4626–4638 *Nucleic Acids Research, 1997, Vol. 25, No. 22*

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. Moser[1], William R. Holley[1], Aloke Chatterjee[1] and I. Saira Mian[1]

NCI-Frederick Cancer Research and Development Center, ABL-Basic Research Program, PO Box B, Building 549, Room 154, Frederick, MD 21702-1202, USA and [1]Life 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-*Cre*I 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 specficity.

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-*Sce*I) 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-*Sce*I and PI-*Tli*I, respectively (10,11). Furthermore, substrate binding of the mutated PI-*Sce*I 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 (Mg^{2+}) 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-*Sce*I, I-*Dmo*I and PI-*Sce*I on their substrates show that they make major and minor groove

*To whom correspondence should be addressed. Tel: +1 301 846 5149; Fax: +1 301 846 6911; Email: dalgaard@ncifcrf.gov

interactions (19–21). Moreover, I-*Sce*I remains bound to one of the two products after cleavage suggesting that it might have additional roles in the recombination event that leads to intron/intein mobility (19). Several regions of two other LAGLIDADG family members, I-*Dmo*I and I-*Por*I, have been identified that are protected from proteases by substrate binding (22).

Some mitochondrial group I intron-encoded LAGLIDADG family members function or have an additional function as maturases (14,23–25). The maturase activity catalyzes folding of the self-splicing intron and may have evolved from the site-specific endonuclease (25). Maturase and site-specific endonuclease activities appear to be separable functions. In I-*Sce*II, mutation of the glycine (Gly) residues in one of the two repeated LAGLIDADG motifs selectively abolishes maturase or endonuclease activity (26).

Hidden Markov models (HMMs) are a statistical modelling method (27–35) that have been applied recently to the problems of characterizing the common features of a family of related sequences, generating a multiple sequence alignment and recognizing related, but divergent sequences present in sequence databases (32,36–42). In previous work (37), a protein splicing domain proposed to be common to inteins and hedgehog proteins was examined using an HMM-based approach. In that study, the endonuclease domain of inteins was not modelled explicitly but was represented simply as an insertion of variable length present at a specific position in the protein splicing domain. In order to gain a more detailed view of the endonuclease domain, a complementary HMM-based study of the endonuclease domain was initiated. During the latter stage of this study, the results of which are presented here, the identification and modelling of inteins as two structurally and functionally distinct domains received experimental support. The three-dimensional structure of the PI-*Sce*I intein is composed of two separate domains (I and II) with different structures and functions (43). The catalytic core of domain I corresponds to the protein splicing domain modelled previously (37) (see also ref. 44) and domain II corresponds to the endonuclease domain examined here. In addition to an intein-encoded LAGLIDADG endonuclease, the three-dimensional structure of the free-standing I-*Cre*I LAGLIDADG endonuclease has been determined by X-ray crystallography (45). The LAGLIDADG endonuclease in PI-*Sce*I forms a compact domain primarily composed of two similar α/β motifs. I-*Cre*I functions as a homodimer whose overall structure is similar to that observed for the LAGLIDADG domain in PI-*Sce*I.

Here, the LAGLIDADG family has been modelled explicitly by training an HMM for this family of endonucleases. An HMM-generated multiple sequence alignment of intein, group I intron, archael intron and free standing ORF family members was utilized for phylogenetic analysis. Potential new family members have been identified and both the common and variable sequence and structural features characterized. Comparison of the alignment of 130 LAGLIDADG family members with the structure of I-*Cre*I has allowed delineation of the essential or core features of the repeated domain present in this family. Another site-specific endonuclease family, the HNH or I-TevIII family, is encoded by group I introns, group II introns as well as numerous other cellular and bacteriophage-encoded enzymes (41,46). Here, the HNH family has been modelled using HMMs and a putative bacterial intein-encoded family member identified, thereby expanding the number and location of elements in this class of endonucleases. Evolutionary implications of the results are discussed.

MATERIALS AND METHODS

Hidden Markov models

Using known LAGLIDADG 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 LAGLIDADG 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 MASPAR MP-2204 with a DEC Alpha 3000/300X frontend at the University of California Santa Cruz (UCSC). A more detailed description of the HMMs trained and used here can be obtained elsewhere (31,52). 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 arbitary 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 LAGLIDADG family ranged in length from ~200 to 300 residues. Inspection of initial HMM-generated alignments for the LAGLIDADG family indicated the emergence of conserved regions in addition to the LAGLIDADG 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 LAGLIDADG 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) LAGLIDADG 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 rnembers (Mj_ORF4, Mj_ORF5 and Mj_ORF6 are new members identified in this work)

Intein-encoded (bacteria, chloroplast, eucarya, archaea)		
Mf_gyrA	*Mycobacterium flavescens*	DNA gyrase subunit A [MFDNAGRA]
Mk_gyrA	*Mycobacterium kansasii*	DNA gyrase subunit A [MKDNAGRA]
Ml_gyrA	*Mycobacterium leprae*	DNA gyrase A subunit [MLDNAGRA]
Ml_pps1	*Mycobacterium leprae*	ORF similar to red algae chloroplast yca3 protein [U00013]
Ml_recA	*Mycobacterium leprae*	recombinase A (recA) [RECA_MYCLE]
‡Mt_recA	*Mycobacterium tuberculosis*	recombinase A (recA) [RECA_MYCTU] ‡PI-*Mtu*I [75]; (E.O. Davis, personal communication)
Mg_gyrA	*Mycobacterium gordonae*	DNA gyrase subunit A [MGDNAGRA]
SP_gyrA	*Synechocystis* sp. PCC6803	DNA gyrase subunit A [SYCSLLE]
SP_polIII	*Synechocystis* sp. PCC6803	DNA polymerase III subunit [D90907]
Ce.c_clpp	*Chlamydomonas eugametos*	choloroplast s14/clpp-like peptidase [CLPP_CHLEU]
Ct_vma1	*Candida tropicalis*	vacuolar ATP synthase [VATA_CANTR]
‡Sc_vma1	*Saccharomyces cerevisiae*	vacuolar ATP synthase [VATA_YEAST] ‡PI-*Sce*I [76, 3]
Mj_arr_1	*Methanococcus jannaschii*	anaerobic ribonucleoside-triphosphate reductase [MJU67527] intein 1
Mj_arr_2	*Methanococcus jannaschii*	anaerobic ribonucleoside-triphosphate reductase [MJU67527] intein 2
Mj_gfpt	*Methanococcus jannaschii*	glutamine-fructose-6-phosphate transaminase [MJU67582]
Mj_pep1	*Methanococcus jannaschii*	ORF MJ0043 similar to P38423 [MJU67462]
Mj_pep2	*Methanococcus jannaschii*	ORF MJ0682 similar to P46850 [MJU67515]
Mj_pep3	*Methanococcus jannaschii*	ORF MJ1124 similar to P32639 [MJU67555]
Mj_polB_1	*Methanococcus jannaschii*	DNA-dependent DNA polymerase family B [MJU67532] intein 1
Mj_polB_2	*Methanococcus jannaschii*	DNA-dependent DNA polymerase family B [MJU67532] intein 2
Mj_ps	*Methanococcus jannaschii*	phosphoenolpyruvate synthase [MJU67503]
Mj_rfC_1	*Methanococcus jannaschii*	replication factor C [MJU67583] intein 1
Mj_rfC_2	*Methanococcus jannaschii*	replication factor C [MJU67583] intein 2
Mj_rfC_3	*Methanococcus jannaschii*	replication factor C [MJU67583] intein 3
Mj_rgyr	*Methanococcus jannaschii*	reverse gyrase [MJU67592]
Mj_rpolA′	*Methanococcus jannaschii*	DNA-dependent RNA polymerase subunit A′ [MJU67547]
Mj_rpolA″	*Methanococcus jannaschii*	DNA-dependent RNA polymerase subunit A″ [MJU67547]
Mj_tfIIB	*Methanococcus jannaschii*	transcription initiation factor IIB [MJU67522]
Mj_tif	*Methanococcus jannaschii*	translation initiation factor (FUN12/bIF-2) [MJU67481]
Mj_ugd	*Methanococcus jannaschii*	UDP-glucose dehydrogenase [MJU67548]
Pf_rnr_1	*Pyrococcus furiosus*	ribonucleotide reductase (rnr) [PFU78098] intein 1
Pf_rnr_2	*Pyrococcus furiosus*	ribonucleotide reductase (rnr) [PFU78098] intein 2
‡PG_pol	*Pyrococcus* sp. GB-D	DNA-dependent DNA polymerase [PSU00707] ‡PI-*Psp*I (F.B. Perler, personal communication)
PK_pol_1	*Pyrococcus* sp. KO	DNA-dependent DNA polymerase [PYWKODPOL]
PK_pol_2	*Pyrococcus* sp. KO	DNA-dependent DNA polymerase [PYWKODPOL]
Tf_pol_1	*Thermococcus fumicolans*	DNA-dependent DNA polymerase [TFDPOLEND] intein 1
Tf_pol_2	*Thermococcus fumicolans*	DNA-dependent DNA polymerase [TFDPOLEND] intein 2
‡Tl_pol_1	*Thermococcus litoralis*	DNA-dependent DNA polymerase [DPOL_THELI] intein 1. ‡PI-*Tli*I [77, 11]
‡Tl_pol_2	*Thermococcus litoralis*	DNA-dependent DNA polymerase [DPOL_THELI] intein 2. ‡PI-*Tli*I [77, 11] (F.B. Perler, personal communication)
Free standing ORFs (mitochdondria, eucarya, archaea)		
Hw.m_ORF	*Hansenula wingei*	ORF g1000989 [D31785, HASMT_18]
Sb.m_RF2	*Saccharomyces bayanus*	RF2 protein [YSCMTRF21, YSCMTRF22]
Sc.m_ORF1	*Saccharomyces cerevisiae*	ORF B386/ORF1 [YM38_YEAST]
Sc.m_RF2	*Saccharomyces cerevisiae*	RF2 protein [B28439]
‡Sc.m_RF3	*Saccharomyces cerevisiae*	RF3 protein [RF3_YEAST] ‡Endo.*Sce*I [78]
Su.m_RF3	*Saccharomyces uvarum*	RF3 protein [RF3_SACUV]
Wm.m_ORF1	*Williopsis mrakii*	ORF [S42735]
Ws.m_ORF1	*Williopsis suaveolens*	ORF [S42738]
Ws.m_ORF3	*Williopsis suaveolens*	ORF [MIWSATP9] (ORF3)
‡Sc_ho	*Saccharomyces cerevisiae*	homothallic switching endonuclease [HO_YEAST] ‡HO endonuclease [79]
Mj_ORF4	*Methanococcus jannaschii*	ORF MJ1098 [MJU67552]
Mj_ORF5	*Methanococcus jannaschii*	ORF MJ0398 [MJU67492]
Mj_ORF6	*Methanococcus jannaschii*	ORF MJ0314 [MJU67486]
Archaeal intron-encoded		
‡Dm_LSU	*Desulfurococcus mobilis*	LSU rRNA [22KD_DESMO] ‡I-*Dmo*I [80]
Pa_SSU	*Pyrobaculum aerophilum*	SSU rRNA [A47399]
‡Po_LSU_1	*Pyrobaculum organotrophum*	LSU rRNA [JC1382] intron ‡I-*Por*I [70]
Po_LSU_2	*Pyrobaculum organotrophum*	LSU rRNA [JC1382] intron 2
Group I intron-encoded (eucarya, chloroplast, mitochondria)		
Tp_LSU	*Trimorphomyces papilionaceus*	LSU rRNA [S34220]
Ce.c_LSU_6	*Chlamydomonas eugametos*	LSU rRNA [S15139] intron 6
‡Ch.c_LSU	*Chlamydomonas humicola*	LSU rRNA [JC1520] ‡I-*Chu*I [81]
Cp.c_SSU_1	*Chlamydomonas pallidostigmatica*	SSU rRNA [S56764]
‡Cp.c_SSU_2	*Chlamydomonas pallidostigmatica*	SSU rRNA [CRECPCPRG] ‡I-*Cpa*II [73]
Ag.m_SSU	*Agrocybe aegerita*	SSU rRNA [AAU54637]
Am.m_cob_2	*Allomyces macrogynus*	cob [AMU41288] intron 2 (ORF246)
Am.m_cob_3	*Allomyces macrogynus*	cob [AMU41288] intron 3 (ORF233)
Am.m_cob_5	*Allomyces macrogynus*	cob [AMU41288] intron 5 (ORF242)
Am.m_cob_6	*Allomyces macrogynus*	cob [AMU41288] intron 6 (ORF361)
Am.m_cox1_8	*Allomyces macrogynus*	cox1 [AMU41288] intron 8 (ORF342)
Am.m_nad5	*Allomyces macrogynus*	nad5 [AMU17010]
Am.m_nad5_1	*Allomyces macrogynus*	nad5 [AMU41288] (ORF261)
Cs.m_cytb	*Chlamydomonas smithii*	cytb [CSAPOCYTB]
En.m_cob_1	*Emericella nidulans*	cob [MBI1_EMENI]
En.m_cox1_2	*Emericella nidulans*	cox1 [C22735]
En.m_cox1_3	*Emericella nidulans*	cox1 [D22735]
Hw.m_cox1	*Hansenula wingei*	cox1 [D31785, HASMT_2] (ORF g1000973)
Km.m_cox1_2	*Kluyveromyces marxianus*	cox1 [S17996] intron 2
Km.m_cox1_3	*Kluyveromyces marxianus*	cox1 [S17997] intron 3
Km.m_cox1_4	*Kluyveromyces marxianus*	cox1 [S17998] intron 4
Kt.m_LSU	*Kluyveromyces thermotolerans*	LSU rRNA [MIKTRRNA]
Mp.m_cox1_4	*Marchantia polymorpha*	cox1 [S25958] intron 4
Mp.m_cox1_8	*Marchantia polymorpha*	cox1 [S25959] intron 8
Ms.m_cox1	*Metridium senile*	cox1 [MSU36783, 1353403] (ORF U36783)
Nc.m_atp6_2	*Neurospora crassa*	atp6 [MINC03]
Nc.m_cob	*Neurospora crassa*	cob [A28755]
Nc.m_cox1_4	*Neurospora crassa*	cox1 [MINCCOIG] intron 4
Nc.m_nad1_1	*Neurospora crassa*	nad1 [S06367] intron 1
Nc.m_nad4L	*Neurospora crassa*	nad4L [S10840] intron 1
Nc.m_nad5_1	*Neurospora crassa*	nad5 [S10841] intron 1
Nc.m_nad5_2	*Neurospora crassa*	nad5 [S10842] intron 2
On.m_LSU	*Ophiostoma novo-ulmi*	LSU rRNA [S55724]
Pa.m_cox1_2	*Podospora anserina*	cox1 [C48327] intron 2
Pa.m_cox1_3	*Podospora anserina*	cox1 [D48327] intron 3
Pa.m_cox1_5	*Podospora anserina*	cox1 [F48327] intron 5
Pa.m_cox1_7a	*Podospora anserina*	cox1 [H48327] intron 7a
Pa.m_cox1_7b	*Podospora anserina*	cox1 [I48327] intron 7b
Pa.m_cox1_8	*Podospora anserina*	cox1 [A38888] intron 8
Pa.m_cox1_9	*Podospora anserina*	cox1 [B38888] intron 9
Pa.m_cox1_10	*Podospora anserina*	cox1 [C38888] intron 10
Pa.m_cox1_11	*Podospora anserina*	cox1 [D38888] intron 11
Pa.m_cox1_12	*Podospora anserina*	cox1 [E38888] intron 12
Pa.m_cox1_13	*Podospora anserina*	cox1 [F38888] intron 13
Pa.m_cox1_15	*Podospora anserina*	cox1 [H38888] intron 15
Pa.m_cox2_2	*Podospora anserina*	cox2 [S09140] intron 2
Pa.m_cytb_1a	*Podospora anserina*	cytb [B48326] intron 1a
Pa.m_cytb_3a	*Podospora anserina*	cytb [E48326] intron 3a
Pa.m_LSU_1	*Podospora anserina*	LSU rRNA [S06606] intron 1
Pa.m_nad1_4	*Podospora anserina*	nad1 [S06059] intron 4 protein 1
Pa.m_nad3_1	*Podospora anserina*	nad3 [YMN3_PODAN] intron 1
Pa.m_nad4_4	*Podospora anserina*	nad4 [YMN4_PODAN] intron 4
Pa.m_nad4L_1	*Podospora anserina*	nad4L [S09134] intron 1
Pa.m_nad4L_2	*Podospora anserina*	nad4L [S09141] intron 2
Pa.m_nad5_1	*Podospora anserina*	nad5 [S09142] intron 1
Pa.m_nad5_2	*Podospora anserina*	nad5 [S09143] intron 2
Pa.m_nad5_3	*Podospora anserina*	nad5 [S09144] intron 3
Pp.m_cox1	*Peperomia polybotrya*	cox1 [MTPPCOXIG]
Pw.m_cox1_1	*Prototheca wickerhamii*	cox1 [PWMCYTOXI] intron 1
Pw.m_cox1_3	*Prototheca wickerhamii*	cox1 [PWU02970] intron 3
‡Sc.m_cob_4	*Saccharomyces cerevisiae*	cob [YMC4_YEAST] intron 4 (bI4) ‡Maturase involved in splicing of bI4 and aI4 [14]
‡Sc.m_cox1_3	*Saccharomyces cerevisiae*	cox1 [YMX3_YEAST] intron 3 (aI3) ‡I-*Sce*III [82]
‡Sc.m_cox1_4	*Saccharomyces cerevisiae*	cox1 [QXBY34] ‡I-*Sce*II (latent maturase) [24, 14]
‡Sc.m_cox1_5α	*Saccharomyces cerevisiae*	cox1 [S27138] intron 5 (aI5-α) ‡I-*Sce*IV [83, 84]
Sc.m_cox1_5β	*Saccharomyces cerevisiae*	cox1 [S27139] (aI5-β)
‡Sc.m_cytb_3	*Saccharomyces cerevisiae*	cytb [CYBM_YEAST] (bI3) ‡Maturase involved in splicing of bI3 [23]
‡Sc.m_LSU	*Saccharomyces cerevisiae*	LSU rRNA [TRAM_YEAST] (ω-omega) ‡I-*Sce*I [7]
Sd.m_cob_2	*Saccharomyces douglasii*	cob [S23208] intron 2 (bI2)
Sd.m_cob_3	*Saccharomyces douglasii*	cob [S23209] (bI3)
Sd.m_cox1_2	*Saccharomyces douglasii*	cox1 [S23208] intron 2
Ss.m_SSU	*Sclerotinia sclerotiorum*	SSU rRNA [SSU07553]
‡Sp.m_cox1_1	*Schizosaccharomyces pombe*	cox1 [YMC1_SCHPO] intron 1. ‡Maturase involved in splicing of cox1_1 (protein induces recombination when expressed in *Escherichia coli*, mobile intron) [85, 86]
Sp.m_cox1_2	*Schizosaccharomyces pombe*	cox1 [YMC2_SCHPO] intron 2
Yl.m_cox3	*Yarrowia lipolytica*	cox3 [YSJATPTRN]

Sequences are grouped according to their origin. For each sequence, its abbreviation, the species name and the protein name are given together with the databank 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.

belong to the family used to train it from those that do not. This is achieved by evaluating how much better a sequence fits a model than some underlying background distribution or (simple) null model (NULL) and assessing the significance of the resultant score. Database searching using the HMM involved computing 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

Fig 1. cont ...

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

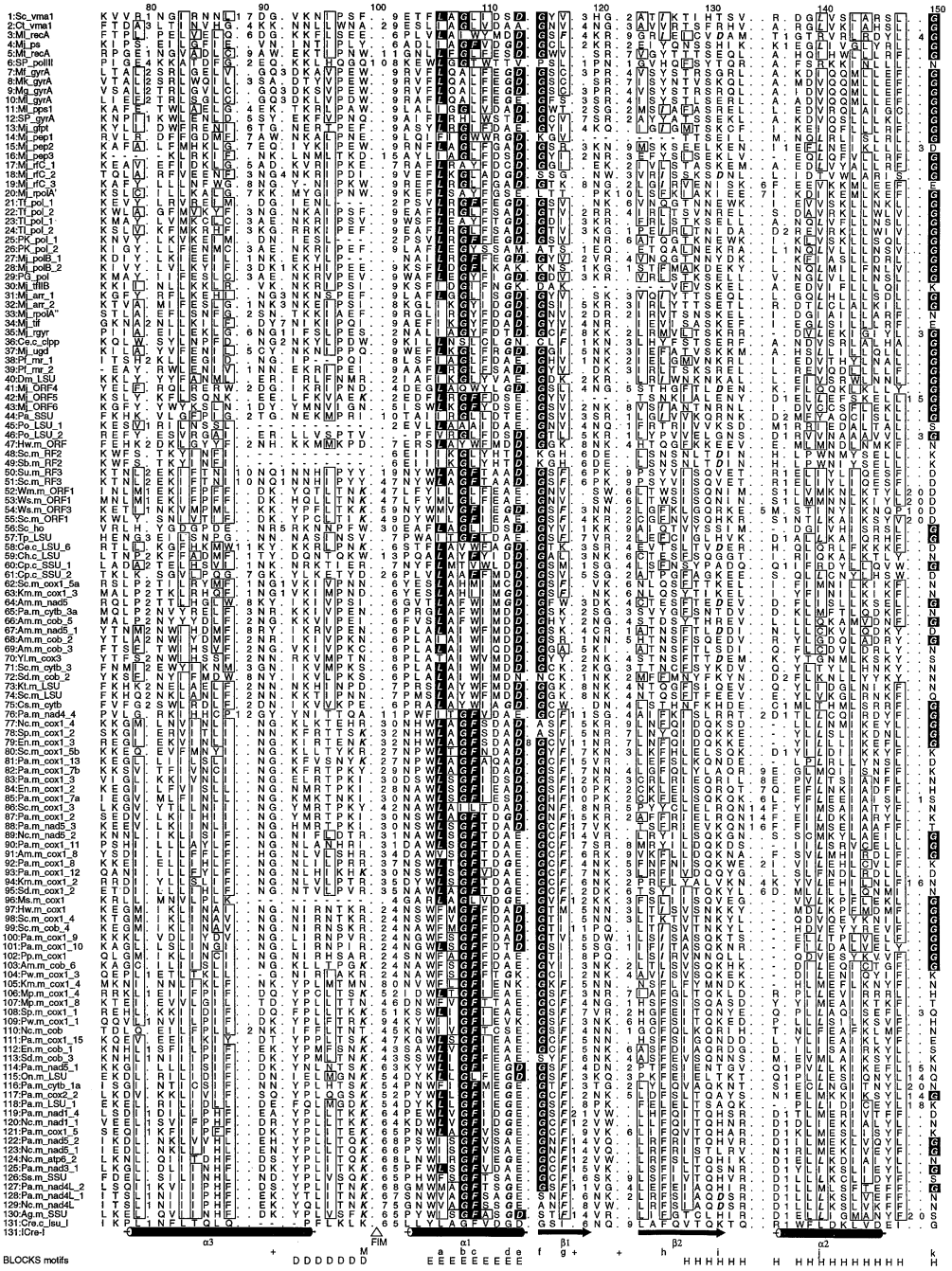


Fig 1. cont ...

sources of systematic error. The MOLPHY suite uses a probabilisitic 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.

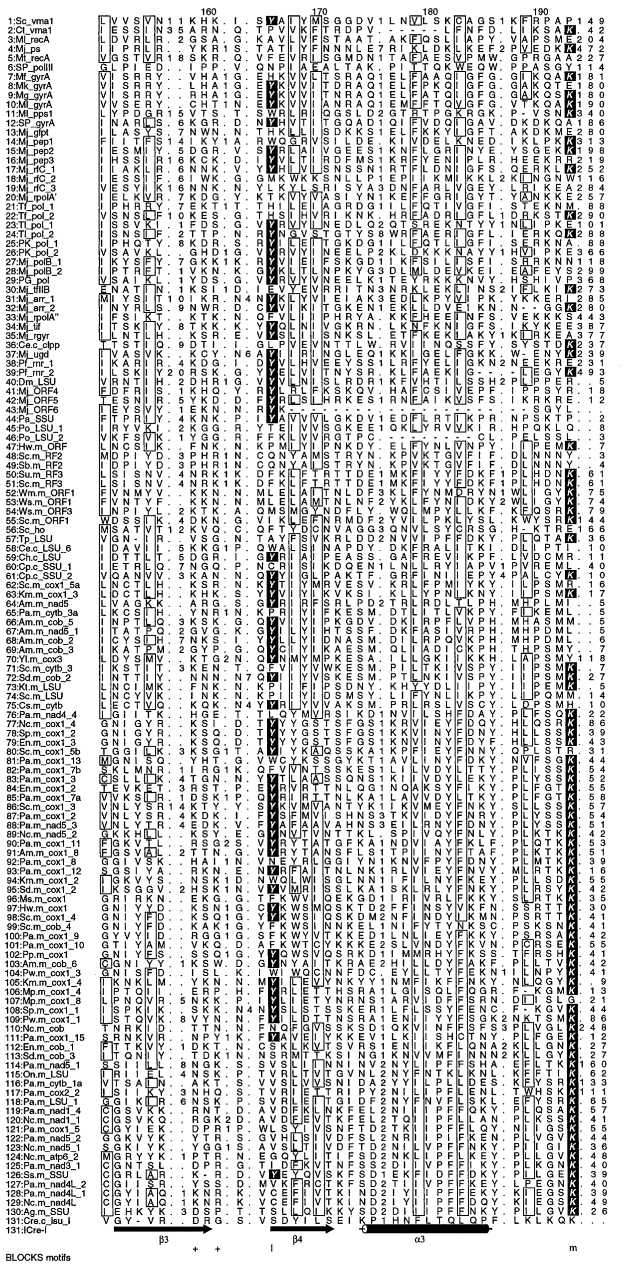


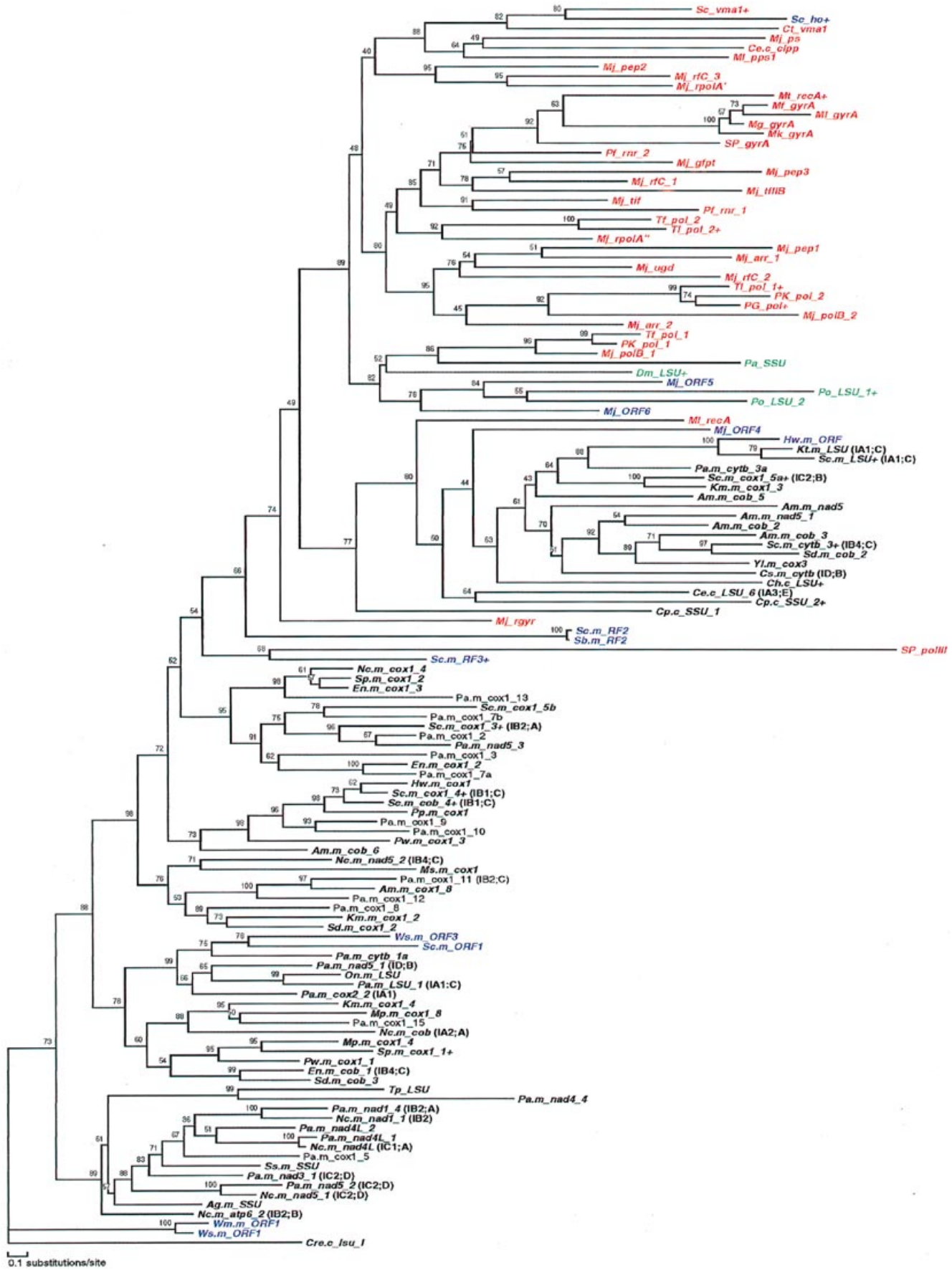
Figure 1. An HMM-generated multiple sequence alignment of the LAGLIDADG family listed in Table 1 (41:Mj_ORF4, 42:Mj_ORF5 and 43:Mj_ORF6 are new members identified in this work). Amino acids conserved in the majority of the sequences are highlighted and columns that are predominantly hydrophobic are boxed. Columns containing '.' correspond to insert states and numbers indicate the lengths of insertions in sequences at that position (if present). The open triangle marks the position of the internal FIM used to model an insertion of variable length. Members of the LAGLIDADG family are comprised of a tandem duplication of a domain containing the LAGLIDADG sequence motif near its N-terminus. The regions before and after the internal FIM form the first (P1) and second (P2) repeated domain respectively. Sequence 131:Cre.c_lsu_l/I-*Cre*I possesses a single copy of the repeated domain and although not part of the HMM training set, is shown aligned to both P1 and P2. Arrows and cylinders represent the β-strands and α-helices taken from the X-ray structure of I-*Cre*I (45). Equivalent conserved residues (columns in bold, italic font) are labelled A–M (P1) and a–m (P2). A number of these positions have been mutated. 98:Sc.m_cox1/I-*Sce*II: G→D mutations at B and F affect endonuclease activity whereas G→D mutations at b and f affect maturase activity (26). 24:Tl_pol_2/PI-*Tli*I: a mutation at E abolishes endonuclease activity (77). 1:Sc_vma1/PI-*Sce*I: D218→N,A at E and D326→N,A at e uncouple DNA binding and DNA cleavage activities; K301→A (column 94) leads to loss of activity (10,43). In 40:Dm_lsu/I-*Dmo*I and/or 45:Po_lsu_1/I-*Por*I, positions marked with + are protected from digestion by protease when substrate is bound (22). For comparison, BLOCKS motifs presents results from a compilation and analysis of intein sequences using a BLOCKS-based rather than HMM-based approach (65). Eight BLOCKS (A–H) were characterized and of these C and E correspond to the LAGLIDADG motifs. An HMM-based analysis of the self-splicing protein domain in inteins (37) indicates that BLOCKS A, B, F and G form part of this particular domain whilst C, D, E and H are part of the LAGLIDADG domain shown here.

RESULTS

LAGLIDADG family

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 ≥48.0. 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 pallidostigmatica* chloroplast LSU rRNA intron protein (37.0, CRECPRRNI2); *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-*Ceu*I) (31.6, DNEI_CHLEU) (63). The remaining sequences all had log-odds scores <29.6 and included *Chlamydomonas reinhardtii* site-specific DNA endonuclease I-*Cre*I (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


Sc_vma1+
Sc_ho+
Ct_vma1
Mj_ps
Ce.c_clpp
Mt_pps1
Mj_pep2
Mj_rfC_3
Mj_rpolA'
Mt_recA+
Mt_gyrA
Mt_gyrA
Mg_gyrA
Mk_gyrA
SP_gyrA
Pf_rnr_2
Mj_gfpt
Mj_pep3
Mj_rfC_1
Mj_hflB
Mj_tif
Pf_rnr_1
Tf_pol_2
Tf_pol_2+
Mj_rpolA''
Mj_pep1
Mj_arr_1
Mj_ugd
Mj_rfC_2
Tf_pol_1+
PK_pol_2
PG_pol+
Mj_polB_2
Mj_arr_2
Tf_pol_1
PK_pol_1
Mj_polB_1
Pa_SSU
Dm_LSU+
Mj_ORF5
Po_LSU_1+
Po_LSU_2
Mj_ORF6
Mt_recA
Mj_ORF4
Hw.m_ORF
Kt.m_LSU (IA1;C)
Sc.m_LSU+ (IA1;C)
Pa.m_cytb_3a
Sc.m_cox1_5a+ (IC2;B)
Km.m_cox1_3
Am.m_cob_5
Am.m_nad5
Am.m_nad5_1
Am.m_cob_2
Am.m_cob_3
Sc.m_cytb_3+ (IB4;C)
Sd.m_cob_2
Yl.m_cox3
Cs.m_cytb (ID;B)
Ch.c_LSU+
Ce.c_LSU_6 (IA3;E)
Cp.c_SSU_2+
Cp.c_SSU_1
Mj_rgyr
Sc.m_RF2
Sb.m_RF2
SP_polIII
Sc.m_RF3+
Nc.m_cox1_4
Sp.m_cox1_2
En.m_cox1_3
Pa.m_cox1_13
Sc.m_cox1_5b
Pa.m_cox1_7b
Sc.m_cox1_3+ (IB2;A)
Pa.m_cox1_2
Pa.m_nad5_3
Pa.m_cox1_3
En.m_cox1_2
Pa.m_cox1_7a
Hw.m_cox1
Sc.m_cox1_4+ (IB1;C)
Sc.m_cob_4+ (IB1;C)
Pp.m_cox1
Pa.m_cox1_9
Pa.m_cox1_10
Pw.m_cox1_3
Am.m_cob_6
Nc.m_nad5_2 (IB4;C)
Ms.m_cox1
Pa.m_cox1_11 (IB2;C)
Am.m_cox1_8
Pa.m_cox1_12
Pa.m_cox1_8
Km.m_cox1_2
Sd.m_cox1_2
Ws.m_ORF3
Sc.m_ORF1
Pa.m_cytb_1a
Pa.m_nad5_1 (ID;B)
On.m_LSU
Pa.m_LSU_1 (IA1:C)
Pa.m_cox2_2 (IA1)
Km.m_cox1_4
Mp.m_cox1_8
Pa.m_cox1_15
Nc.m_cob (IA2;A)
Mp.m_cox1_4
Sp.m_cox1_1+
Pw.m_cox1_1
En.m_cob_1 (IB4;C)
Sd.m_cob_3
Tp_LSU
Pa.m_nad4_4
Pa.m_nad1_4 (IB2;A)
Nc.m_nad1_1 (IB2)
Pa.m_nad4L_2
Pa.m_nad4L_1
Nc.m_nad4L (IC1;A)
Pa.m_cox1_5
Ss.m_SSU
Pa.m_nad3_1 (IC2;D)
Pa.m_nad5_2 (IC2;D)
Nc.m_nad5_1 (IC2;D)
Ag.m_SSU
Nc.m_atp6_2 (IB2;B)
Wm.m_ORF1
Ws.m_ORF1
Cre.c_lsu_I
0.1 substitutions/site

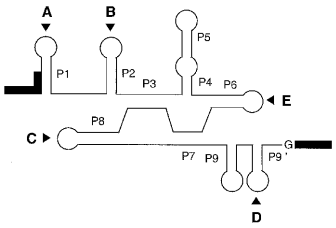


Figure 2. (Opposite) Phylogenetic tree for the LAGLIDADG family based upon the alignment present in the electronic appendix (Fig. 1 shows an alignment of some of these sequences). Intein-encoded, free standing ORFs, archaeal intron-encoded and group I intron-encoded sequences are in red, blue, green 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.

three-dimensional structure of either PI-*Sce*I or I-*Cre*I, 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 constrast, 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, futher 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.

Figure 2 shows the LAGLIDADG family tree and gives an indication of the phylogenetic relationship between the endonucleases. The vast majority of elements branch according to the host elements they are encoded by: group I introns (black), archaeal introns (green) and inteins (red). Free standing ORFs (blue) do not branch together suggesting that they originated from

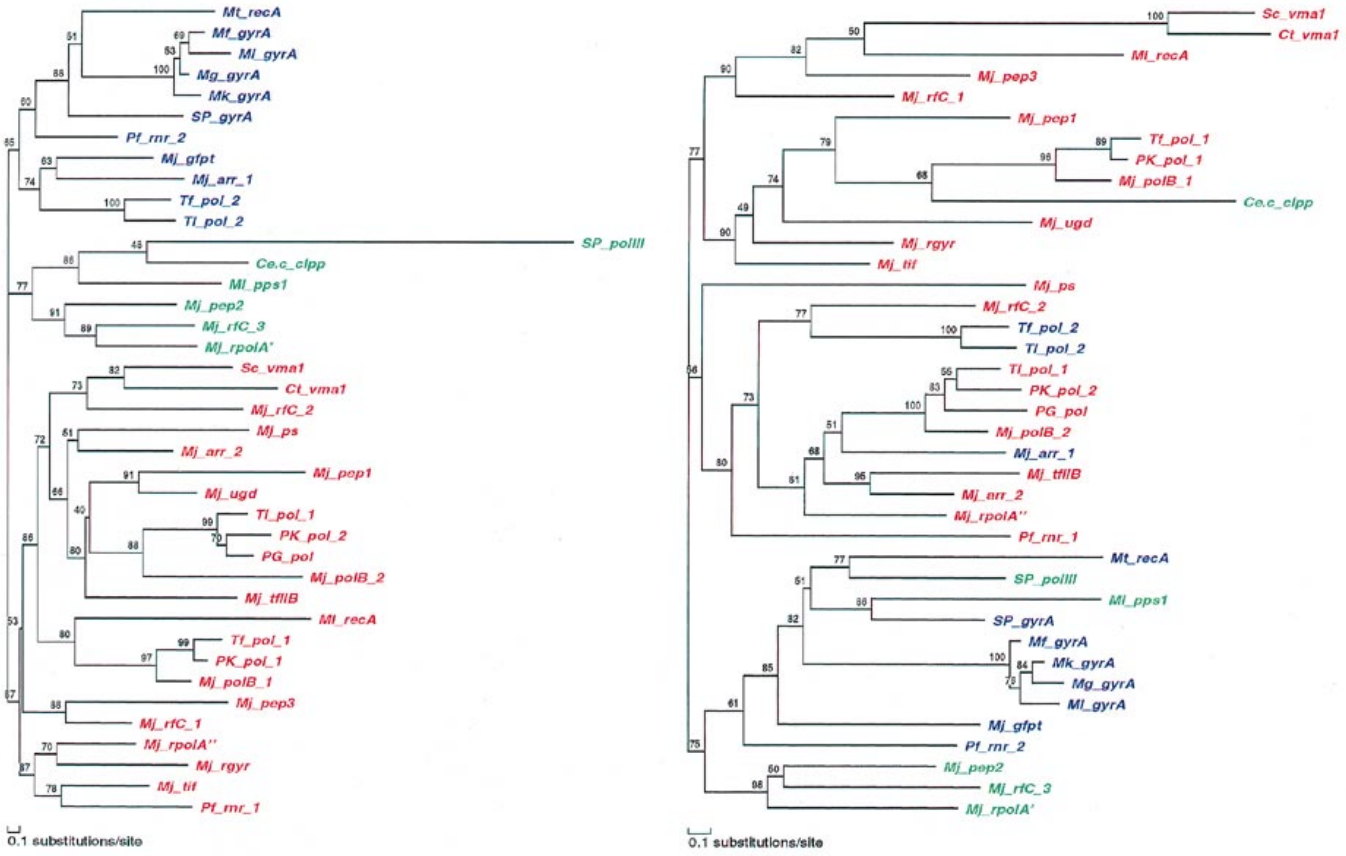


Figure 3. Comparison of the phylogenetic trees for the LAGLIDADG family endonuclease (left) and protein splicing (right) domains of inteins. Sequences in the three major branches of the left hand tree are colored red, green and blue and this scheme is used to color sequences in the other trce. The sequences 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).

Table 2. List of the HNH family members used to train an HMM with new members identified in this work shown in italic

Intein-encoded (bacteria)		
SP.p_gyrB	*Synechocystis* PCC6803	gyrB [D90908]
ORFs (bacteria, bacteriophage, virus, chloroplast, mitochondria, transposon, plasmid)		
AP_adx	*Anabaena 7120* PCC7120	xisA involved in developmental switch [68] [APU38537] (anaredoxin gene)
Bs_yxid	*Bacillus subtilis*	ORF [YXID_BACSU]
Ec_mcra	*Escherichia coli*	5-methylcytosine-specific restriction enzyme A (acts at ^{Me}CCGG sequences) [MCRA_ECOLI]
Ec_yajd	*Escherichia coli*	ORF [YAJD_ECOLI]
Pa_PYS1	*Pseudomonas aeruginosa*	pyocin S1 (killer protein) [PYS1_PSEAE]
Pa_PYS2	*Pseudomonas aeruginosa*	pyocin S2 (killer protein) [PYS2_PSEAE]
SP_sll1193	*Synechocystis* PCC6803	ORF sll1193 [D90901]
SP_sll2007	*Synechocystis* PCC6803	ORF sll2007 [D90908]
SP_slr2080	*Synechocystis* PCC6803	ORF slr2080 [D90910]
SP_slr0393	*Synechocystis* PCC6803	ORF slr0393 [D90915]
LLH_ORF168	Bacteriophage LL-H	ORF168 [LLHORF168A]
BPT3_5.3	Bacteriophage T3	ORF 5.3 [POT3111G]
BPT4_y03e	Bacteriophage T4	ORF [Y03E_BPT4]
BPT4_y04i	Bacteriophage T4	ORF [Y04I_BPT4]
BPT4_y05h	Bacteriophage T4	ORF [Y05H_BPT4]
BPT4_y13m	Bacteriophage T4	ORF [Y13M_BPT4]
BPT7_2.8	Bacteriophage T7	ORF 2.8 [PET7XX]
BPT7_3.8	Bacteriophage T7	ORF 3.8 [Y38_BPT7]
BPT7_7.7	Bacteriophage T7	ORF 7.7 [T7CG]
C31_ORF2	Bacteriophage ϕC31	ORF2 [APHIC31C]
C31_ORF7	Bacteriophage ϕC31	ORF7 [S38919]
LAM_EA31	Bacteriophage λ	ORF EA31 [VE31_LAMBD]
Phi41_ORFL3	Bacteriophage ϕ41	ORFL3 [B4PORFSX]
R1T_ORF26	Bacteriophage r1t	ORF26 [BRU38906]
R1T_ORF41	Bacteriophage r1t	ORF41 [BRU38906]
Pbc_A422R	Paramecium bursaria Chlorella virus 1	ORF A422R [PBU42580]
Pbc_A87R	Paramecium bursaria Chlorella virus 1	ORF A87R [PBU42580]
Cm.c_ORF2	*Chlamydomonas moewusii*	ORF2 [CHCMPSBA]
Sg.m_mutS	*Sarcophyton glaucum*	mutS protein homolog [S58881]
Pa.t_AP41	*Pseudomonas aeruginosa*	pyocin AP41 large chain [S30271]
Ec.t_cea2	*Escherichia coli*	colicin E2 [CEA2_ECOLI]
Ec.t_cea7	*Escherichia coli*	ColE7-K317 colicin E7 [S22453]
Ec.t_cea8	*Escherichia coli*	colicin E8 [CEA8_ECOLI]
Ec.t_cea9	*Escherichia coli*	ColE9-J colicin E9 [PQ0032]
Intron-encoded (bacteria, bacteriophage, chloroplast, mitochondria, transposon)		
Av_ORF	*Azotobacter vinelandii*	ORF [S35081] group II intron
CS_ORF	*Calothrix sp.*	ORF [S35080] group II intron
CS_ORF2	*Calothrix sp.* PCC7601	ORF [S40013] group II intron (protein 2)
So.c_RDPO	*Scenedesmus obliquus*	probable reverse transcriptase [RDPO_SCEOB]
PhiE_DPO	Bacteriophage ϕE	DNA polymerase [S50092] group intron I
RB3_nrdB	Bacteriophage RB3	nrdB [A61182] group I intron (I-TevIII)
SPO1_DPO	Bacteriophage SPO1	DNA polymerase [A36077] group I intron
SPP1_ORF36.1	Bacteriophage SPP1	[BSSPP1] group I intron
SP82_DPO	Bacteriophage SP82	DNA polymerase [S50094] group I intron
Cm.c_psbA	*Chlamydomonas moewusii*	psba [YCX1_CHLMO] group I intron
Pl.m_LSU1	*Pylaiella littoralis*	LSU rDNA [S58503] group II intron
Pl.m_LSU2	*Pylaiella littoralis*	LSU rDNA [S58504] group II intron
Cd.t_ORF	*Clostridium difficile*	[CDIIORF] group II intron

Sequences are grouped according to their environment and for each one, its abbreviation, the species name and the protein name are given together with the databank code in '[ ]'. Other abbreviations are as follows. .c, chloroplast; .m, mitochondria; .t, transposon sequence; .p, plasmid sequence.

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, eucaryotic 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 genes. 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 (37,66). This model for the origin of mobile group I introns was based on the observations that (i) group I introns encode several types of endonucleases and (ii) the ORFs are inserted at several different positions in the catalytic core of the introns (Fig. 2, insert). The phylogenetic analysis here provides additional insights into the frequencies of such events. Group I introns have been classified into different subclasses (67); A correlation between the host intron subclass, the position of the ORF within the catalytic core of the host intron and the branching pattern of the endonucleases would suggest that each insertion site in the catalytic core corresponded to one event. In contrast, Figure 2 shows that closely related endonucleases are inserted in different position within the catalytic core of group I introns and in different intron subclasses. For example, the related group I intron-encoded endonucleases in the lower cluster are present in introns that belong to subclasses IB2, IC1 and IC2 and are inserted in position A, B and D in the catalytic core of the introns. This suggests that invasion of an endonuclease domain into the catalytic core of a group I intron has happened more frequently than estimated by the number of insertion sites and families of endonucleases. Since our analysis suggests there has been no exchange of endonucleases domains between the different classes of elements, it is most likely that the source of these domains was other group I introns.

Two different endonuclease families have been shown to be encoded by inteins (see below). Only one site of insertion has been observed to date. Thus, a comparison of insertion sites and the branching pattern is not possible. Instead, Figure 3 shows separate phylogenetic trees for the protein splicing domain and LAGLIDADG endonuclease domain of inteins. 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 that has been characterized using an HMM. Table 2 lists the HNH famly members used to train an HMM. Only these sequences had log-odds scores >22.6, all other sequences had scores <15.0. All sequences with log-odds scores >22.6 are classified as belonging to the HNH family and consist 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

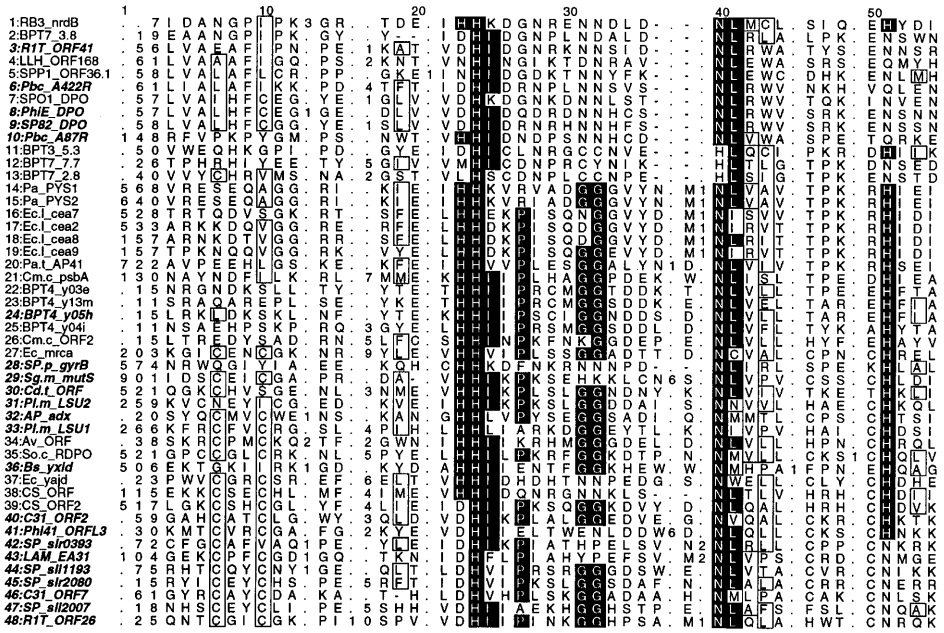


Figure 4. An HMM-generated multiple sequence alignment of HNH family listed in Table 2 with new members identified in this work shown in a different font. Amino acids conserved in the majority of the sequences are highlighted and columns that are predominantly hydrophobic are boxed. Columns containing '.' correspond to insert states and numbers indicate the lengths of insertions in sequences at that position (if present).

developmentally controlled DNA rearrangement (68). 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-*Sce*I (43) and I-*Cre*I (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-*Cre*I only possesses one LAGLIDADG motif and acts as a homodimer, PI-*Sce*I is a monomer containing two domains whose overall structure is similar structure to each I-*Cre*I 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-*Cre*I (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-*Cre*I 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-*Cre*I, 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_2/I-*Por*I and 40:Dm_LSU/I-*Dmo*I (+ 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-*Cre*I, 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-*Cre*I 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

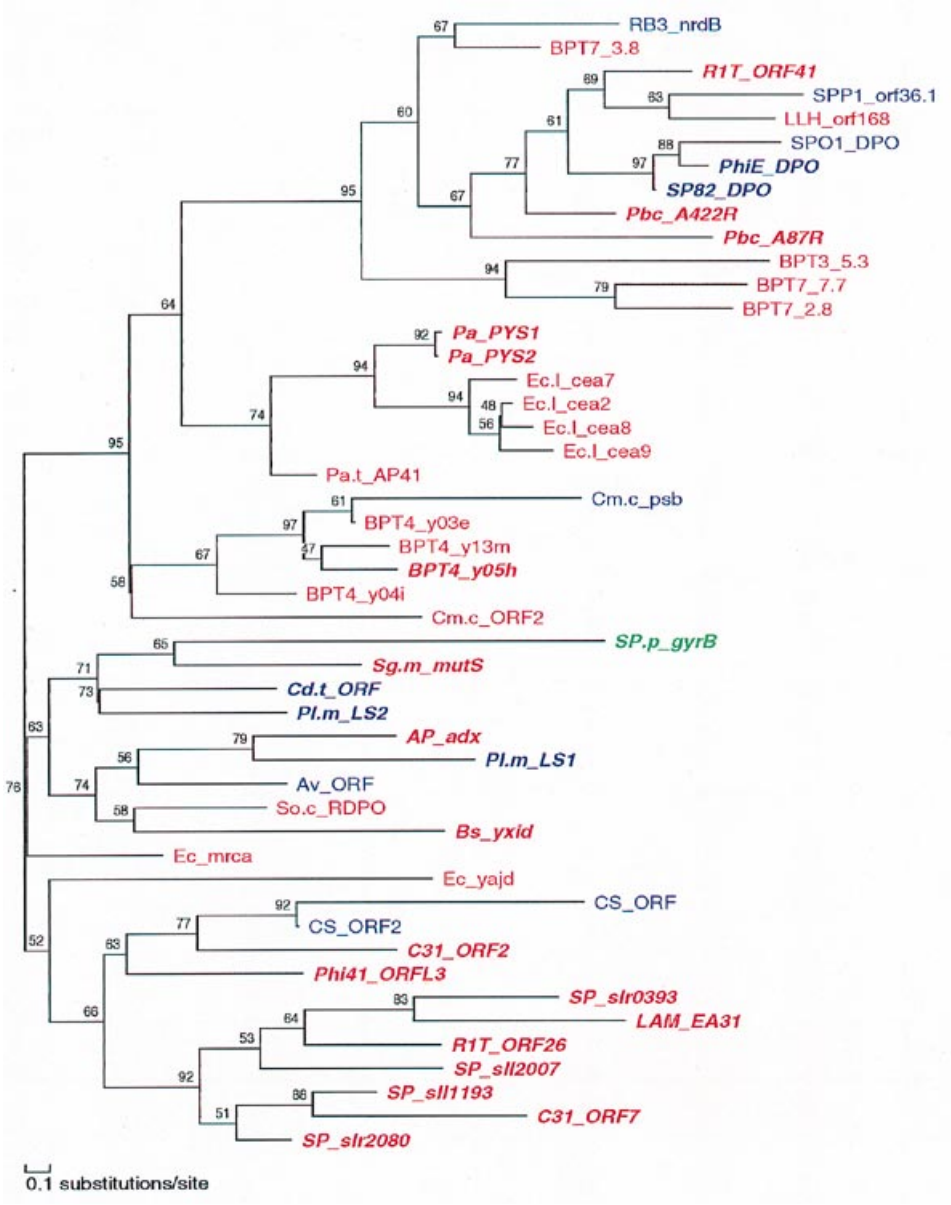


Figure 5. Phylogenetic tree for the HNH family members listed in Table 2 and based upon the alignment shown in Figure 4. Intein-encoded, ORF and intron-encoded endonucleases are coloured green, red and blue respectively. New endonucleases identified in this work are shown in an italic font.

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 β2-α2 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-*Cre*I and its substrate, the β2-α2 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 *Eco*RV (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-*Sce*I (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.

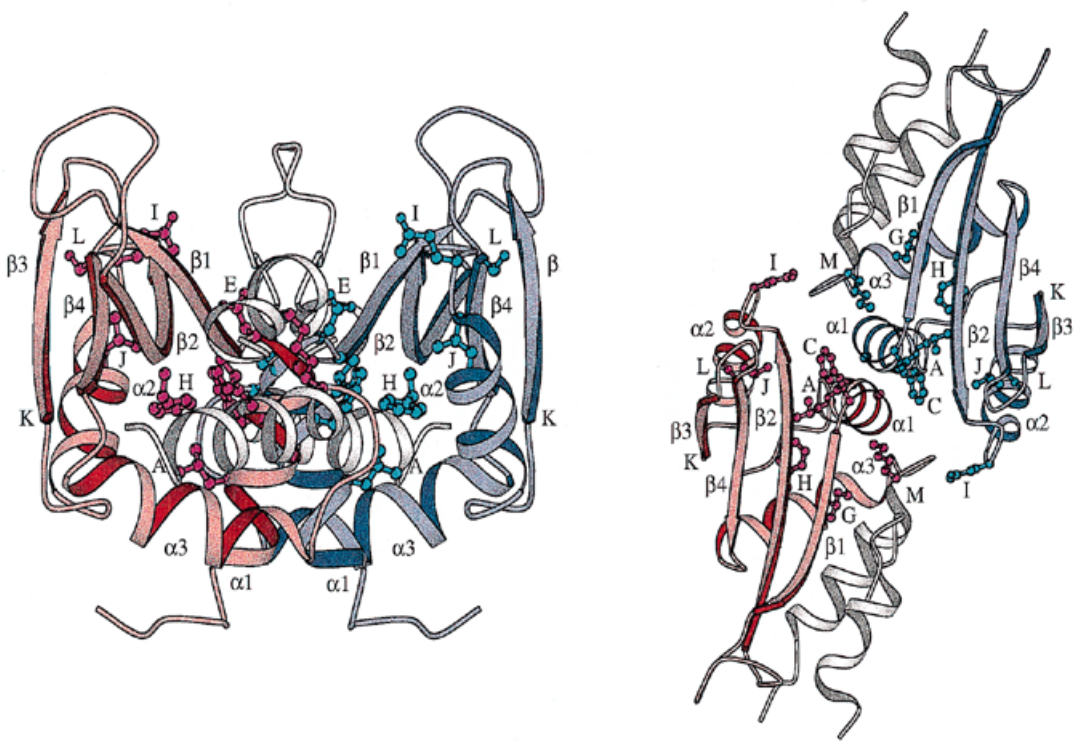


Figure 6. Ribbon diagrams of the I-*Cre*I homodimer (45) showing the residues conserved in P1 or P2 (cyan or magenta) in the two monomers (blue and red). The positions labelled A–M and a–m and secondary structure designations are taken from Figure 1 and elsewhere (45). For clarity, not all positions are labelled. The regions in grey are not part of the LAGLIDADG HMM and do not form the core of the repeated domain present in the LAGLIDADG family.

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 occured 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 succesfully 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-*Cre*I 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

1 Dujon,B. (1989) *Gene*, **82**, 91–114.
2 Ågaard,C., Dalgaard,J. and Garrett,R. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 12285–12289.

3 Gimble,F. and Thorner,J. (1992) *Nature*, **357**, 301–306.
4 Mueller,J., Bryk,M., Loizos,N. and Belfort,M. (1994) Homing endonulceases. In Linn,S.M., Lloyd,R.S. and Roberts,R.J. (eds), *Nucleases*. Cold Spring Harbor Press, Cold Spring Harbor, New York, pp. 111–143.
5 Dujon,B., Belfort,M., Butow,R., Jacq,C., Lemieux,C., Perlman,P. and Vogt,V. (1989) *Gene*, **82**, 115–118.
6 Dujon,B. (1980) *Cell*, **20**, 185–197.
7 Colleaux,L., D'Auriol,L., Betermier,M., Cottarel,G., Jacquier,A., Galibert,F. and Dujon,B. (1986) *Cell*, **44**, 521–533.
8 Belfort,M., Reaban,M., Coetzee,T. and Dalgaard,J. (1995) *J. Bacteriol.*, **177**, 3897–3903.
9 Hensgens,L., Bonen,L., de Haan,M., van der Horst,G. and Grivell,L. (1983) *Cell*, **32**, 379–389.
10 Gimble,F. and Stephens,B. (1995) *J. Biol. Chem.*, **270**, 5849–5856.
11 Perler,F., Comb,D., Jack,W., Moran,L., Qiang,B., Kucera,R., Benner,J., Slatko,B., Nwankwo,D. and Hempstead,S. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 5577–5581.
12 Dalgaard,J., Garrett,R. and Belfort,M. (1994) *J. Biol. Chem.*, **269**, 28885–28892.
13 Monteilhet,C., Perrin,A., Thierry,A., Colleaux,L. and Dujon,B. (1990) *Nucleic Acids Res.*, **18**, 1407–1413.
14 Wenzlau,J., Saldanha,R., Butow,R. and Perlman,P. (1989) *Cell*, **56**, 421–430.
15 Marshall,P. and Lemieux,C. (1992) *Nucleic Acids Res.*, **20**, 6401–6407.
16 Sargueil,B., Hatat,D., Delahodde,A. and Jacq,C. (1990) *Nucleic Acids Res.*, **18**, 5659–5665.
17 Wernette,C., Saldanha,R., Perlman,P. and Butow,R. (1990) *J. Biol. Chem.*, **265**, 18976–18982.
18 Lykke-Andersen,J., Thi-Ngoc,H. and Garrett,R. (1994) *Nucleic Acids Res.*, **22**, 4583–4590.
19 Perrin,A., Buckle,M. and Dujon,B. (1993) *EMBO J.*, **12**, 2939–2947.
20 Gimble,F. and Wang,J. (1996) *J. Mol. Biol.*, **263**, 163–180.
21 Ågaard,C., Awayez,M. and Garrett,R. (1997) *Nucleic Acids Res.*, **25**, 1523–1530.
22 Lykke-Andersen,J., Garrett,R. and Kjems,J. (1996) *Nucleic Acids Res.*, **24**, 3982–3989.
23 Lazowska,J., Claisse,M., Gargouri,A., Kotylak,Z., Spyridakis,A. and Solonimski,P. (1989) *J. Mol. Biol.*, **205**, 275–289.
24 Delahodde,A., Goguel,V., Becam,A., Creusot,F., Perea,J., Banroques,J. and Jacq,C. (1989) *Cell*, **56**, 431–441.
25 Lambowitz,A. and Perlman,P. (1990) *Trends Biochem. Sci.*, **15**, 440–444.
26 Henke,R., Butow,R. and Perlman,P. (1995) *EMBO J.*, **14**, 5094–5099.
27 Bahl,L., Jelinek,F. and Mercer,R. (1983) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **5**, 179–190.
28 Lee,K.-F. (1989) Automatic speech recognition: the development of the SPHINX system, Kluwer Academic, Boston, MA.
29 Rabiner,L. and Juang,B. (1986) *IEEE ASSP Magazine*, **3**, 4–16.
30 Rabiner,L. (1989) *Proceedings of the IEEE*, **77**, 257–286.
31 Krogh,A., Brown,M., Mian,I., Sjölander,K. and Haussler,D. (1994) *J. Mol. Biol.*, **235**, 1501–1531.
32 Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 1059–1063.
33 Eddy,S., Mitchison,G. and Durbin,R. (1995) *J. Comp. Biol.*, **2**, 9–23.
34 Eddy,S. (1996) *Curr. Opin. Struct. Biol.*, **6**, 361–365.
35 Fujiwara,Y., Asogawa,M. and Konagaya,A. (1994) *Intelligent Syst. Mol. Biol.*, **2**, 121–129.
36 Mian,I. (1997) *Nucleic Acids Res.*, **25**, 3187–3195.
37 Dalgaard,J., Moser,M., Hughey,R. and Mian,I. (1997) *J. Comp. Biol.*, **4**, 193–214.
38 Bateman,A. and Chothia,C. (1996) *Curr. Biol.*, **6**, 1544–1547.
39 Bateman,A., Eddy,S. and Chothia,C. (1996) *Protein Sci.*, **5**, 1939–1941.
40 Hazes,B. (1996) *Protein Sci.*, **5**, 1490–1501.
41 Shub,D., Goodrich-Blair,H. and Eddy,S. (1994) *Trends Biochem. Sci.*, **19**, 402–404.
42 Grundy,W., Bailey,T., Elkan,C. and Baker,M. (1997) *Biochem. Biophys. Res. Commun.*, **231**, 760–766.
43 Duan,X., Gimble,F. and Quiocho,F. (1997) *Cell*, **89**, 555–564.
44 Tanaka-Hall,T., Porter,J., Young,K., Koonin,E., Beachy,P. and Leahy,D. (1997) *Cell*, **91**, 85–97.
45 Heath,P., Stephens,K., Monnat,R. and Stoddard,B. (1997) *Nature Struct. Biol.*, **4**, 468–476.
46 Gorbalenya,A. (1994) *Protein Sci.*, **3**, 1117–1120.
47 Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) *J. Mol. Biol.*, **215**, 403–410.
48 Mian,I., Moser,M., Holley,W. and Chatterjee,A. (1997) *J. Comp. Biol.*, in press
49 Herbert,A., Alfken,J., Kim,Y.-G., Mian,I., Nishijura,K. and Rich,A. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 8421–8426.
50 Moser,M., Holley,W., Chatterjee,A. and Mian,I. (1997) *Nucleic Acids Res.*, in press.
51 Mian,I. and Moser,M. (1997) *Biochem. Mol. Med.*, in press.
52 Hughey,R. and Krogh,A. (1996) *Comput. Appl. Biosci.*, **12**, 95–107. The hidden Markov model software can be accessed at URL http://www.cse.ucsc.edu/research/compbio/sam.html
53 Brown,M., Hughey,R., Krogh,A., Mian,I., Sjölander,K. and Haussler,D. (1993) *Intelligent Syst. Mol. Biol.*, **1**, 47–55.
54 Sjölander,K., Karplus,K., Brown,M., Hughey,R., Krogh,A., Mian,I. and Haussler,D. (1996) *Comput. Appl. Biosci.*, **12**, 327–345.
55 Altschul,S. (1991) *J. Mol. Biol.*, **219**, 555–565.
56 Barrett,C., Hughey,R. and Karplus,K. (1997) *Comput. Appl. Biosci.*, **13**, 191–199.
57 NCI (1997) NRP (Non-Redundant Protein) and NRN (Non-Redundant Nucleic Acid) Database. Distributed on the Internet via anonymous FTP from ftp.ncifcrf.gov, under the auspices of the National Cancer Institute's Frederick Biomedical Supercomputing Center.
58 Barton,G. (1993) *Protein Engng.*, **6**, 37–40.
59 Maciukenas,M. (1992) Treetool: an interactive tool for displaying, editing and printing phylogenetic trees, Currently, Treetool is modified and maintained by Mike McCaughey, Ribosomal Database Project, University of Illinois. It is available from ftp://rdp.life.uiuc.edu/rdp/programs/TreeTool.
60 Kraulis,P. (1991) *J. Appl. Crystallog.*, **24**, 946–950.
61 Adachi,J. (1995) Modelling of molecular evolution and maximum likelihood inference of molecular phylogeny. PhD dissertation, Institute of Statistical Mathematics, Tokyo.
62 Adachi,J. and Hasegawa,M. (1992) MOLPHY: Programs for Molecular Phylogenetics, I. PROTML: Maximum Likelihood Inference of Protein Phylogeney Corrlputer Science Monographs 27 Institute of Statistical Matllematics, Tokyo. MOLPHY is available from ftp://sunmh.ism.ac.jp/pub/molphy.
63 Marshall,P. and Lemieux,C. (1991) *Gene*, **104**, 241–245.
64 Durrenberger,F. and Rochaix,J. (1991) *EMBO J.*, **10**, 3495–3501.
65 Perler,F., Olsen,G. and Adam,E. (1997) *Nucleic Acids Res.*, **25**, 1087–1093.
66 Loizos,N., Tillier,E. and Belfort,M. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11983–11987.
67 Michel,F. and Westhof,E. (1990) *J. Mol. Biol.*, **216**, 585–610.
68 Lammers,P., Golden,J. and Haselkorn,R. (1986) *Cell*, **44**, 905–911.
69 Dalgaard,J. and Garrett,R. (1992) *Gene*, **121**, 103–110.
70 Garrett,R., Ågaard,C., Andersen,M., Dalgaard,J., Lykke-Andersen,J., Phan,H., Trevisanato,S., Östergaard,L., Larsen,N. and Leffers,H. (1994) *Systematic Appl. Microbiol.*, **16**, 180–191.
71 Baldwin,G., Vipond,I. and Halford,S. (1995) *Biochemistry*, **34**, 705–714.
72 Vipond,I., Baldwin,G. and Halford,S. (1995) *Biochemistry*, **34**, 697–704.
73 Turmel,M., Mercier,J., Côte,V., Otis,C. and Lemieux,C. (1995) *Nucleic Acids Res.*, **23**, 2519–2525.
74 Pietrokovski,S. (1994) *Protein Sci.*, **3**, 2340–2350.
75 Davis,E., Jenner,P., Brooks,P., Colston,M. and Sedgwick,S. (1992) *Cell*, **71**, 201–210.
76 Bremer,M., Gimble,F., Thorner,J. and Smith,C. (1992) *Nucleic Acids Res.*, **20**, 5484–5490.
77 Hodges,R., Perler,F., Noren,C. and Jack,W. (1992) *Nucleic Acids Res.*, **20**, 6153–6157.
78 Nakagawa,K., Morishima,N. and Shibata,T. (1991) *J. Biol. Chem.*, **266**, 1977–1984.
79 Kostriken,R. and Heffron,F. (1984) *Cold Spring Harbor Symposia Quant. Biol.*, **49**, 89–96.
80 Dalgaard,J., Garrett,R. and Belfort,M. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 5414–5417.
81 Côte,V., Mercier,J., Lemieux,C. and Turmel,M. (1993) *Gene*, **129**, 69–76.
82 Perea,J., Desdouets,C., Schapria,M. and Jacq,C. (1993) *Nucleic Acids Res.*, **21**, 358–360.
83 Moran,J., Wernette,C., Mecklenburg,K., Butow,R. and Perlman,P. (1992) *Nucleic Acids Res.*, **20**, 4069–4076.
84 Seraphin,B., Faye,G., Hatat,D. and Jacq,C. (1992) *Gene*, **113**, 1–8.
85 Manna,F., Massardo,D., Giudice,L., Buonocore,A., Nappo,A., Alifano,P., Shöfer,B. and Wolf,K. (1991) *Curr. Genet.*, **19**, 295–299.
86 Shafer,B., Wilde,B., Massardo,D., Manna,F., Giudice,L. and Wolf,K. (1994) *Curr. Genet.*, **25**, 336–341.