherwise exposed to environmental challenges such as ionizing radiation or DNA-damaging reagents. Colicin domains usually display a modular, multi-domain architecture. In most cases, the N-terminal domain is usually responsible for translocation, the central domain facilitates receptor binding and the C-terminal domain represents the active cytotoxic agent. Once the colicin has been introduced into the cytoplasm of the target cell, the cytotoxic domain acts via a mechanism that is dictated by its unique structure and function. Various colicin systems incorporate cytoxic domains that are capable of RNA degradation, membrane

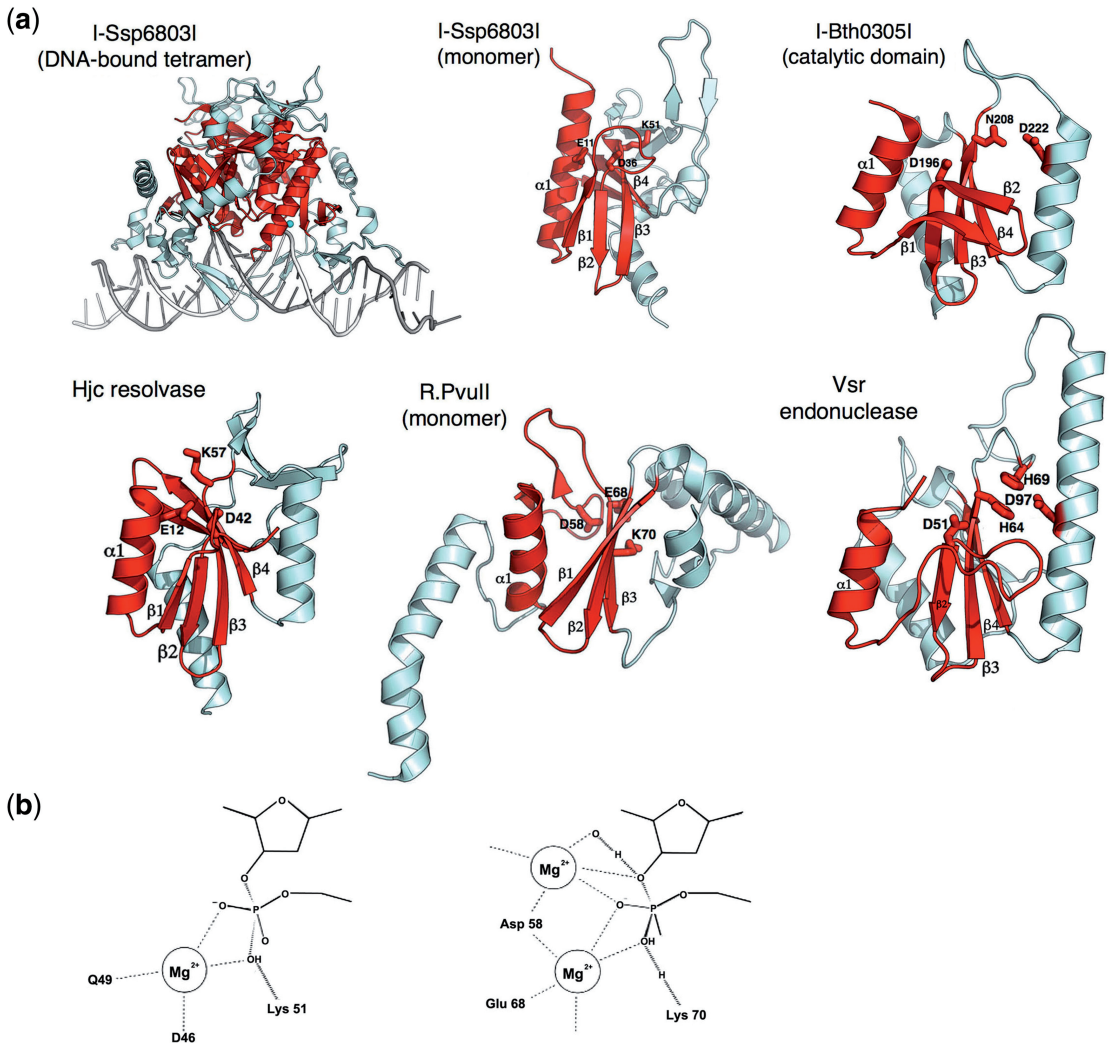

Figure 4. The PD-(D/E)xK motif. (**a**)As has been observed for the ββα-metal motif (Figure 3), HEs found in two widely different genomic niches (phage and eubacteria) have evolved that contain different versions of the PD-(D/E)xK nuclease motif. In the traditional motif (I-Ssp6803I and its closest structural relatives), a minimum of two acidic residues and one basic residue are position within a mixed α/β topology (colored in red) and participate in divalent metal coordination and promote general acid/base catalysis and/or transition state stabilization. The I-Ssp6803I HE (found within a cyanobacterium) forms a tetrameric structure in which two copies of the enzyme fold contact the DNA-target half-sites, while another two copies promote the overall quaternary interactions necessary to properly position the two binding surfaces at opposite ends of the target. Similar subunit architectures and active site chemistries are observed both for many different DNA-active enzymes. Some of the mostly closely related such enzymes to the I-Ssp6803I HE, identified using automated searches of the PDB database, are an archael Holliday junction resolvase (Hjc) and the R.PvuII and R.SfiI REases. In contrast, the phage-derived I-Bth0305I HE displays a common core-folded topology as the traditional PD-(D/E)xK nucleases, but displays a significantly diverged active-site architecture and presumed mechanism of DNA cleavage that most closely resembles the very short patch repair (Vsr) endonuclease (42). Such HEs have been termed the 'EDxHD' family (3) to denote their conservation pattern of active-site residues and to distinguish them from the bacterial HEs.(**b**) The generic mechanism of DNA hydrolysis by the PD-(D/E)xK containing nucleases involves the activation of a metal-bound water either through direct interaction with a basic side chain, which acts as a general base, or through a water-mediated contact. PD-(D/E)xK nucleases display considerable variation both in the number and exact position of bound metal ions during catalysis, as well as the exact structural position within the core protein fold of each catalytic residue. The I-Ssp6803I HE appears to bind one divalent metal ion (left), while the R.PvuII REase is thought to bind metal ions within at least two distinct sites during catalysis (right).

depolarization, inhibition of murein synthesis or (in the case discussed below) non-specific DNAse activity. To protect against self-cytoxic activity, cells producing colicins often co-express an inhibitor protein that physically sequesters and blocks the action of the cytotoxic domain until release from the host.

The active sites of monomeric DNAse colicins contain an HNH-nuclease motif (Figure 3), corresponding to a ββα-metal active site, which has been described by a variety of crystallographic analyses of colicins E7 and E9 (16–19). The residues of the HNH motif are found in a concave crevice in the surrounding protein fold that is believed to provide space for binding of double-stranded DNA in a sequence non-specific manner. Several of the residues in the active site of these enzymes coordinate a single-divalent metal ion that is required to stabilize the phosphoanion transition state and the 3′ oxygen leaving group of the reaction. An absolutely conserved

histidine residue acts as a general base for the reaction, specifically to activate a water nucleophile.

The active sites of bacterial colicins, as well as non-specific microbial endonucleases such as the secreted nuclease from *Serratia marcencens* were found to display similar architectures to the active site of the *Physarum polycephalum* His-Cys box HE I-PpoI (7,10) (Figure 3). Whereas the colicin nucleases display relatively small, compact-domain architectures that reflect their function as non-specific DNA-degradation enzymes, I-PpoI contains several structural elaborations beyond the HNH motif and associated ββα-metal core fold that are required for dimerization and for sequence-specific DNA recognition.

The observation that the HNH-nuclease motif is broadly distributed across both HEs and a variety of distantly related host proteins was further illustrated by the DNA-bound crystal structure of the phage-derived HE I-HmuI (9). Unlike I-PpoI, that enzyme and a large number of related phage HEs (20,21) display monomeric structures in which their HNH-catalytic nuclease domains are tethered to independent DNA-binding regions via an overall protein-domain organization that is unique from either bacterial colicins or the His-Cys-box HEs.

RESTRICTION-MODIFICATION

Bacterial genomes contain a wide variety of genetic systems that are believed to act biologically to protect their hosts against phage infections, as well as other potential sources of incoming foreign DNA (22). The best studied of these correspond to restriction-modification (RM) systems, which include reading frames that encode restriction endonuclease (REase) enzymes that recognize short nucleotide sequences with extremely high fidelity (23). Many, if not all, bacterial genera possess multiple RM systems (22); in each one the REase acts in concert with a cognate DNA-modification activity that chemically modifies the same target sequence within the host genome (usually via base methylation within the same target-site sequence) so that cleavage is effectively blocked.

RM enzyme systems are classified according to their subunit composition and their mechanism of recognition and action on DNA (24). Class II RM systems are small and do not require ATP hydrolysis or the action of motor proteins for target-site recognition, DNA cleavage or modification. In most (but not all) class II systems, the REases act independently of their cognate methyltransferase to cleave their specific DNA targets. Several thousand of class II REases have been biochemically characterized (25), and many more have been identified during the course of microbial genomic sequencing and annotation efforts around the world.

In contrast to HEs, REases usually recognize short sequences (generally 4–8 bp in length) with high fidelity (26). A large number of crystallographic analyses of various type II REase/DNA complex have demonstrated that the REase typically contacts the target DNA sequence with a mechanism that includes the formation of a large number (15–20) of directional hydrogen bonds that specifically participate in recognition of the individual bases through the major and/or the minor groove (27).

In addition to their fundamental protective role in the bacterial host, the genes encoding at least some REases and their associated modification enzymes have also been proposed to act as selfish DNA (28). According to this theory, loss of the modification activity leads to cell death via residual activity of the restriction enzyme, and thereby imposes a form of negative selection against elimination of RM systems.

A large percentage of well-characterized REases belong to the PD-(D/E)xK structural superfamily (Figure 4), in which metal ions (coordinated by the conserved acidic residues of the motif) participate in activation of the hydroxyl nucleophile and stabilization of the phosphoanion transition state, and the basic residue facilitates charge stabilization and/or proton transfer steps of the reaction. The exact mechanism and number of metal ions required for catalysis for almost any unique type II REase is usually somewhat ambiguous.

In general, REases appear to undergo rapid divergence, and different REase families exhibit very little sequence similarity (24). Despite their low sequence similarity, it has been proposed that most if not all PD-(D/E)xK type II REases are descended from a common ancestor by divergent evolution (29). As expected, the active site is the most structurally conserved region in PD-(D/E)xK endonucleases, albeit with obvious cases in which the position of individual catalytic residues have been 'swapped' between different structural elements in the active-site architecture.

The I-Ssp6803I HE was the first HE to be shown to contain a PD-(D/E)xK core fold and to resemble REases from that family (Figure 4) (11,12). This HE and its close homologues are generally encoded in cyanobacteria. The enzyme forms a tetramer in solution; upon sequence recognition, two subunits make contact with the DNA while the other two provide additional quaternary structural interactions that allow interaction of the protein across its long DNA target. This allows the HE to recognize a pseudo-palindromic target sequence 23 bp in length. When compared to the type II REases that have been visualized crystallographically, I-Ssp6803I particularly resembles the R.PvuII REase, with an RMSD of 3.3 Å over aligned Cα atoms (12) (Figure 4). Despite their similar size and tertiary folds, the mechanism of DNA-target site recognition by the two enzymes is highly diverged, with I-Ssp6803I recognizing a long target with highly variable degrees of fidelity exhibited at individual DNA base pairs (in contrast to recognition of a 6 bp target with absolute fidelity by R.PvuII). Even though they recognize very dissimilar target sites with very different balances of overall specificity and fidelity, I-Ssp6803I makes approximately the same number of nucleotide specific contacts as R.PvuII does to its target.

In addition to the PD-(D/E)xK REase enzyme superfamily, a significant number of additional type II restriction enzymes contain either GIY-YIG (Figure 1) or the ββα-metal (Figure 3) catalytic cores and active sites motifs (4,30–32). The DNA-bound structures of the GIY-YIG REases R.Eco29kI and R.Hpy188I have been

solved (13,14), which has allowed direct comparisons with the structure and proposed catalytic mechanism of the GIY-YIG HE I-TevI (33). The catalytic core of a GIY-YIG endonuclease follows a 'β-β-α-β-α' topology where the first two β strands contain the residues GIY and YIG. The active-site architecture and proposed mechanism of phosphoryl hydrolysis resembles that of the HNH enzymes, with the notable exception that the first tyrosine residue in the GIY-YIG motif is proposed to act as the immediate general base for activation and formation of the hydroxyl nucleophile, with adjacent basic residues involved in reducing the pK_a of the tyrosine side chain and thus increasing its reactivity.

The catalytic domain of the I-TevI GIY-YIG HE represents a minimal, compact nuclease core fold, corresponding to its role as a modular nuclease domain with minimal sequence specificity (Figure 1). Specifically, I-TevI prefers the sequence 5′ CNNN/G-3′ for efficient cleavage (with and / representing the bottom- and top-strand nicking sites respectively). The structures of R.Eco29kI and R.Hpy188I demonstrate that the requirement for high-fidelity DNA recognition has been met through the incorporation of additional structural elements around and within the catalytic core fold. R.Eco29kI has an extended DNA-binding loop immediately after the second β strand of the GIY-YIG motif, as well as a unique α helix inserted between the two β strands. This unique helix lies on the surface of the protein, distant from both the active site and the bound DNA; it appears to have a purely structural role in the protein fold and does not directly participate in the site of catalysis. The sequence identity between the catalytic core domain of R.Eco29kI and the nuclease domain of I-TevI is 12% and the structure superposition has an RMSD of about 2.9 Å for backbone atoms (13).

Structures of two HNH-containing REases (R.PacI and R.Hpy99I) have also been determined (Figure 3) and the HNH motif within R.KpnI has also been well-characterized biochemically (32,34,35). These REases are all homodimers containing one ββα-metal motif per subunit. Similar to the I-PpoI HE, the DNA-bound co-crystal structures of R.PacI and R.Hpy99I indicate that those enzymes contain two bound zinc ions per protein subunit; however, all three enzymes have evolved different additional structural elaborations around their active sites and equally unique DNA-binding modes. Whereas the I-PpoI enzyme recognizes a 14 bp target site, again with moderate fidelity at several positions, the restriction enzymes recognize considerably shorter target sites with absolute fidelity. The heart of the Hpy99I protein forms a structure that wraps around its target site, aligning the helices from the catalytic site ββα-metal motif almost perpendicular with the DNA-duplex axis. In contrast, PacI binds via an elongated fold. In that structure, two subunits and the ββα-metal motif aligned almost parallel to the DNA duplex.

DNA REPAIR

Nucleotide excision functions

UvrABC is a multienzyme complex found in *E. coli* and other bacteria that are involved in 'short patch' nucleotide excision repair in response to DNA damage at individual bases. The sequence of events in the UvrABC-mediated damage recognition and nucleotide excision reaction are relatively well established (36). UvrC, working in conjunction with UvrA and UvrB, mediates two-strand scission events on the same DNA strand, with one cleavage event located four nucleotides 3′ of the lesion, and the second eight nucleotides 5′ to the lesion. The two-strand cleavage events generate a 12-nt fragment of DNA containing the lesion. After incision, DNA helicase II (UvrD) releases UvrC and the excised oligonucleotide. DNA polymerase I then resynthesizes the excised strand and removes UvrB from the non-damaged DNA strand in the process. DNA ligase I joins the synthesized DNA to the template finishing the nucleotide excision repair pathway.

Bioinformatic analyses and homology searches using the sequence of *E. coli* UvrC revealed a bacterial homolog named Cho (36). This protein is homologous to the N-terminal region of UvrC and can initiate 3′DNA-strand cleavage, but not 5′cleavage. As previously demonstrated for UvrC, Cho is also dependent on UvrAB but UvrC and Cho interact with different UvrB domains. Cho and UvrC are both encoded in several bacterial species including *E. coli*, but the greater majority of bacteria contain only a recognizable copy of UvrC. In some organisms, such as mycoplasma and *Borrelia burgdorferi*, only Cho is found. In these cases, a 5′-strand cleavage activity might originate from an additional exonuclease domain found on Cho or from the exonuclease activity of an alternative enzyme. This may be plausible as Cho proteins of the mycoplasma species are larger than those of *E. coli*.

The nucleotide excision repair proteins UvrC and Cho share homology with the catalytic domain of the GIY-YIG family of HEs, as typified by the I-TevI HE (37). The two proteins roughly follow a structural motif of α1-β1-β2-α2-α3-β3-α4-α5 (Figure 1). At the center of each structure is a β sheet that contains the GIY-YIG catalytic motif on β1 and β2. The catalytic domain of UvrC and the catalytic domain of I-TevI have relatively low-sequence identity of 15%. Given their low-sequence identity, it is notable that the two structures superimpose with an RMSD of 2.2 Å for 60 of 89 possible Cα atoms (37). While the two structures have a nearly identical topology, there are clear differences in their secondary and tertiary structure. First, an additional helix, α1, is present in the UvrC structure compared to I-TevI. This helix is likely structural and appears not to be involved in catalysis, because residues that form the helix are not conserved among various UvrC homologues. Secondly, the region spanning α2 and β3, which includes α3, is not structurally conserved compared to I-TevI. Nevertheless, a residue that stabilizes the hydrophobic core of the domain superimposes between the two structures (Ile45 from UvrC and Leu 56 in I-TevI). Finally, the terminal helix α5 in the motif is found in neither I-TevI nor all UvrC homologs.

Mismatch repair functions

In the first step of DNA mismatch repair in bacteria, MutS binds to base pair mismatches and to small insertion/

deletion loops (38). MutS is a functional heterodimer with one monomer binding the mismatch, and the other binding non-specifically to the surrounding DNA. Each subunit also contains an ATPase domain that interacts with the DNA-binding domain. The MutS–DNA–ATP complex then interacts with MutL which also binds DNA and ATP. Interaction of MutL with DNA is mediated primarily through MutS and occurs independently of ATP hydrolysis. ATP hydrolysis by MutL is then required for interaction with many of the downstream proteins required for completion of mismatch repair, one of which is termed the **V**ery **S**hort patch **R**epair protein or 'Vsr'.

Unlike other mismatch repair proteins, Vsr recognizes mismatches in the context of a longer sequence. Through recruitment by MutL, this single-strand endonuclease preferentially targets T/G mismatches within hemi-methylated 5′-CTWGG/5′CCWGG sequences where W is an A or a T [the 3′C of CCWGG sequences is the substrate for the bacterial DNA-cytosine methyltransferase (Dcm)] (39). Vsr cleaves the DNA 5′ of the mismatched T, so that after removal of downstream bases, DNA polymerase I may perform templated DNA resynthesis, creating a short repair patch. DNA ligase then reseals the DNA patch into the DNA backbone.

In a recent analysis of environmental metagenomic sequence data collected by the Global Ocean Sampling project, a novel type of fractured gene was discovered corresponding to separately encoded halves of self-splicing inteins that interrupt individual host genes in the same locus (40). The inteins were frequently found to be interrupted by open reading frames that do not exhibit significant sequence similarity to previously characterized HE families. Further analysis indicated that the uncharacterized open reading frames were associated with introns, inteins, or as free-standing genes. In total 15 members, including two in previously annotated genes in the NCBI-sequence database, were described. Limited sequence homology to the catalytic domain of Vsr endonucleases was detected in the C-terminal region of the translated protein sequences of these genes (40). The established catalytic residues from Vsr endonucleases were conserved across all members of the new gene family. These residues include an essential aspartate that coordinates a catalytic magnesium ion, a histidine thought to act as a general base, and a proximal aspartate residue. Inferred from the presence of endonuclease catalytic residues within the domain, this gene family was hypothesized to encode a novel lineage of HEs. The activity, specificity, and structure have been characterized for one representative member of this family, I-Bth0305I (3). The crystal structure of the catalytic-domain support a similar mechanism for DNA-strand cleavage and confirms that members of this HE family share a common ancestor with the Vsr mismatch repair endonuclease (Figure 4).

This newly discovered HE family has been named the 'EDxHD' family after conserved catalytic residues. Vsr endonucleases and the 'EDxHD' HEs display a type II restriction enzyme topology that has significantly diverged from the traditional 'PD-(D/E)xK' motif and appears to employ an activated histidine as a general base (3). In contrast, the lysine residue in the PD-(D/E)xK motif is often assigned this role in the catalytic mechanism. Further subtle divergence of catalytic mechanism is indicated by an additional highly conserved acidic residue in the active-site region. Apart from these two exceptions, the enzyme has maintained most the features of this unique active-site arrangement. The observed bipartite arrangement of the catalytic domain is not common with Vsr but the relationship between the two proteins is clear when comparing global topologies.

DNA CROSSOVER RESOLUTION

Four-way DNA (Holliday) junctions are branch-points generated by the interconnection of four helices during strand exchange events that are necessary for various DNA integration, transposition, and recombination processes (41). Four-way junctions are resolved by junction resolving enzymes to create duplex products. These nucleases are highly specific for the structure of DNA junctions where they initiate cleavage at the four-way junction. Junction-resolving enzymes have been isolated from a number of different organisms ranging from bacteria, bacteriophages, archaea, yeast, and mammalian cells and their viruses.

In comparing the crystal structure of the I-Ssp6803I HE to previously determined macromolecular structures, a similar core fold corresponds to the archael Holliday-junction resolving enzyme (Figure 4) (12). Specifically, the Hjc enzyme from *Pyrococcus furiosus* aligns with an RMSD of 2.4 Å (1.9 Å across the catalytic core) (12). Whereas I-SspI forms a tetramer to bind a long duplex DNA target, four-way junction resolving enzymes form a dimer to recognize the junction itself. This is accomplished through the creation of two DNA-binding channels that are 30 Å in length, formed on both sides of the dimer. These channels are positively charged and make extensive contact with the arms containing the 5′ ends of the continuous strands. This results in the burial of 4180 $Å^2$ of solvent accessible protein surface and the channels hold the DNA arms in a perpendicular orientation (41). The relationship of the catalytic core between a HE and a four-way junction resolving enzyme suggests a common ancestor even with the different oligomeric state found in each of the two proteins.

POST-TRANSCRIPTIONAL SPLICING AND MATING SWITCHING

Whereas all of the examples provided above appear to represent situations where modern day HEs and contemporary host proteins have diverged from ancient common ancestors, there exist at least two cases where established HEs appear to have developed secondary biological activities and roles in the host, which in time led to the original invasive DNA function giving way entirely to an important host-specific role. For example, many HEs also participate in the post-transcriptional splicing of their host intron, by assisting the folding of their cognate RNA intron—a function termed 'maturase' activity (43–50). In some cases, such maturases have retained their original HE activity and thus, moonlight between both activities (51) where in other cases, the HE activity has been lost—in

some cases through a single, presumably recent-point mutation that can be easily reverted to restore endonuclease activity (47).

In a separate example, some HEs have been adopted by the host to act directly as free-standing endonucleases that drive biologically important gene conversion events. For example, the HO endonuclease in yeast, which is responsible for the mating-type genetic switch in that organism, is a LAGLIDADG protein which appears to be derived from an intein-associated HE (52).

GENETIC REGULATION

The DNA-binding properties of HEs appears to facilitate their ability to be utilized, either directly or as a result of evolutionary repurposing, as genetic regulators. For example, the I-TevI HE moonlights as a transcriptional repressor, acting to suppress its own expression (53,54). At least two examples have been described in the literature of considerably more distant relationships between HEs and genetic regulators: the WhiA/DUF199 family of bacterial sporulation factors and the eukaryotic SMAD proteins.

Transcriptional regulation via WhiA/DUF199

The initiation of mRNA synthesis depends ultimately on factors that interact with specific elements in gene promoters (55). These proteins are composed of a wide variety of usually separable DNA-binding and transcriptional activation domains. The DNA-binding subregions of many transcription factors consist of 60–100 amino acids and are necessary but not sufficient for transcriptional activation. These regions are tethered to transcriptional activation domains that are required for the initiation of transcription, presumably through recruitment of RNA polymerase.

One family of putative bacterial transcription factors named DUF199 is present in all Gram-positive bacteria (56). One representative member of this family, WhiA, was observed in bioinformatic and structural studies to contain a core LAGLIDADG-sequence motif and corresponding fold and topology at its N-terminal region, tethered to a C-terminal helix-turn-helix domain (57,58). The WhiA protein is essential for sporulation in *Streptomyces coelicolor* and related *Streptomycete* strains, and appears to regulate expression of multiple sporulation-specific 'Whi' genes (56). Notably, WhiA regulates expression of its own reading frame and at least one other sporulation-specific transcript (ParAB2), and appears to interact with and regulate the activity of the sporulation-specific sigma factor WhiG (59). All Gram-positive bacteria contain similar Whi operons including a single recognizable DUF199/WhiA protein. This conservation suggests that WhiA homologs function in a similar manner.

The similarities and differences between WhiA sequence and structure relative to its closest bacterial homologs and more distantly related LAGLIDADG HEs are displayed in Figure 2. Analysis of the structure elucidates how unique evolutionary pressures that are placed upon a genetic regulator versus those placed on an invasive endonuclease might produce individually tailored structures and biochemical features that are appropriate for each function. The protein-fold topology observed in monomeric LAGLIDADG HEs is observed in the N-terminal region of WhiA. Monomeric LAGLIDADG HEs are composed of two structurally similar domains, each containing an αββαββ core that are connected by a short peptide linker. The closest structural homolog of WhiA, identified by the DALI webserver, is the I-DmoI HE, which is an archaeal enzyme encoded within a mobile group I intron. The two sequences have low-sequence identity of 13% and the structures superimpose with an α-carbon RMSD across all aligned residues of 2.4 Å (58). Conserved elements include those residues that comprise the two LAGLIDADG helices that form the core of the domain interface. Intimate packing between backbone atoms in the helices resulted in helices that are closely superimposable.

A key difference between LAGLIDADG HEs and WhiA family members is that the WhiA proteins lack acidic residues at the base of the LAGLIDADG helices that coordinate metal ions in HEs. In I-DmoI (60), these conserved residues correspond to D20 and E117 and are essential for catalysis. Other catalytic residues, such as K42 and K120 in I-DmoI, are not conserved in WhiA. These residues are basic residues that are involved in transition-state stabilization in HEs. These positions are occupied by a histidine and methionine (H54 and M125, respectively) in the WhiA structure and are similarly non-conserved in close homologs. As a consequence, WhiA family members cannot be endonucleases and do not digest DNA in controlled experiments.

The mechanism of DNA recognition and binding by WhiA LAGLIDADG domains might differ significantly from that displayed by the same domains in the HE. Enzymes such as I-DmoI make extensive contacts with their DNA substrates using a pair of antiparallel β sheets and associated loops. These structural elements make interactions with the DNA backbone with individual nucleotide base pairs across the entire DNA target. Each LAGLIDADG domain recognizes a single-DNA half-site using DNA-contact surfaces that are uniformly positively charged. The only exception to this surface is the presence of conserved metal coordinating acid residues in the active sites at the center of the domain interface.

The surface of WhiA corresponding to the DNA-binding surface of the N-terminal domain in traditional LAGLIDADG HE displays significant negative surface charge. Also, the C-terminal LAGLIDADG domain displays positively charged surface that extends well beyond its β-sheet region. Consequently, the DUF199/WhiA protein family is expected to interact with its DNA target in a different manner from the mode of DNA binding exhibited by LAGLIDADG HEs such as I-DmoI; it is quite possible that the LAGLIDADG domain in the WhiA/DUF199 family has entirely surrendered DNA-binding function to the helix-turn-helix domain and is instead involved in protein–protein interactions required for its role as a gene expression regulator.

Transcriptional regulation via Smad Proteins

SMADs are intracellular proteins that are involved in transducing signals to the nucleus, in response to the presence of various growth factors, in order to activate expression of the TGF-beta gene (61). The DNA-binding domain of the Smad transcriptional regulator in the TGF-B signaling cascade has been found to resemble the overall topology of the His-Cys-Box HE I-PpoI (62). Smad consists of two domains, MH1 and MH2. The MH2 domain is homologous to a large family of nuclear signaling protein–protein interaction domains in eukaryotes and prokaryotes. A presumably unique spatial structure of the MH1 domain earned it a unique fold classification in the SCOP database. A combination of sequence and structure-based analyses show that the MH1 domain is homologous to the His-Cys-Box HE family (Figure 3). The structural similarity was first detected by the DALI server with a 16% sequence identity and an RMSD of 3.3 Å between 78 aligned α-carbons (62).

The structural organization of I-PpoI follows three subdomain architectures with two subdomains having structural equivalents in MH1 Smad. Notably, the first subdomain is a triple-stranded β-sheet that binds in the major groove of DNA; the turn between β strands incorporates the active site Arg61. Further, MH1 and I-PpoI have similar secondary structural elements in the same topological connection and spatial arrangement. From this global comparison, it is clear that they possess the same fold (62) and likely share a common ancestor.

CONCLUSIONS

The ability to recognize and interact with nucleic acid targets in a specific manner and to modify their structure, organization and/or sequence through tightly controlled catalysis of phosphoryl hydrolysis and transfer reactions is one of the most fundamental and universal set of functions to be assumed by proteins in the modern biological universe. Of the hundreds of recognized and accepted unique protein folds to have been visualized to date, a large number encompass a subset of proteins that are in some way involved in nucleic acid chemistry, organization or metabolism. A considerably smaller number of protein folds, including those described in this review, are specifically tasked with the fundamental function of phosphodiester-bond cleavage via hydrolysis.

The structure–function relationships between modern HEs and their contemporaneous cousins found in host genomes provides a tempting opportunity to suggest that these particular nuclease families represent particularly early or 'ancient' proteins folds, or that certain modern families of nucleic acid-acting enzymes or genetic regulators may have arisen from ancestral mobile elements. However, it should be noted that solid evidence of either hypothesis is, at best, scarce. Bioinformatics-based studies of the establishment, distribution and evolutionary history of protein folds throughout the known biological kingdoms (63) does not appear to identify a set of the likely 'most ancient' protein folds that coincides with the HE structural families; nor is there any obvious evidence for the presence of mobile endonuclease ancestors prior to the establishment of enzyme activities that are involved in the most fundamental aspects of genomic maintenance and fidelity.

Nevertheless, the structural relationships observed between HEs (a significant number of which are unique to phage) and eukaryotic protein factors involved in DNA metabolism or gene expression speaks clearly to the historical intersection and divergence of prokaryotic, eukaryotic and archael genomes, as well as the phage and viruses that act upon each of those kingdoms. David Shub (64) stated that 'The odd thing about bacteriophages is how frequently they surprise us', while a 2004 review by Howard Ochman (65) outlined the multiple ways in which genes associated with parasitic or selfish elements, particularly from phage, are often adopted by their hosts for a wide variety of biological functions. That phage can be a rich source of HE reading frames and closely related protein factors is made clear by the fact that 15 separate HE genes correspond to 11% of the total coding sequence of the T4-phage genome (66). Given the examples described in this review, a related observation might be that ancient battles for space and resources, including events reduced all the way down to introns competing for common genomic insertion sites, have likely provided a crucible of evolutionary innovation that has lead to a least part of the modern repertoire of nucleic acid enzymes and other factors found throughout the biosphere.

FUNDING

Funding for open access charge: National Institutes of Health (Grant R01 GM49857).

Conflict of interest statement. One coauthor (B.L.S.) is a cofounder of a biotech startup (Pregenen, Inc) that works on the engineering of LAGLIDADG homing endonucleases for applications requiring targeted gene modification.

REFERENCES

1. Dujon,B. (1989) Group I introns as mobile genetic elements: facts and mechanistic speculations – a review. *Gene*, **82**, 91–114.
2. Stoddard,B.L. (2011) Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. *Structure*, **19**, 7–15.
3. Taylor,G.K., Heiter,D.F., Pietrokovski,S. and Stoddard,B.L. (2011) Activity, specificity and structure of I-Bth0305I: a representative of a new homing endonuclease family. *Nucleic Acids Res.*, **39**, 9705–9719.
4. Dunin-Horkawicz,S., Feder,M. and Bujnicki,J.M. (2006) Phylogenomic analysis of the GIY-YIG nuclease superfamily. *BMC Genomics*, **7**, 98.
5. Dalgaard,J.Z., Klar,A.J., Moser,M.J., Holley,W.R., Chatterjee,A. and Mian,I.S. (1997) Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucleic Acids Res.*, **25**, 4626–4638.
6. Chevalier,B., Monnat,R.J. and Stoddard,B.L. (2005) LAGLIDADG Homing Endonucleases. In: Belfort,M., Wood,D., Derbyshire,V. and Stoddard,B. (eds), *Homing endonucleases and inteins*, Vol. 16. Springer, Berlin, pp. 34–47.

7. 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.
8. Laganeckas,M., Margelevicius,M. and Venclovas,C. (2011) Identification of new homologs of PD-(D/E)XK nucleases by support vector machines trained on data derived from profile-profile alignments. *Nucleic Acids Res*., **39**, 1187–1196.
9. Shen,B.W., Landthaler,M., Shub,D.A. and Stoddard,B.L. (2004) DNA binding and cleavage by the HNH homing endonuclease I-HmuI. *J. Mol. Biol*., **342**, 43–56.
10. Galburt,E.A., Chevalier,B., Tang,W., Jurica,M.S., Flick,K.E., Monnat,R.J. and Stoddard,B.L. (1999) A novel endonuclease mechanism directly visualized for I-PpoI. *Nat. Struct. Biol*., **6**, 1096–1099.
11. Orlowski,J., Boniecki,M. and Bujnicki,J.M. (2007) I-Ssp6803I: the first homing endonuclease from the PD-(D/E)XK superfamily exhibits an unusual mode of DNA recognition. *Bioinformatics*, **23**, 527–530.
12. Zhao,L., Bonocora,R.P., Shub,D.A. and Stoddard,B.L. (2007) The restriction fold turns to the dark side: a bacterial homing endonuclease with a PD-(D/E)-XK motif. *EMBO J*., **26**, 2432–2442.
13. Mak,A.N., Lambert,A.R. and Stoddard,B.L. (2010) Folding, DNA recognition, and function of GIY-YIG endonucleases: crystal structures of R.Eco29kI. *Structure*, **18**, 1321–1331.
14. Sokolowska,M., Czapinska,H. and Bochtler,M. (2011) Hpy188I-DNA pre- and post-cleavage complexes – snapshots of the GIY-YIG nuclease mediated catalysis. *Nucleic Acids Res*., **39**, 1554–1564.
15. Lancaster,L.E., Wintermeyer,W. and Rodnina,M.V. (2007) Colicins and their potential in cancer treatment. *Blood Cells Mol. Dis*., **38**, 15–18.
16. 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 HNH endonucleases. *J. Mol. Biol*., **324**, 227–236.
17. Garinot-Schneider,C., Pommer,A.J., Moore,G.R., Kleanthous,C. and James,R. (1996) Identification of putative active-site residues in the DNase domain of colicin E9 by random mutagenesis. *J. Mol. Biol*., **260**, 731–742.
18. 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*, **7**, 91–102.
19. Pommer,A.J., Cal,S., Keeble,A.H., Walker,D., Evans,S.J., Kuhlmann,U.C., Cooper,A., Connolly,B.A., Hemmings,A.M., Moore,G.R. *et al*. (2001) Mechanism and cleavage specificity of the H-N-H endonuclease colicin e9. *J. Mol. Biol*., **314**, 735–749.
20. Drouin,M., Lucas,P., Otis,C., Lemieux,C. and Turmel,M. (2000) Biochemical characterization of I-CmoeI reveals that this H-N-H homing endonuclease shares functional similarities with H-N-H colicins. *Nucleic Acids Res*., **28**, 4566–4572.
21. Landthaler,M., Shen,B.W., Stoddard,B.L. and Shub,D.A. (2006) I-BasI and I-HmuI: two phage intron-encoded endonucleases with homologous DNA recognition sequences but distinct DNA specificities. *J. Mol. Biol*., **358**, 1137–1151.
22. Labrie,S.J., Samson,J.E. and Moineau,S. (2010) Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol*., **8**, 317–327.
23. Nathans,D. and Smith,H.O. (1975) Restriction endonucleases in the analysis and restructuring of dna molecules. *Annu. Rev. Biochem*., **44**, 273–293.
24. Bujnicki,J.M. (2001) Understanding the evolution of restriction-modification systems: clues from sequence and structure comparisons. *Acta Biochim. Pol*., **48**, 935–967.
25. Orlowski,J. and Bujnicki,J.M. (2008) Structural and evolutionary classification of Type II restriction enzymes based on theoretical and experimental analyses. *Nucleic Acids Res*., **36**, 3552–3569.
26. Pingoud,A. and Jeltsch,A. (2001) Structure and function of type II restriction endonucleases. *Nucleic Acids Res*., **29**, 3705–3727.
27. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*., **29**, 2860–2874.
28. Naito,T., Kusano,K. and Kobayashi,I. (1995) Selfish behavior of restriction-modification systems. *Science*, **267**, 897–899.
29. Fuxreiter,M. and Simon,I. (2002) Protein stability indicates divergent evolution of PD-(D/E)XK type II restriction endonucleases. *Protein Sci*., **11**, 1978–1983.
30. Ibryashkina,E.M., Zakharova,M.V., Baskunov,V.B., Bogdanova,E.S., Nagornykh,M.O., Den'mukhamedov,M.M., Melnik,B.S., Kolinski,A., Gront,D., Feder,M. *et al*. (2007) Type II restriction endonuclease R.Eco29kI is a member of the GIY-YIG nuclease superfamily. *BMC Struct. Biol*., **7**, 48.
31. Kaminska,K.H., Kawai,M., Boniecki,M., Kobayashi,I. and Bujnicki,J.M. (2008) Type II restriction endonuclease R.Hpy188I belongs to the GIY-YIG nuclease superfamily, but exhibits an unusual active site. *BMC Struct. Biol*., **8**, 48.
32. Saravanan,M., Bujnicki,J.M., Cymerman,I.A., Rao,D.N. and Nagaraja,V. (2004) Type II restriction endonuclease R.KpnI is a member of the HNH nuclease superfamily. *Nucleic Acids Res*., **32**, 6129–6135.
33. VanRoey,P., Meehan,L., Kowalski,J.C., Belfort,M. and Derbyshire,V. (2002) Catalytic domain structure and hypothesis for function of GIY-YIG intron endonuclease I-TevI. *Nat. Struct. Biol*., **9**, 806–811.
34. Shen,B.W., Heiter,D.F., Chan,S.H., Wang,H., Xu,S.Y., Morgan,R.D., Wilson,G.G. and Stoddard,B.L. (2010) Unusual target site disruption by the rare-cutting HNH restriction endonuclease PacI. *Structure*, **18**, 734–743.
35. Sokolowska,M., Czapinska,H. and Bochtler,M. (2009) Crystal structure of the beta beta alpha-Me type II restriction endonuclease Hpy99I with target DNA. *Nucleic Acids Res*., **37**, 3799–3810.
36. Van Houten,B., Croteau,D.L., DellaVecchia,M.J., Wang,H. and Kisker,C. (2005) 'Close-fitting sleeves': DNA damage recognition by the UvrABC nuclease system. *Mutat. Res*., **577**, 92–117.
37. Truglio,J.J., Rhau,B., Croteau,D.L., Wang,L., Skorvaga,M., Karakas,E., DellaVecchia,M.J., Wang,H., Van Houten,B. and Kisker,C. (2005) Structural insights into the first incision reaction during nucleotide excision repair. *EMBO J*., **24**, 885–894.
38. Polosina,Y.Y. and Cupples,C.G. (2010) MutL: conducting the cell's response to mismatched and misaligned DNA. *Bioessays*, **32**, 51–59.
39. Polosina,Y.Y., Mui,J., Pitsikas,P. and Cupples,C.G. (2009) The *Escherichia coli* mismatch repair protein MutL recruits the Vsr and MutH endonucleases in response to DNA damage. *J. Bacteriol*., **191**, 4041–4043.
40. Dassa,B., London,N., Stoddard,B.L., Schueler-Furman,O. and Pietrokovski,S. (2009) Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. *Nucleic Acids Res*., **37**, 2560–2573.
41. Lilley,D.M. (2010) The interaction of four-way DNA junctions with resolving enzymes. *Biochem. Soc. Trans*., **38**, 399–403.
42. Tsutakawa,S.E., Jingami,H. and Morikawa,K. (1999) Recognition of a TG mismatch: the crystal structure of very short patch repair endonuclease in complex with a DNA duplex. *Cell*, **99**, 615–623.
43. Delahodde,A., Goguel,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.
44. 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.
45. Goguel,V., Delahodde,A. and Jacq,C. (1992) Connections between RNA splicing and DNA intron mobility in yeast mitochondria: RNA maturase and DNA endonuclease switching experiments. *Mol. Cell. Biol*., **12**, 696–705.
46. Henke,R.M., Butow,R.A. and Perlman,P.S. (1995) Maturase and endonuclease functions depend on separate conserved domains of the bifunctional protein encoded by the group I intron aI4 alpha of yeast mitochondrial DNA. *EMBO J*., **14**, 5094–5099.
47. Szczepanek,T. and Lazowska,J. (1996) Replacement of two non-adjacent amino acids in the S. cerevisiae bi2 intron-encoded RNA maturase is sufficient to gain a homing-endonuclease activity. *EMBO J*., **15**, 3758–3767.

48. Ho,Y., Kim,S.J. and Waring,R.B. (1997) A protein encoded by a group I intron in Aspergillus nidulans directly assists RNA splicing and is a DNA endonuclease. *Proc. Natl Acad. Sci. USA*, **94**, 8994–8999.
49. Geese,W.J., Kwon,Y.K., Wen,X. and Waring,R.B. (2003) In vitro analysis of the relationship between endonuclease and maturase activities in the bi-functional group I intron-encoded protein, I-AniI. *Eur. J. Biochem.*, **270**, 1543–1554.
50. Longo,A., Leonard,C.W., Bassi,G.S., Berndt,D., Krahn,J.M., Hall,T.M. and Weeks,K.M. (2005) Evolution from DNA to RNA recognition by the bI3 LAGLIDADG maturase. *Nat. Struct. Mol. Biol.*, **12**, 779–787.
51. Chatterjee,P., Brady,K.L., Solem,A., Ho,Y. and Caprara,M.G. (2003) Functionally distinct nucleic acid binding sites for a group I intron encoded RNA maturase/DNA homing endonuclease. *J. Mol. Biol.*, **329**, 239–251.
52. Jin,Y., Binkowski,G., Simon,L.D. and Norris,D. (1997) Ho endonuclease cleaves MAT DNA in vitro by an inefficient stoichiometric reaction mechanism. *J. Biol. Chem.*, **272**, 7352–7359.
53. Liu,Q., Derbyshire,V., Belfort,M. and Edgell,D.R. (2006) Distance determination by GIY-YIG intron endonucleases: discrimination between repression and cleavage functions. *Nucleic Acids Res.*, **34**, 1755–1764.
54. Edgell,D.R., Derbyshire,V., Van Roey,P., LaBonne,S., Stanger,M.J., Li,Z., Boyd,T.M., Shub,D.A. and Belfort,M. (2004) Intron-encoded homing endonuclease I-TevI also functions as a transcriptional autorepressor. *Nat. Struct. Mol. Biol.*, **11**, 936–944.
55. Mitchell,P.J. and Tjian,R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, **245**, 371–378.
56. Ainsa,J.A., Ryding,N.J., Hartley,N., Findlay,K.C., Bruton,C.J. and Chater,K.F. (2000) WhiA, a protein of unknown function conserved among gram-positive bacteria, is essential for sporulation in Streptomyces coelicolor A3(2). *J. Bacteriol.*, **182**, 5470–5478.
57. Knizewski,L. and Ginalski,K. (2007) Bacterial DUF199/COG1481 proteins including sporulation regulator WhiA are distant homologs of LAGLIDADG homing endonucleases that retained only DNA binding. *Cell Cycle*, **6**, 1666–1670.
58. Kaiser,B.K., Clifton,M.C., Shen,B.W. and Stoddard,B.L. (2009) The structure of a bacterial DUF199/WhiA protein: domestication of an invasive endonuclease. *Structure*, **17**, 1368–1376.
59. Kaiser,B.K. and Stoddard,B.L. (2011) DNA recognition and transcriptional regulation by the WhiA sporulation factor. *Sci. Rep.*, **1**, 156.
60. Silva,G.H., Dalgaard,J.Z., Belfort,M. and Van Roey,P. (1999) Crystal structure of the thermostable archaeal intron-encoded endonuclease I-DmoI. *J. Mol. Biol.*, **286**, 1123–1136.
61. Heldin,C.H., Miyazono,K. and ten Dijke,P. (1997) TGF-beta signalling from cell membrane to nucleus through SMAD proteins. *Nature*, **390**, 465–471.
62. Grishin,N.V. (2001) Mh1 domain of Smad is a degraded homing endonuclease. *J. Mol. Biol.*, **307**, 31–37.
63. Kim,K.M. and Caetano-Anolles,G. (2010) Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. *Mol. Biol. Evol.*, **27**, 1710–1733.
64. Shub,D.A. (2003) Q & A. *Curr. Biol.*, **13**, R858–R859.
65. Daubin,V. and Ochman,H. (2004) Start-up entities in the origin of new genes. *Curr. Opin. Genet. Dev.*, **14**, 616–619.
66. Edgell,D.R., Gibb,E.A. and Belfort,M. (2010) Mobile DNA elements in T4 and related phages. *Virol. J.*, **7**, 290.