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

Jacob Z. Dalgaard*, Amar J. Klar, Michael J. Moser1, William R. Holley1, Alok Chatterjee1 and I. Saira Mian1

NCI-Frederick Cancer Research and Development Center, ABL-Basic Research Program, PO Box B, Building 549, Room 154, Frederick, MD 21702-1202, USA and 1Life Sciences Division (Mail Stop 29-100), Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

Received July 9, 1997; Revised and Accepted October 3, 1997

ABSTRACT

The LAGLIDADG and HNH families of site-specific DNA endonucleases encoded by viruses, bacteriophages as well as archaeal, eucaryotic nuclear and organellar genomes are characterized by the sequence motifs ‘LAGLIDADG’ and ‘HNH’, respectively. These endonucleases have been shown to occur in different environments: LAGLIDADG endonucleases are found in inteins, archaeal and group I introns and as free standing open reading frames (ORFs); HNH endonucleases occur in group I and group II introns and as ORFs. Here, statistical models (hidden Markov models, HMMs) that encompass both the conserved motifs and more variable regions of these families have been created and employed to characterize known and potential new family members. A number of new, putative LAGLIDADG and HNH endonucleases have been identified including an intein-encoded HNH sequence. Analysis of an HMM-generated multiple alignment of 130 LAGLIDADG family members and the three-dimensional structure of the I-CreI endonuclease has enabled definition of the core elements of the repeated domain (∼90 residues) that is present in this family of proteins. A conserved negatively charged residue is proposed to be involved in catalysis. Phylogenetic analysis of the two families indicates a lack of exchange of endonucleases between different mobile elements (environments) and between hosts from different phylogenetic kingdoms. However, there does appear to have been considerable exchange of endonuclease domains amongst elements of the same type. Such events are suggested to be important for the formation of elements of new specificity.

INTRODUCTION

Three different types of insertional elements have been shown to be mobile via a similar mechanism (1–4). These elements are inteins, group I introns and archaeal introns and their mobility is termed homing because each element is specific for a particular gene (5). Homing occurs when two genomes are juxtaposed but only one possesses the mobile element. The only activity the element provides is a DNA site-specific endonuclease capable of recognizing and cleaving the intein/intron allele of the gene. For example, the single group I intron (OMEGA) in the mitochondrial rrnL gene of Saccharomyces cerevisiae encodes a site-specific DNA endonuclease (I-SceI) capable of recognizing and cleaving the intron−large subunit (LSU) rRNA gene at the insertion position (6,7). Repair of the double-stranded (ds) break by cellular enzymes via a recombination event employing the intein/intron allele as donor results in the element being copied to the recipient genome. Although the aforementioned elements are mobile by the same mechanism, they are unrelated with respect to their splicing mechanism: inteins are spliced at the protein level by an auto-catalytic reaction, group I introns are self splicing and archaeal introns are spliced by cellular enzyme(s) (8). However, most of the site-specific endonucleases they encode are related (4).

The founding members of the largest family of site-specific endonucleases are mitochondrial group I intron-encoded proteins possessing two copies of the conserved amino acid motif LAGLIDADG (9). Additional members of this LAGLIDADG family have been identified as open reading frames (ORFs) as well as being encoded by inteins, group I introns and archaeal introns (4). The precise function of the LAGLIDADG motif is unknown. Mutation of the first and second aspartic acid (Asp) residues abolishes the endonuclease activity of PI-SceI and PI-TliI, respectively (10,11). Furthermore, substrate binding of the mutated PI-SceI is unaffected suggesting these Asp residues are involved in catalysis (10). Purification and characterization of several members indicates that the only co-factor necessary for the enzymes is magnesium (Mg2+) ions (12). Several studies have characterized the interaction between these enzymes and their substrates. All the enzymes have long recognition sites (15–30 bp) that are cleaved in a central position leading to a 4 bp 3′-overhang (3,12–15). Mutational analysis of the recognition sites has shown that the enzymes can tolerate variation of the recognition sequence (7,15–18). Footprinting of I-SceI, I-DmoI and PI-SceI on their substrates show that they make major and minor groove...
MATERIALS AND METHODS

Hidden Markov models

Using known LAGLIDAGG and HNH family members, the BLAST suite of programs (47) were run with default parameters and a merged, non-redundant collection of sequences derived from PIR, SwissProt and translated GenBank. Database sequences were considered to exhibit a statistically significant similarity to the query if smallest sum probability $P(N) \leq 0.05, P(N)$ being the lowest probability ascribed to any set of high scoring segment pairs for each database sequence. HMMs were trained for the LAGLIDAGG and HNH families by the procedure outlined below and used subsequently for phylogenetic studies. Efforts were made to ensure training resulted in HMMs capable of yielding alignments such that known enzymatic elements aligned. A similar approach to that employed here has been used to model other protein domains (36,37,48–51).

For each family, HMM was created using the SAM (Sequence Alignment and Modeling Software System) suite running on a MASPAPR MP-2204 with a DEC Alpha 3000/300X in front end at the University of California Santa Cruz (UCSC). A more detailed description of the HMMs trained and used here can be obtained elsewhere (31,32). HMMs may be viewed as profiles recast within a probabilistic framework and consist of a series of nodes corresponding to columns in a multiple sequence alignment for a set of sequences. The architecture of the HMM captures most of the features of a family of related sequences. In an HMM, use of a match state indicates that a sequence has a residue in that column whereas using a delete state denotes that the sequence does not. Insert states allow sequences to have additional residues between columns and represent regions of the sequence that are not part of the core elements of the family being modelled. To improve the ability of the HMM to generalize, to fit sequences not employed for training, Dirichlet mixture priors (53,54) were employed. Free Insertion Modules (FIMs) were utilized at the beginning and end of the HMM to allow an arbitrary number of insertions at either end to accomodate family members that occurred as domains within larger sequences.

The starting training set of BLAST-derived sequences for the LAGLIDAGG family ranged in length from ~200 to 300 residues. Inspection of initial HMM-generated alignments for the LAGLIDAGG family indicated the emergence of conserved regions in addition to the LAGLIDAGG motifs. Furthermore, the training set sequences appeared to be comprised of a tandem duplication of a domain ~90–100 residues long with each domain containing a copy of the LAGLIDAGG motif near its N-terminus. Differences in length between sequences could be accounted for by the presence of a region of variable length between the two domains. Therefore, an internal FIM was employed to accomodate an insertion at this position during subsequent rounds of HMM training. This internal FIM demarcates the boundary between the first (P1) and second (P2) LAGLIDAGG motif containing domains.

Any sequence can be compared to an HMM by calculating the likelihood that the sequence was generated by that model. Taking the negative (natural) logarithm of this likelihood gives the NLL score. For sequences of equal length, the NLL scores measures how ‘far’ they are from the model and can be used to select sequences that are from the same family. To assess the specificity and sensitivity of an HMM, it can be used in database discrimination experiments to distinguish between sequences that
Table 1. List of the LAGLIDADG family members (Mj ORF4, Mj ORF5 and Mj ORF6 are new members identified in this work)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Family name</th>
<th>Origin</th>
<th>Abbreviation</th>
<th>Species</th>
<th>Protein name</th>
<th>Database code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mj ORF4</td>
<td>Mj ORF4</td>
<td>Mitochondria</td>
<td>Mj ORF4</td>
<td>Melilotus japonicus</td>
<td>ATPase subunit 6</td>
<td>[MjATP6]</td>
</tr>
<tr>
<td>Mj ORF5</td>
<td>Mj ORF5</td>
<td>Mitochondria</td>
<td>Mj ORF5</td>
<td>Melilotus japonicus</td>
<td>Cytochrome b</td>
<td>[MjCYTB]</td>
</tr>
<tr>
<td>Mj ORF6</td>
<td>Mj ORF6</td>
<td>Mitochondria</td>
<td>Mj ORF6</td>
<td>Melilotus japonicus</td>
<td>Cytochrome oxidase subunit I</td>
<td>[MjCOX1]</td>
</tr>
</tbody>
</table>

Sequences are grouped according to their origin. For each sequence, its abbreviation, the species name and the protein name are given together with the database code in ‘[ ]’. * denotes proteins where the enzymatic activity has been characterized and whose enzymatic name is given in the third column.

Other abbreviations are as follows. .c, chloroplast; .m, mitochondria; atp6, ATPase subunit 6; cob, cytochrome b; cox1, cytochrome oxidase subunit I; cox2, cytochrome oxidase subunit II; cox3, cytochrome oxidase subunit III; cytb, apocytochrome b; nad1, NADH dehydrogenase subunit 1; nad3, NADH dehydrogenase subunit 3; nad4, NADH dehydrogenase subunit 4; nad5, NADH dehydrogenase subunit 5; rRNA, ribosomal RNA; LSU, large subunit; SSU, small subunit.

This cycle of ‘search, align and retrain’ was repeated until no new sequences were identified in databases up to 2014. A database search was performed and based upon examination of the log-odds scores and an HMM-generated alignment, new family members were identified, added to the training set and the HMM retrained.

The sequences are grouped according to their origin. For each sequence, its abbreviation, the species name and the protein name are given together with the database code in ‘[ ]’. * denotes proteins where the enzymatic activity has been characterized and whose enzymatic name is given in the third column.

Other abbreviations are as follows. .c, chloroplast; .m, mitochondria; atp6, ATPase subunit 6; cob, cytochrome b; cox1, cytochrome oxidase subunit I; cox2, cytochrome oxidase subunit II; cox3, cytochrome oxidase subunit III; cytb, apocytochrome b; nad1, NADH dehydrogenase subunit 1; nad3, NADH dehydrogenase subunit 3; nad4, NADH dehydrogenase subunit 4; nad5, NADH dehydrogenase subunit 5; rRNA, ribosomal RNA; LSU, large subunit; SSU, small subunit.

The log-odds (NLL-NULL) (55,56) scores for all sequences in a non-redundant protein database obtained from the NCI (57) and updated weekly at UCSC. The significance of log-odds scores can be ascertained by evaluating E, the expected number of false positives above a given log-odds score in a given database search. However, since the NULL model does not consider the score distribution for all ‘random’ sequences, the E value calculated by SAM is not a true estimate of E but represents an upper bound. Taking into account the number of sequences in this database (~230 000 different proteins in early 1997) and an expected number of false positives of 0,01, a significant log-odds score is 22.6. Scores higher than this value denote fewer false positives.

A database search was performed and based upon examination of the log-odds scores and an HMM-generated alignment, new family members were identified, added to the training set and the HMM retrained. This cycle of ‘search, align and retrain’ was repeated until no new sequences were identified in databases up
to January 1997. Multiple models were trained and the final (best) retained for further study. As a consequence of the problems in calculating a true estimate for $E$, the approach employed here emphasises training an HMM that discriminates between training and non-training set sequences, i.e., one in which the gap in log-odds scores between the lowest scoring training set sequence and the highest scoring non-training set (database) sequence is relatively large (usually $>5.0$) and the absolute log-odds score for the lowest training set sequence is $>22.6$ ($E = 0.01$).

Figures showing multiple sequence alignments, phylogenetic trees and ribbon diagrams of molecules were produced using ALSCRIPT (58), Treetool (59) and MOLSCRIPT (60), respectively.

**Phylogenetic analysis**

HMM-generated multiple sequence alignments of the training sets were utilized as the starting points for phylogenetic studies. The alignments only contained match and delete states and insertions (including the FIMs) were excluded. Insert states are not modelled by an HMM and because the regions in a sequence they represent are the most divergent parts of the molecules, they are likely to be
sources of systematic error. The MOLPHY suite uses a probabilistic procedure for inferring phylogenetic relationships (61,62). PROTML, the main program in MOLPHY, infers evolutionary trees from amino acid sequences by means of a maximum likelihood method. The star decomposition algorithm of PROTML 2.3 and the default JTT model was used to determine automatically an initial tree from an HMM-generated multiple alignment. Starting from this tree, repeated local rearrangements were employed to search for better topologies. Amongst these final trees, the one with the highest likelihood was selected. Local bootstrap probabilities (LBPs) for branches in the final tree indicate the bootstrap probability of that branch when the other parts of the tree are correct. Because of the large number of LAGLIDADG family members from all environments (130 in total), it was not possible to generate an initial tree using the Star Decomposition algorithm. Instead, a maximum likelihood distance matrix was calculated and NJdist (neighbour joining) used to compute a tree which was then subjected to local rearrangement as described earlier.
A primary aim of this study was to create and use a specific and sensitive HMM for the LAGLIDADG family. This involved training an HMM that minimized the number of false positives (sequences incorrectly identified by the HMM as belonging to the family) and false negatives (sequences not identified by the HMM as belonging to the family). Table 1 lists the LAGLIDADG family members used to train the HMM. Of ~230 000 sequences in the final non-redundant protein database searched using the HMM, only these training sequences had log-odds scores >88.0. The next highest scoring sequence (47.0) was a fragment of the group 1 intron sequence Sp.m_cox1_2 in Table 1 (databank code A25568). Each of the subsequent highest scoring sequences appeared to contain a single copy of the repeated domain. These sequences, which were excluded from the training set, are *Acanthamoeba castellanii* mitochondrial LSU rRNA intron protein ymf46 (log-odds score 43.0, databank code S46445); *Prototheca wickerhamii* mitochondrial cox1 intron ORF ymf44 (42.6, PWU02970); *A.castellanii* mitochondrial LSU rRNA intron protein ymf48 (37.9, S46447); *Chlamydomonas palidostigmatica* chloroplast LSU rRNA intron protein (37.0, CRECPRN12); *A.castellanii* mitochondrial LSU rRNA intron protein ymf47 (35.4, S46446); *Plasmodium falciparum* plastid-like DNA Clp protein which exhibits some similarity to Sd.m_cob_3 in Table 1 (33.8, PFCOMPIRB); *Chlamydomonas eugametos* LSU rRNA intron 1 protein (site-specific DNA endonuclease I-CreI) (31.6, DNEI_CHLEU) (63). The remaining sequences all had log-odds scores <29.6 and included *Chlamydomonas reinhardtii* site-specific DNA endonuclease I-CreI (23.7, DNEI_CHLRE) (64).

There may be potential LAGLIDADG family members or closely related sequences amongst sequences with log-odds scores >29.6 but these false negatives would have diverged from the training set used here to a degree that the current HMM is too specific and thus unable to classify them as belonging to the family. Here, sequences with log-odds scores >47.0 are classified as belonging to the LAGLIDADG family and consist of those listed in Table 1. New members identified in this work are three free-standing archaeal ORFs Mj.ORF4, Mj.ORF5 and Mj.ORF6. Figure 1 shows an HMM-generated alignment of members of the LAGLIDADG family and other data. Although the HMM was trained without knowledge of, or reference to, the
Figure 2. (Opposite) Phylogenetic tree for the LAGLIDADG family based upon the alignment present in the electronic appendix (Fig. 1 shows an alignment of some of these sequences). Intein-encoded, free standing ORFs, archaeal intron-encoded and group I intron-encoded sequences are in red, blue, green and black respectively. + denotes proteins whose enzymatic activity has been characterized (sequences marked with ¥ in Table 1). The 12 intron-encoded endonucleases from the mitochondrial cox1 gene of Podospora anserina are shown in a different (black) font. For group I intron-encoded sequences where data are available, the host intron subtype (IA1, IB1, IB2, IC2 or ID) and the location of the endonuclease within the catalytic core (position A, B, C, D or E) as defined elsewhere (67) are given in parenthesis. A schematic diagram of the secondary structure of group I introns (67) is shown above. Exons are depicted as black boxes, P1–P9 denote the various helical elements and A–E denote the location of the endonuclease in the catalytic core.

Figure 3. Comparison of the phylogenetic trees for the LAGLIDADG family endonuclease (left) and protein splicing (right) domains of inteins. Sequences in the three major branches of the left hand tree are colored red, green and blue and this scheme is used to color sequences in the other tree. The sequences are listed in Table 1. The trees are based upon the alignment shown in Figure 1 and an alignment of the protein splicing domain of inteins taken from elsewhere (37).

Overall, there is considerable divergence amongst the sequences: only 8% of the positions (16 out of 193) are highly conserved (positions labelled B, D, E, F, H, I, J, K, a, b, c, e, f, k, l, m in Fig. 1). These highly conserved positions include the LAGLIDADG motifs (B–F/b–f). Position E/e corresponds to a functionally important Asp residue (10,11). Position I/i is located in a negatively charged region defined elsewhere (9). There are several conserved hydrophobic positions (boxed). Whilst the gross features of the alignment such as the locations of conserved regions are unlikely to change, further refinement of the HMM and inclusion of LAGLIDADG sequences that have not yet been deposited in the databanks are likely to revise and improve the detailed aspects of the model as well as identify new family members. The HMM-generated alignment represents the current best estimate for the features that characterize this family.

Three-dimensional structure of either PI-SceI or I-CreI, it has the capacity to model the core elements of the family because insertions are generally confined to regions between secondary structure elements. Examination of the alignment indicates that the four BLOCKS labelled C, D, E and H that have been described elsewhere (65) characterize only some of the conserved regions in the family. In contrast, the HMM describes both the variable and conserved regions of the complete P1 and P2 repeated domains.

Overall, there is considerable divergence amongst the sequences: only 8% of the positions (16 out of 193) are highly conserved (positions labelled B, D, E, F, H, I, J, K, a, b, c, e, f, k, l, m in Fig. 1). These highly conserved positions include the LAGLIDADG motifs (B–F/b–f). Position E/e corresponds to a functionally important Asp residue (10,11). Position I/i is located in a negatively charged region defined elsewhere (9). There are several conserved hydrophobic positions (boxed). Whilst the gross features of the alignment such as the locations of conserved regions are unlikely to change, further refinement of the HMM and inclusion of LAGLIDADG sequences that have not yet been deposited in the databanks are likely to revise and improve the detailed aspects of the model as well as identify new family members. The HMM-generated alignment represents the current best estimate for the features that characterize this family.

Three-dimensional structure of either PI-SceI or I-CreI, it has the capacity to model the core elements of the family because insertions are generally confined to regions between secondary structure elements. Examination of the alignment indicates that the four BLOCKS labelled C, D, E and H that have been described elsewhere (65) characterize only some of the conserved regions in the family. In contrast, the HMM describes both the variable and conserved regions of the complete P1 and P2 repeated domains.
different types of elements. The branching pattern suggests that endonuclease domains are unlikely to have been exchanged between different classes of host elements. There is no strong evidence, either, for exchange of elements between hosts from different phylogenetic kingdoms: intein-encoded endonucleases generally branch according to their origin (bacterial, eukaryotic and archaeal); archaeal intron-encoded endonucleases cluster together and group I intron sequences (black) branch according to whether they are chloroplast or mitochondrial (indicated by .c or .m). However, phylogenetic analysis does provide strong support for frequent transposition of the elements during evolution. Transposition is the process whereby elements invade new positions/host genes in the genome. Support for transposition comes from the absence of a correlation between the branching pattern of the endonucleases and the host genomes. For example, the lowest cluster of related group I intron-encoded endonuclease in Figure 2 are present in an array of unrelated and distantly related genes (LSU and SSU rRNA, atp6, nad1, nad3, nad4L, nad5). The simplest explanation for this observation is that transposition occurred frequently during evolution.

It has been suggested that mobile group I introns and inteins arose by invasion of a site-specific endonuclease into a self-splicing intron or intein (28:SP .p_gyrB) (37). This is the first report of an intein that encodes an endonuclease of the HNH family and verifies the presence of a member in a bacterial phylogenetic domain. A correlation between the two trees would have suggested that the two domains were combined only once during evolution. However, several major rearrangements in the branching pattern are present suggesting that the LAGLIDADG domains have been shuffled between inteins during evolution also.

**HNH family**

The strongest support for mobile group I introns having arisen several times during evolution by acquisition of site-specific DNA endonucleases comes from the observation that they encode endonucleases belonging to several different families, the two most common being the LAGLIDADG and HNH families (4). The result presented next show that a putative intein identified in an earlier work (37) encodes an endonuclease of the HNH family and consists of those listed in Table 2. There may be HNH members amongst sequences with log-odds scores <22.6 but these false negatives may have diverged to a degree that the current HMM is too specific and thus unable to classify them as belonging to the family. Figure 4 shows an alignment of the HNH family and verifies the presence of a member in a bacterial intein (28:SP_p_gyrB) (37). This is the first report of an intein that does not encode an endonuclease of the LAGLIDADG family. This observation shows that inteins encode endonucleases belonging to at least two families and supports the suggestion that mobile inteins evolved by invasion of a protein splicing domain by a site-specific endonuclease (37,43).

A number of new HNH family members have been identified here and include several bacteriophage-encoded proteins, some of which are site-specific DNA endonucleases involved in packaging, as well as a bacterial enzyme (AP_adx) involved in a...
Figure 4. An HMM-generated multiple sequence alignment of HNH family listed in Table 2 with new members identified in this work shown in a different font. Amino acids conserved in the majority of the sequences are highlighted and columns that are predominantly hydrophobic are boxed. Columns containing ‘.’ correspond to insert states and numbers indicate the lengths of insertions in sequences at that position (if present).

Figure 5 shows a phylogenetic tree for the sequences shown in Table 2. The enzymes are present in bacteria, mitochondria, chloroplasts, a virus, bacteriophages and a plasmid and are either free standing ORFs or are encoded by a transposon, group I or II introns and an intein. Thus, HNH family members can be one domain of a multifunctional enzyme or form the complete protein. As with the LAGLIDADG family, the HNH tree (Fig. 5) indicates a lack of correlation between the branching pattern of these enzymes and the cellular function and host suggesting a high degree of transposition/genetic mobility of these endonucleases during evolution.

DISCUSSION

This study has focused on a divergent family of proteins that occurs in all three phylogenetic kingdoms as well as organelles and whose members are intein-encoded, free standing ORFs, archaeal intron-encoded and group I intron-encoded. A statistical model, an HMM, was trained that captured the core elements of this LAGLIDADG family and identified several new members amongst the 130 sequences characterized as belonging to this family. Analysis of an HMM-generated alignment and the three-dimensional structures of PI-SceI (43) and I-CreI (45) support an earlier suggestion that the LAGLIDADG family is comprised of a repeated domain (22,69). These domains, termed P1 and P2, are conserved at the level of both primary sequence and structure. Whilst I-CreI only possesses one LAGLIDADG motif and acts as a homodimer, PI-SceI is a monomer containing two domains whose overall structure is similar structure to each I-CreI monomer.

Figure 6 shows the highly conserved residues present in the alignment of 130 LAGLIDADG family members mapped onto the three-dimensional structure of I-CreI (residues in bold in Fig. 1 and labelled A–M and a–m). Comparison of the structure and the alignment indicates that the α1-β1-β2-α2-β3-β4-α3 region of I-CreI comprises the core of the ~90 residue long repeated domains (P1 or P2) common to this family of proteins. The alignment shows that P1 and P2 are separated by a linker region that varies in length from zero (46:Po_LSU_2) to 108 residues (6:SP_polIII). In I-CreI, the β1-β2 and β3-β4 loops have been suggested to make sequence-specific interactions with the major groove of DNA (45). In the LAGLIDADG family, these loops are the regions of the P1/P2 core that exhibit the greatest variation in terms of sequence length (0–49 and 1–28 residues) as well as low sequence conservation. The proposition here that the β1-β2 and β3-β4 loops may generally be important in substrate recognition is supported by data which show that they are protected from protease digestion by substrate binding in 46:Po_LSU/I-PorI and 40:Dm_LSU/I-DmoI (+ in Fig. 1) (70).

The majority of the highly conserved residues in Figure 1 appear to be important largely for the hydrophobic core of P1 or P2 (A, C, G, J, H) and as potential signals for the generation of specific secondary structure elements (K). The relative organization of the repeated domains in the monomeric LAGLIDADG family members examined here is likely to be similar to that of the two monomers in the LAGLIDADG motif containing endonucleases which act as dimers. In I-CreI, the first seven residues of the LAGLIDADG motifs that include the conserved positions B and D are involved in formation of the dimer interface whilst the last two residues are believed to be involved in formation of the active site (45). Like I-CreI where B and D are Gly and Ala, the LAGLIDADG members with two domains also possess similar small amino acids suggesting that these residues play a similar role in the interaction between the two repeated domains of the monomers and may be crucial in the formation or positioning of the active site(s). Although P1 and P2 are likely to be similar in
Figure 5. Phylogenetic tree for the HNH family members listed in Table 2 and based upon the alignment shown in Figure 4. Intein-encoded, ORF and intron-encoded endonucleases are coloured green, red and blue respectively. New endonucleases identified in this work are shown in an italic font.

terms of structure and function, subtle differences may be important for activity (see for example, endonucleases that have tryptophan at $c$ in Figure 1 and which branch with intein-encoded endonucleases in Figure 2).

The frequent occurrence of one or more negatively charged residues in the $\beta_{2}-\alpha_{2}$ loop, most notably position $i/i$ in Figure 1, may provide some insight into the catalytic mechanism of the LAGLIDADG endonucleases. In a model for the interaction between I-CreI and its substrate, the $\beta_{2}-\alpha_{2}$ loop is proposed to be in close proximity to the phosphate backbone at the position where cleavage is expected to occur (45). Therefore, it is possible that position $i/i$ could be involved in catalysis. In conjunction with the acidic residue of the LAGLIDADG motif (E/e), positions $i/i$ could each be involved in the formation of a single Mg$^{2+}$ binding site. If this is the case, then the enzyme would have two metal binding sites that would form two active sites capable of cleaving the two strands as has been suggested for EcoRV (71,72). Data supporting this model come from the observation that several LAGLIDADG endonucleases cleave only one strand of the substrate at low Mg$^{2+}$ concentrations (19,73). This model for catalysis differs substantially from that of Gimble and colleagues (10,43) who suggest that the enzyme only has one active site that catalyzes the cleavage of both strands. It should be noted that the residue proposed to be involved in stabilizing the doubly charged pentavalent transition state in PI-SceI (43), Lys 301 (column 94 in Fig. 1), exhibits only limited conservation amongst the 130 LAGLIDADG sequences.

The results here present an opportunity to address the relationship between endonucleases encoded by different classes of elements. The correlation between branching pattern and sequence origin suggests limited or no exchange of endonucleases between different elements and between hosts belonging to different kingdoms.
However, the lack of correlation between host genes and branching pattern suggests a substantial loss of mobile elements over time and that transposition to new positions has occurred on many occasions during evolution.

Comparison of the phylogenetic relationships between host elements and the endonucleases leads us to propose that the formation of elements of altered specificity and transposition might involve shuffling of endonuclease domains between related elements. Such shuffling events seem to have occurred several times during evolution and could be the result of heterologous recombination events. Although such events would be expected to be rare, the propagation of a successfully created element of altered specificity would be ensured by its mobility. This hypothesis is also supported by the observation here of an intein encoding an endonuclease of the HNH family (41,46).

Although several families of endonucleases are encoded by group I introns, this is the first example of an intein encoding an endonuclease not belonging to the LAGLIDADG family. The existences of such an intein and of inteins that lack any site-specific endonuclease domain (37,65,74) supports the theory that the protein splicing and endonuclease domains of inteins are of different evolutionary origins (37,43). It remains to be seen whether endonucleases other than those belonging to the families studied here or other domains are encoded by inteins.

The focus here has been on LAGLIDADG and HNH family members that are homing endonucleases encoded by inteins, group I introns and archaeal introns. It should be emphasised that these two families include members from all three phylogenetic kingdoms, organelles, viruses, bacteriophages, plasmids and transposons. Furthermore, these endonucleases are involved in an array of cellular processes such as homing of site-specific elements including inteins and archaeal and group I intron; retrotransposition of group II introns; induction of recombination in mitochondria; differentiation controlled DNA rearrangements in bacteria and eucarya; phage packaging and bacterial toxins. This broad spectrum of hosts and functions and the phylogenetic evidence for their genetic mobility, shuffling and evolution of de novo functions highlights the important roles endonucleases have played in the evolutionary processes that have shaped both proteins and organisms.

ACKNOWLEDGEMENTS

We thank Barry Stoddard for providing us with the coordinates of I-CreI and our colleagues at UCSC for use of computer hardware and software. This work was supported by the Danish Natural Science Research Council (J.Z.D); the National Cancer Institute, DHHS, with ABL (J.Z.D, A.K.); National Science Foundation grant DBI-9408579 (W.R.H.) and the Director, Office of Energy Research, Office of Biological and Environmental Research, Division of the US Department of Energy under Contract No. DE-AC03-76F00098 (M.J.M, W.R.H, A.C., I.S.M.). The data and multiple alignments are available in electronic form upon request.

REFERENCES
