Modular organization of inteins and C-terminal autocatalytic domains

SHMUEL PIETROKOVSKI

Fred Hutchinson Cancer Research Center, Seattle, Washington 98109

(RECEIVED June 4, 1997; ACCEPTED August 1, 1997)

Abstract

Analysis of the conserved sequence features of inteins (protein “introns”) reveals that they are composed of three distinct modular domains. The N-terminal (N) and C-terminal (C) domains are predicted to perform different parts of the autocatalytic protein splicing reaction. An optional endonuclease domain (EN) is shown to correspond to different types of homing endonucleases in different inteins. The N domain contains motifs predicted to catalyze the first steps of protein splicing, leading to the cleavage of the intein N terminus from its protein host. Intein N domain motifs are also found in C-terminal autocatalytic domains (CADs) present in hedgehog and other protein families. Specific residues in the N domain of intein and CADs are proposed to form a charge relay system involved in cleaving their N-termini. The intein C domain is apparently unique to inteins and contains motifs that catalyze the final protein splicing steps: ligation of the intein flanks and cleavage of its C terminus to release the free intein and spliced host protein. All intein EN domains known thus far have dodecapeptide (DOD, LAGLI-DADG) type homing endonuclease motifs. This work identifies an EN domain with an HNH homing-endonuclease motif and two new small inteins with no EN domains. One of these small inteins might be inactive or a “pseudo intein.” The results suggest a modular architecture for inteins, clarify their origin and relationship to other protein families, and extend recent experimental findings on the functional roles of intein N, C, and EN motifs.

Keywords: charge relay system; C-terminal autocatalytic domains; HO endonuclease; homing endonucleases; inteins; protein splicing

Inteins are proteins that catalyze self-excision and ligation of various host proteins, with a standard polypeptide bond. Organisms from all three domains of life contain inteins; they have been found in yeast and algal chloroplasts, in mycobacteria and cyanobacteria, and in thermophilic archaea. So far, no essential role has been shown for inteins. All their identified functions involve their own preservation and maintenance, with no apparent benefit to the host protein and organism. At least some inteins are multifunctional, being able to both catalyze their own protein splicing and to home a copy of their gene into intein-less alleles (Gimble & Thorner, 1992; Chong et al., 1996). The protein splicing function can be further divided into cleavage of the intein N-terminal end and ligation of the two host flanks (exteins) followed by cleavage of the intein C-terminal end (Xu & Perler, 1996).

Other proteins are known to perform some of these functions in a similar fashion. C-terminal autocatalytic domains, found in eukaryotic hedgehog developmental proteins (Hammerschmidt et al., 1997) and in various *Caenorhabditis elegans* proteins (Burglin,

1996; Porter et al., 1996a), cleave themselves from the proteins, modulating the activity of the N-terminal parts. Different types of homing endonucleases, residing in self-splicing and archaea introns and as free-standing genes, are responsible for the mobility of their genes (Belfort & Roberts, 1997). Together with this similarity in function, inteins also share various sequence motifs with CADs and a family of homing endonucleases (Hirata et al., 1990; Perler et al., 1992; Davis et al., 1994; Pietrokovski, 1994; Koonin, 1995; Burglin, 1996; Fsihi et al., 1996).

This work examines the conserved sequence motifs of inteins and CADs and their presence in other proteins. A general model for the modular organization of inteins and CADs is presented, and specific domains are suggested to perform similar discrete functions in different types of proteins. Conserved residues identified in this work at the intein N' terminal domain are proposed to be part of a charge relay system that catalyzes the initial steps of protein splicing. Identification of new inteins with atypical sequence features shows that inteins optionally harbor different types of endonuclease domains.

Results and discussion

Intein motifs can be organized in functional domains

All 40 publicly available intein sequences were analyzed for conserved regions (motifs). Inteins were found to contain 10 motifs,

Reprint requests to: Shmuel Pietrokovski, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, Washington 98109; e-mail: pietro@sparky.fhcr.org.

Abbreviations: aa, amino acid(s); C, carboxyl-terminal domain; CAD, C-terminal autocatalytic domain; DOD, dodecapeptide; EN, endonuclease; N, amino-terminal domain.

including 2 that had not been reported before. Intein motifs can be grouped in three domains according to their location and inferred function (Fig. 1). (1) An N domain that includes four motifs and spans 100–150 aa (Fig. 2; Table 1). Mutations in this domain have been shown to affect the first step in protein splicing, an N-S/O acyl shift in the peptide bond connecting the N terminus of the intein and the N-terminal extein (Hirata & Anraku, 1992; Hodges et al., 1992; Cooper et al., 1993; Chong et al., 1996; Xu & Perler, 1996; Kawasaki et al., 1997). (2) A C domain includes two motifs in the C-terminal 25–40 aa. Residues in these motifs are necessary for catalyzing the next steps of protein splicing: branch formation and resolution to free the intein and ligated exteins (Chong et al., 1996; Xu & Perler, 1996; Kawasaki et al., 1997). (3) A central EN domain typically consisting of four motifs. This domain is 190–420 aa in size and is optional. It is missing from a few inteins (present work and Perler et al., 1997) and protein splicing was shown for inteins that naturally (Telenti et al., 1997) or artificially (Chong & Xu, 1997) lack this domain. Until now, the domain was only known to include motifs similar to those of dodecapeptide (DOD, LAGLI-DADG) homing endonucleases (Hirata et al., 1990; Perler et al., 1992; Davis et al., 1994; Pietrovski, 1994, 1996a; Bult et al., 1996; Fsihi et al., 1996). Mutations of the EN motifs affect the activity of the intein endonuclease (Hodges et al., 1992; Gimble & Stephens, 1995).

Two new N-domain motifs were identified in this study (N2 and N4). The two motifs are similar to each other and their main attribute is a conserved acidic residue usually preceded by a glycine (Fig. 2; Table 1). Motif N4 is more conserved than motif N2, being longer and less diverse. Nevertheless, the N2 motif is reliably assigned (P value $1 \cdot 10^{-17}$; Schuler et al., 1991) and can be identified in all inteins. Motif N4 could not be identified in three of the four eukaryotic inteins, in inteins Tli pol-2, Mja pol-1, and their alleles, and in intein Mja PEPSyn (Table 1).

Charge relay system in intein N' terminal cleavage

Six motifs were identified in CAD sequences from hedgehog and *C. elegans* proteins (Fig. 2). Four of the motifs are similar to the intein N-domain motifs and appear in the same order (Table 2). Sequence similarity was formerly reported for two pairs of N-terminal motifs from the families (Koonin, 1995; Burglin, 1996).

Searching databases of multiple alignments with the CAD motifs as queries (Pietrovski, 1996b) identified another notable identity. CAD motif 4 was found similar to a conserved region of cutinases (Fig. 3). Cutinases are serine esterases with catalytic

triad active sites (Martinez et al., 1994). The cutinase histidine active site residue is aligned with an invariant histidine crucial for the CAD proteolytic cleavage (Lee et al., 1994; Valentini et al., 1997). Limited sequence similarity of this CAD histidine to catalytic histidines of serine proteases was noted previously (Lee et al., 1994) and a charge relay system was suggested for the first step of the CAD autocleavage reaction (Porter et al., 1996a).

The relationship between the intein N-domain and CAD motifs (Table 2) suggests similar roles for the common motifs. Experimental evidence indicates a similar autocatalytic reaction: an acyl shift of a specific Cys or Ser peptide bond to an ester bond that is cleaved by a nucleophilic attack (Chong et al., 1996; Porter et al., 1996a; Xu & Perler, 1996). This attack occurs in inteins through the C domain by the Cys, Ser, or Thr residue flanking the intein C-terminal end (position 9 in motif C1) and apparently by a cholesterol molecule in CADs (Porter et al., 1996b). In both inteins and CADs, an invariant histidine (positions 10 in motif N3 and 9 in motif CAD 4, respectively) was shown essential for autoproteolytic activity (Lee et al., 1994; Kawasaki et al., 1997; Valentini et al., 1997). This histidine, together with a conserved acidic aa, might be part of a catalytic triad that catalyzes the acyl shift, the nucleophilic attack, or both by a charge relay system. This scheme is supported by the position of this histidine in the Sce VMA structure (Duan et al., 1997) and by the motif similarity of the conserved CAD histidine to the active site histidine from a family of catalytic triad hydrolases.

CADs and the intein N domain likely share a common origin. Intein motifs N2 and N4 and CAD motifs 2 and 5 are also similar to each other (Table 2; Fig. 2). The simplest scenario is that both pairs originated from a duplicated motif in the common ancestor of CAD and intein proteins. The two resulting motifs in each family then diverged from each other. Motifs CAD 2 and intein N4 are highly similar to each other (Table 2) and have a well-conserved Gly-Asp dipeptide that can be part of the suggested catalytic triad. Thus, the sequence order of the suggested catalytic triad residues is different in inteins (His-Asp) and in CADs (Asp-His). This difference in order is observed among hydrolase families with catalytic triad active sites (Ollis et al., 1992; Rawlings & Barrett, 1994). The roles of motifs intein N2 and CAD 5 are unclear. N2 might complement the absence of motif N4 in some inteins, act with the invariant His in a different context than N4, or perform a different function. The relative position of motif N2 to the His in motif N3 in the Sce VMA intein structure is unavailable. Although the structure has been published (Duan et al., 1997), the coordinates have not been released publicly to date (PDB tracking number T11189, listed on hold until April 1998).

Fig. 1. Intein domains. Motifs are depicted as rectangles. The longest and shortest distances between motifs are represented by the lines between the motifs. The optional Endonuclease domain is shown with DOD type motifs. Motifs corresponding to those in Pietrovski (1994) and Perler et al. (1997) are: N1, A; N3, B; EN1, C (DOD1); EN2, D; EN3, E (DOD2); EN4, H; C2, F; C1, G. Note that the C-domain motifs are numbered from the C-terminal end of the proteins.

C-terminal autocatalytic domains

Fig. 2. CAD and intein N-domain motifs. Intein (top) and CAD (bottom) motifs are shown as sequence logos (Henikoff et al., 1995). The height of each amino acid is proportional to its conservation at that position, after the sequences have been weighted and frequencies adjusted by the expected amino acid composition. Amino acids are colored according to their chemical and physical properties. Minimal and maximal distances between adjacent motifs are shown between them and the proteolytic cleavage sites are indicated.

Table 1. Intein N-domain motifs^a

		N1		N2		N3		N4	
Sce VMA	1	CFAGKTNVLMADG	(7)	IEVG N KV	(42)	LLKFTCNAT H ELVV	83		
Ctr VMA	1	CFTKGTQVMMADG	(7)	IEVG D KV	(37)	LMDFTVSAD H KLIL	78		
Ceu clpP	1	CLTSDHTVLTRG	(6)	VTLD D KV	(33)	GVDLFVTPN H RMVY	73		
Ppu dnaB	1	CISKFSHIMWSHV	(24)	LDQ G EAF	(16)	EKYLELTSN H KILT	(3)	WQRCQQLLCNDMITTQ	93
Mtu recA	1	CLAEGRTRIFDPVT	(3)	HRIE D VV	(40)	GAIVWATPD H KVLT	(3)	WRAAGELRKGDRAVQP	96
Mle recA	1	CMNYSTRVTLADG	(2)	EKIG K IV	(47)	KSQFAATPN H LIRT	(3)	WTEAGNLIAGDRVLAV	102
Mle pps1	1	CLTADARINVKGK	(7)	VQPG D EV	(31)	GRALEATGN H QFLV	(15)	WAPLEEIESG P IAVA	103
Mfl gyrA	1	CVTGDALVRLPFG	(2)	VRLR D VV	(43)	GYEVTGTAN H PLLC	(11)	WKLTEEIRPGDHVVVQ	106
Mgo gyrA	1	CLTGDALVRLPFG	(2)	MRIG D VA	(43)	GYEVTGTSN H PVLC	(11)	WMLIEEIRPD D YVVVQ	106
Mka gyrA	1	CVTGDALVRLPFG	(2)	MRIAD V V	(43)	GYEVTGTAN H PLLC	(11)	WKLIEEIHPPD D YVALQ	106
Mle gyrA	1	CVSGNSLVRLPFG	(2)	IRIG D IV	(43)	GYEITGTSN H PLLC	(11)	WKLIGEIRSGD D YVVVQ	106
Mxe gyrA	1	CITGDALVALPEG	(2)	VRIAD D IV	(43)	GLRVTGTAN H PLLC	(11)	WKLIDBIKPGD D YAVIQ	106
Ssp dnaB	1	CISGDSLISLAST	(3)	VSIK D LL	(40)	GRTIKATAN H RFLT	(3)	WKRLDELSLK E HIALP	96
Tli pol-1	1	SILPNEWLPPIEN	(5)	VKIG E FI	(61)	GRKINITAG H SLFT	(7)	EVSGDGIKEGD L IVAP	123
Psp pol-1	1	SILPEEWVPLIKN	(5)	FRIG D EV	(61)	GRKITITEG H SLFV	(7)	EATGEDVKIG D LLAVP	123
Psp pol-3	1	SILPEEWLPVLEE	(5)	VRIG E LI	(61)	GRRIKITSG H SLFS	(7)	EVTGDELKPGD L VAVP	123
Tli pol-2	1	SVSGESEIIIRQN	(5)	VKIK D LF	(53)	SWYIDVTED H SLIG	92		
Tfu pol-2	1	SVTGDTEVTIRRN	(5)	VPIE K LF	(53)	SWYLDVTED H SLIG	92		
Psp pol-2	1	CHPADTKVVVKGK	(7)	VQEG D YV	(23)	INGLKCTPN H KLPV	64		
Tfu pol-1	1	CHPADTKVIVKGK	(7)	VREG D YV	(23)	INGLKCTPN H KLPV	64		
Mja pol-1	1	CHPKGTKVVVKGK	(7)	VKEG N YV	(23)	VNGLKCTPN H KIPL	64		
Pfu RIR1-1	1	CHDGKAKIIFENE	(15)	KHLG E FY	(54)	GTKILTSPW H PFFV	(8)	EKRADELKEGD I LIGG	127
Pfu RIR1-2	1	CVVGDTRILTPEG	(1)	LKA E EIF	(75)	GYEITATLD H KLMT	(3)	WKEVGKLKEGD K ILLP	129
Mja hyp1	1	CVPPDPLLILENG	(6)	IKVGD K V	(30)	PEEIIILT P HPVYA	(37)	WIIAKDLKVG D IVVYP	123
Mja hyp2	1	CLTSNSKILTDDG	(5)	EKL K EKL	(37)	GRVLEGSKD H PVLT	(3)	YVPMGMLKEGD D VIVY	95
Mja helica.	1	CLNANTEILQESG	(6)	LNK D EKV	(32)	GLEITTT P NHIFLV	(7)	EKEAKDLKVG D YVATV	95
Mja pol-2	1	SILPDDEYLTIIEE	(5)	VKIG E YI	(61)	GRTIKVTRG H SLFK	(7)	EVKGDDVRFGD L IVVP	123
Mja GF6P	1	CLHPDPTYVILPDG	(7)	IDED E VL	(31)	FSELITT G E H KLFV	(5)	EKCVKDLNG S ELIGVV	93
Mja IF2	1	CLMPHEKVLTEYG	(1)	IKIED L F	(55)	WHSITTT P HPFLT	(3)	WIKAEINIKKG M YVAIP	109
Mja RNR-1	1	SLGRDELIFIKEG	(5)	CKIG E AI	(64)	GTSIIIVT D HSLFN	(23)	NPYDVEYRIG D YIETN	142
Mja RNR-2	1	SLPYDEKILIFEN	(5)	VKIG E FV	(62)	GKKVRVTGD H SVFT	(8)	EVKASDLKVG D FIITP	125
Mja PEPSyn.	1	CIEGDAKILDRG	(10)	VKN G EKL	(39)	KDTIKITPD H KFPV	83		
Mja RFC-1	1	CLTGDTKVIVNGE	(1)	REIG E VI	(50)	GRELKVTTY H PLLI	(8)	WEKAENLKVGD K LATP	109
Mja RFC-2	1	SVSKDTPILVKID	(33)	LTVD E NF	(24)	GGYIELTGN H SIMM	(7)	AKKASDIKVG D CFLSF	114
Mja RFC-3	1	CLTGDAKITLPDE	(2)	IKIED F I	(42)	GREIEATGD H KFLT	(3)	WKEVYELKED E VLVY	97
Mja RpoA''	1	CVDGDTTVLLDGK	(1)	IKIK D LE	(43)	GREIIATED H PFYT	(3)	RKRCGELKVG D EVIIY	97
Mja RpoA'	1	SLPYEEKIIIEKEG	(5)	VEIG K LV	(59)	GREITATPY H SFVI	(7)	PVKGSELKIG D RIPVV	121
Mja TFIIB	1	SVDYNEPIIIKEN	(5)	VKIG E LI	(58)	NKKVRVTRS H SVFT	(7)	PIRVDELKVG D ILVLA	120
Mja r-gyr	1	CLTPDPTYVVLGDG	(2)	ETIED D IV	(41)	NYELKATPD H CLLV	(7)	WIPAKDIKENDYIAMP	100
Mja UDPGD	1	CFHPDEVLFIDRG	(5)	ITFK E LF	(42)	GREIKITKD H PVVI	(7)	IKLTSVKEGD K VILP	104

^aNumbers flanking the first and last motif of each intein denote their first and last aa positions, respectively. Numbers in parenthesis between motifs denote their distance from each other in aa. Residues in positions of known and suggested functional importance are in bold. Species abbreviations: Sce, *S. cerevisiae*; Ctr, *Candida tropicalis* (yeast); Ceu, *Chlamydomonas eugametos*; Ppu, *Porphyra purpurea* (Red alga); Mtu, *Mycobacterium tuberculosis*; Mle, *Mycobacterium leprae*; Mfl, *Mycobacterium flavescens*; Mgo, *Mycobacterium goodii*; Mka, *Mycobacterium kansasii*; Mxe, *Mycobacterium xenopi*; Ssp, *Synechocystis* spp. (Cyanobacteria); Tli, *Thermococcus litoralis*; Psp, *Pyrococcus* spp.; Tfu, *Thermococcus fumicolans*; Pfu, *Pyrococcus furiosus*; Mja, *Methanococcus jannaschii*. Protein abbreviations: VMA, vacuolar ATPase catalytic subunit; clpP, Clp protease catalytic subunit; dnaB, DnaB helicase; gyrA, gyrase subunit A; pol, DNA polymerase; RIR1, ribonucleoside-diphosphate reductase alpha chain; hyp, hypothetical; helica, helicase; GF6P, glutamine-fructose 6-phosphate transaminase; IF2, translation initiation factor bIF-2; RNR, anaerobic ribonucleoside-triphosphate reductase; PEPSyn, phosphoenolpyruvate synthase; RFC, replication factor C 37-kDa subunit; RpoA'', RNA polymerase subunit A''; RpoA', RNA polymerase subunit A'; TFIIB, transcription factor IIB; r-gyr, reverse gyrase; UDPGD, UDP-glucose dehydrogenase.

Intein endonuclease domain can be of different types

Database searches with the intein motifs (Petrokovski, 1994) identified previously unreported inteins. One intein was identified in the DNA gyrase B subunit (gyrB) from the cyanobacterium *Synechocystis* sp. strain PCC6803. This intein has similar domain organization to the dnaX intein from the same species (Liu & Hu, 1997). Both have the intein N- and C-domain motifs, but are

missing the typical EN-domain motifs. However, both inteins have a ~300 aa region between the N and C domains (Fig. 4 and not shown). This region in the Ssp gyrB intein contains an HNH type endonuclease motif (Gorbalenya, 1994; Shub et al., 1994) (Fig. 4). Thus far, the central region in the Ssp dnaX intein was not found similar to any protein sequence or to contain motifs from any known homing endonucleases. The similarity to a dodecapeptide motif reported in this intein (Liu & Hu, 1997) is calculated to be

Table 2. Similarity between intein and CAD motifs

Motifs		Aligned positions	Z-score ^a	Expected occurrences by chance in searching 100 blocks
Intein N1	CAD 1	1–13/1–13	8.1	$\leq 1.0 \cdot 10^{-6}$
Intein N2	CAD 2	1–7/7–13	4.4	$8.8 \cdot 10^{-2}$
	CAD 5	1–7/6–12	5.9	$4.1 \cdot 10^{-3}$
Intein N3	CAD 4	2–14/1–13	7.4	$1.5 \cdot 10^{-4}$
Intein N4	CAD 2	2–15/2–15	7.9	$\leq 1.0 \cdot 10^{-6}$
	CAD 5	2–15/1–14	6.5	$1.2 \cdot 10^{-3}$
Intein N2	Intein N4	1–7/7–13	5.1	$2.1 \cdot 10^{-2}$
CAD 2	CAD 5	5–15/4–14	4.9	$3.9 \cdot 10^{-2}$

^aThe score cutoff is 4.0 Z-score units and corresponds to the top $1.6 \cdot 10^{-3}$ percentile of chance scores (Pietrokovski, 1996b).

insignificant—more than half of the current Swissprot database sequences (>29,500) had better scores with that motif's block.

Two new short inteins without an EN domain were identified. One is in a klbA virulence protein of *Methanococcus jannaschii* (MJ0781, gi|1591480). This is the 19th intein in the genome of this methanogenic archaeon (Bult et al., 1996). The klbA intein is 168 aa long and only contains the intein N- and C-domain motifs [chance expectant value in the *M. jannaschii* proteins database is

$1.9 \cdot 10^{-10}$ (Henikoff & Henikoff, 1994)]. The intein is probably not functional because its N-terminal residue is Ala, rather than the Cys or Ser found in all other inteins. Mutations of the N-terminal aa in different inteins (Hirata & Anraku, 1992; Hodges et al., 1992; Cooper et al., 1993; Chong et al., 1996; Xu & Perler, 1996) (including ones to Ala) and the protein splicing mechanism (Chong et al., 1996; Xu & Perler, 1996) demonstrate the necessity for an aa that could undergo an N-S/O acyl shift at the N-termini of inteins. The intein klbA host protein might be inactive due to the presence of an impaired intein. Nevertheless, the klbA function might be performed in the cell by one of two other klbA-similar *M. jannaschii* putative proteins (MJ0900, gi|1591575, and the protein formed by joining MJ1287 and MJ1288, gi|1591924 and gi|1591925). The klbA intein may be the first example of a “pseudo intein.”

The other new short intein was found in *Methanobacterium thermoautotrophicum* strain ΔH. It is the only intein identified in the 1.75-Mb genome of this lithoautotrophic, thermophilic archaeon (Smith et al., 1997). The intein is in the ribonucleoside-diphosphate reductase alpha chain (RIR1) protein (MTH652) and is only 134 aa long, the shortest intein found to date. Like the Mja klbA intein, it only contains the intein N- and C-domain motifs [chance expectant value (Henikoff & Henikoff, 1994) in the *M. thermoautotrophicum* proteins database is $1.8 \cdot 10^{-11}$] (Fig. 5). The Mth RIR1 intein is integrated in the same point as the *Pyrococcus furiosus* RIR1-2 intein (Riera et al., 1997), but the two inteins are not particularly similar to each other (Smith et al., 1997).

Fig. 3. Similarity of motif CAD 4 to a cutinase catalytic triad active site. **A:** Horizontal striped bar shows the aligned positions. The motif similarity is 6.3 Z-score units, expected to occur by chance $8.3 \cdot 10^{-2}$ times searching 3,838 blocks (the number of blocks in the searched database). The cutinase motif is from the PRINTS 15.0 database (Attwood et al., 1994), motif 5 of entry CUTINASE. Motif positions are numbered according to PDB (Bernstein et al., 1977) entry 1CUS of *Fusarium solani pisi* cutinase. Asterisks mark the aspartate and histidine active site residues. **B:** Active site of *F. solani pisi* cutinase. Catalytic triad residues are labeled, colored black, and shown with side chains. The region found similar to motif CAD 4 is colored orange. The structure is from PDB entry 1CUS.

Fig. 4. *Synechocystis* sp. *gyrB* intein. On top is the *Ssp gyrB* intein sequence (gi|1652801) with intein motifs boxed and denoted according to Figure 1. Arrow points to the intein integration site in the host *gyrB* protein, also marked with an asterisk. The integration region is aligned with the *Bacillus subtilis gyrB* (sp|P05652) protein sequence. The expectant value for the intein motifs hit in the NCBI nonredundant database is $5.6 \cdot 10^{-21}$ and $1.2 \cdot 10^{-4}$ for the HNH hit querying the Blocks databases augmented with HNH blocks. The BLAST *p*-value for the *Synechocystis* and *B. subtilis gyrB* sequence alignment is $3.0 \cdot 10^{-266}$.

These atypical inteins show that the EN domain is not only optional, missing from some inteins (Perler et al., 1997), but can be of different types. The *Ssp gyrB* intein EN domain has an HNH homing endonucleases motif in contrast to all other intein EN domains that have DOD homing endonuclease motifs. Thus, different inteins seem to have acquired different types of endonucleases. This strengthens the analogy between inteins and self-splicing introns. These group I and II introns have a similar distribution to inteins (Belfort et al., 1995) and frequently contain a homing endonuclease that enables them to home into intron-less alleles. The intron-encoded homing endonucleases are of several different types, including DOD and HNH (Belfort & Perlman, 1995). In inteins, the structural modularity of the EN domain is shown in the *Sce VMA* intein structure, where the central sequence region (includ-

ing all the EN motifs) is folded into a compact separate domain (Duan et al., 1997).

Endonuclease activity can provide mobility to intein genes (Gimble & Thorner, 1992), could be uncoupled from the protein splicing activity (Hodges et al., 1992; Kawasaki et al., 1997; Liu & Hu, 1997), and is not required for protein splicing (Chong & Xu, 1997; Telenti et al., 1997). This can explain the occurrence of inteins lacking the EN domain. Absence or loss of endonuclease activity will not immediately select against an intein because it can remove itself from the host protein. However, in the long run, this intein will not be able to disseminate itself and is more likely to be lost if it does not provide some selective advantage. One explanation for the diverse distribution of inteins is their genetic mobility (Belfort et al., 1995; Perler et al., 1997). This trait is very effective

Fig. 5. *M. thermoautotrophicum* RIR1 intein. Data are shown as in Figure 4. *Tac RIR1* denotes *Thermoplasma acidophilum* ribonucleotide reductase. (gi|1657993). The BLAST *p*-value for the *M. thermoautotrophicum* and *T. acidophilum* RIR1 sequence alignment is $1.4 \cdot 10^{-205}$.

in propagating and retaining genetic elements with no apparent selective advantage ["selfish DNA" (Crick, 1979)]. Another trait that may aid the persistence of inteins is the usual presence of their genes in highly conserved regions of essential proteins. Removing the intein DNA requires precise excision to avoid modifying crucial parts of the host gene. All of this does not preclude the possibility that some inteins evolved to have some selective advantage for their host (Davis et al., 1994; Navas, 1996; Pietrokovski, 1996a).

Conclusions

Findings presented here and other recent data suggest the following observations on intein evolution. Inteins and CADs likely share a common ancestor. It is interesting that CADs are found in proteins of multicellular organisms (indeed in proteins that control development of such organisms), whereas inteins are found in proteins of unicellular organisms and of organelles. The earliest inteins most probably lacked an EN domain and later acquired different types of homing endonuclease domains. Present-day inteins lacking an EN domain may never have acquired one or subsequently lost it. The HO endonuclease apparently is a derivative of an intein with a DOD-type EN domain adapted for mating type switch in the progenitor of the *Saccharomyces* genus.

Careful analysis of the intein sequence characteristics has been useful in elucidating their mode of function. Notable progress was achieved in identifying mutations and conditions modifying the protein splicing activity. Further advancement could be made by replacing or including proteins of interest in the optional intein EN domain. Inducible and "porter" inteins could be the basis for innovative biological tools.

Material and methods

Sequences

Sequence accession prefixes "gi|" and "sp|" denote entries from the NCBI nonredundant and the SwissProt protein databases, respectively. Intein protein sequences used in constructing the multiple alignment blocks are from the compilation of Perler et al. (1997), supplemented by four more sequences available in December 1996 (accession gi|1655695 and Riera et al., 1997). Nine intein alleles (Perler et al., 1997) with more than 55% aa identity to other inteins were excluded from determining the conserved regions (motifs) and later added to the multiple alignments of the motifs (blocks). This procedure avoids biasing the motif determination toward over-represented sequences, but allows using all sequence information in the final blocks. CAD sequences were taken from the C-terminal domain of hedgehog family proteins (for a recent compilation see Hammerschmidt et al., 1997) and of *C. elegans* proteins (Burglin, 1996; Porter et al., 1996a).

Block construction

Protein blocks were first found by the BlockMaker (Henikoff et al., 1995) and MEME (Bailey & Elkan, 1994) systems. Other regions in the sequences were then aligned using the MACAW program (Schuler et al., 1991).

Database searches

Sequence database searches were done on the NCBI WWW server using program BLASTP version 1.4.9 (Altschul et al., 1990). Block

searches of sequence databases were done with the Blimps and Multimat programs (Henikoff et al., 1995). The searched sequence databases were NCBI nonredundant protein database (release February 9, 1997), the *M. jannaschii* protein database (Bult et al., 1996), the *Synechocystis* sp. protein database (Kaneko et al., 1996), and the *M. thermoautotrophicum* database (Smith et al., 1997). The LAMA program (Pietrokovski, 1996b) was used for block versus block comparisons and block database searches. Information on the new inteins identified in this work was deposited in the Intein Database (Perler et al., 1997).

Note added in proof

Recent papers appearing after the acceptance of this work contain results relevant to its conclusions. A paper by Dalgaard et al. (1997) reports some of the new inteins described in this paper and provides a multiple alignment of inteins and CAD sequences using different sequence analysis methods. Derbyshire et al. (1997) constructed functional inteins lacking the central endonuclease domain, providing another example that all the information for protein splicing is found in the N- and C-terminal domains and that the EN domain is optional for this reaction. Tanaka Hall et al. (1997) determined the crystal structure of a hedgehog CAD domain and also provided a multiple sequence alignment of inteins and CADs. The CAD structure is indeed found to be similar to the yeast intein structure and to comprise two homologous, and apparently duplicated, subdomains. The active site is suggested to consist of the N-terminal Cys, an Asp, Thr, and His residues (respectively, positions 1 in motif CAD1, 9 in CAD3, and 6 and 8 in CAD4). This confirms the suggested active site role of the histidine. However, the Gly-Asp site suggested in this paper to be part of the active site (positions 10–11 in CAD2) have apparent structural roles. Only part of the CAD active site is probably relevant to the intein protein splicing active site because inteins protein splice without cholesterol, the CAD Asp active site residue has no apparent corresponding residue in intein sequences, and its mutation in a hedgehog CAD only affected the cholesterol-stimulated reaction (Tanaka Hall et al., 1997).

Acknowledgments

I thank F. Perler, A. Telenti, R. Roberts, M. Belfort, D. Leahy, and V. Derbyshire for providing manuscripts before publication and E.A. Greene, S. Henikoff, R.J. Monnat, Jr., and Z. Kelman for commenting on the manuscript. S.P. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. This work was supported by an NIH grant (GM29009).

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410.
- Attwood TK, Beck ME, Bleasby AJ, Parry-Smith DJ. 1994. PRINTS—A database of protein motif fingerprints. *Nucleic Acids Res* 22:3590–3596.
- Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings of Second International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, California: AAAI Press. pp 28–36.
- Belfort M, Perlman PS. 1995. Mechanisms of intron mobility. *J Biol Chem* 270:30237–30240.
- Belfort M, Reaban ME, Coetzee T, Dalgaard JZ. 1995. Prokaryotic introns and inteins: A panoply of form and function. *J Bacteriol* 177:3897–3903.
- Belfort M, Roberts R. 1997. Homing endonucleases—Keeping the house in order. *Nucleic Acids Res* 25:3379–3388.

- Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol* 112:535–542.
- Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Weidman JF, Fuhrmann JL, Presley EA, Nguyen D, Utterback TR, Kelley JM, Peterson JD, Sadow PW, Hanna MC, Cotton MD, Hurst MA, Roberts KM, Kaine BP, Borodovsky M, Klenk HP, Fraser CM, Smith HO, Woese CR, Venter JC. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273:1058–1073.
- Burglin TR. 1996. Warthog and groundhog, novel families related to hedgehog. *Curr Biol* 6:1047–1050.
- Chong S, Shao Y, Paulus H, Benner J, Perler FB, Xu MQ. 1996. Protein splicing involving the *Saccharomyces cerevisiae* VMA intein: The steps in the splicing pathway, side reactions leading to protein cleavage, and establishment of an in vitro splicing system. *J Biol Chem* 271:22159–22168.
- Chong S, Xu MQ. 1997. Protein splicing of the *Saccharomyces cerevisiae* VMA intein without the endonuclease motifs. *J Biol Chem* 272:15587–15590.
- Cooper AA, Chen YJ, Lindorfer MA, Stevens TH. 1993. Protein splicing of the yeast TFP1 intervening protein sequence: A model for self-excision. *EMBO J* 12(6):2575–2583.
- Crick F. 1979. Split genes and RNA splicing. *Science* 204:264–271.
- Dalgaard JZ, Moser MJ, Hughey R, Mian IS. 1997. Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J Comput Biol* 4:193–214.
- Davis EO, Thangaraj HS, Brooks PC, Colston MJ. 1994. Evidence of selection for protein introns in the recAs of pathogenic mycobacteria. *EMBO J* 13:699–703.
- Derbyshire V, Wood DW, Wu W, Dansereau J, Dalgaard JZ, Belfort M. 1997. Genetic definition of a protein splicing domain: Functional mini-inteins support structure predictions and a model for intein evolution. *Proc Natl Acad Sci USA* 94:11466–11471.
- Duan X, Gimble FS, Quijcho FA. 1997. Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell* 89:555–564.
- Fsihi H, Vincent V, Cole ST. 1996. Homing events in the *gyrA* gene of some mycobacteria. *Proc Natl Acad Sci USA* 93:3410–3415.
- Gimble FS, Stephens BW. 1995. Substitutions in conserved dodecapeptide motifs that uncouple the DNA binding and DNA cleavage activities of PI-SceI endonuclease. *J Biol Chem* 270:5849–5856.
- Gimble FS, Thorne J. 1992. Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. *Nature* 357:301–306.
- Gorbalenya AE. 1994. Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. *Protein Sci* 3:1117–1120.
- Hammerschmidt M, Brook A, McMahon AP. 1997. The world according to hedgehog. *Trends Genet* 13:14–21.
- Henikoff S, Henikoff JG. 1994. Protein family classification based on searching a database of blocks. *Genomics* 19:97–107.
- Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S. 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* 163:GC 17–26.
- Hirata R, Anraku Y. 1992. Mutations at the putative junction sites of the yeast VMA1 protein, the catalytic subunit of the vacuolar membrane H(+)-ATPase, inhibit its processing by protein splicing. *Biochem Biophys Res Commun* 188:40–47.
- Hirata R, Ohsumi Y, Nakano A, Kawasaki H, Suzuki K, Anraku Y. 1990. Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. *J Biol Chem* 265:6726–6733.
- Hodges RA, Perler FB, Noren JN, Jack WE. 1992. Protein splicing removes intervening sequences in an archaea DNA polymerase. *Nucleic Acids Res* 20:6153–6157.
- Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirokawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Raki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* 3:109–136.
- Kawasaki M, Nogami S, Satow Y, Ohya Y, Anraku Y. 1997. Identification of three core regions essential for protein splicing of the yeast vma1 protozyme. A random mutagenesis study of the entire VMA1-derived endonuclease sequence. *J Biol Chem* 272:15668–15674.
- Koonin EV. 1995. A protein splice-junction motif in hedgehog family proteins. *Trends Biochem Sci* 20:141–142.
- Lee JJ, Ekker SC, von KDP, Porter JA, Sun BI, Beachy PA. 1994. Autoproteolysis in hedgehog protein biogenesis. *Science* 266:1528–1537.
- Liu XQ, Hu Z. 1997. Identification and characterization of a cyanobacterial DnaX intein. *FEBS Lett* 408:311–314.
- Martinez C, Nicolas A, van Tilbeurgh H, Egloff MP, Cudrey C, Verger R, Cambillau C. 1994. Cutinase, a lipolytic enzyme with a preformed oxyanion hole. *Biochemistry* 33:83–89.
- Navas J. 1996. Genetic tools in pathogenic nocardioform actinomycetes. *Micobiologia* 12:297–304.
- Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschuere KHG, Goldman A. 1992. The alpha/beta hydrolase fold. *Protein Eng* 5:197–211.
- Perler FB, Comb DG, Jack WE, Moran LS, Qiang B, Kucera RB, Benner J, Slatko BE, Nwankwo DO, Hempstead SK, Carlow CKS, Jannasch H. 1992. Intervening sequences in an Archaea DNA polymerase gene. *Proc Natl Acad Sci USA* 89:5577–5581.
- Perler FB, Olsen GJ, Adam E. 1997. Compilation and analysis of intein sequences. *Nucleic Acids Res* 25:1087–1093.
- Pietrokovski S. 1994. Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein Sci* 3(12):2340–2350.
- Pietrokovski S. 1996a. A new intein in Cyanobacteria and its significance for the spread of inteins. *Trends Genet* 12:287–288.
- Pietrokovski S. 1996b. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res* 24:3836–3845.
- Porter JA, Ekker SC, Park WJ, von Kessler DP, Young KE, Chen CH, Ma Y, Woods AS, Cotter RJ, Koonin EV, Beachy PA. 1996a. Hedgehog patterning activity: Role of a lipophilic modification mediated by the carboxy-terminal autoprocessing domain. *Cell* 86:21–34.
- Porter JA, Young KE, Beachy PA. 1996b. Cholesterol modification of hedgehog signaling proteins in animal development. *Science* 274:255–259.
- Rawlings ND, Barrett AJ. 1994. Families of serine peptidases. *Methods Enzymol* 244:19–61.
- Riera J, Robb FT, Weiss R, Fontecave M. 1997. Ribonucleotide reductase in the archaeon *Pyrococcus furiosus*: A critical enzyme in the evolution of DNA genomes? *Proc Natl Acad Sci USA* 94:475–478.
- Schuler GD, Altschul SF, Lipman DJ. 1991. A workbench for multiple alignment construction and analysis. *Protein Struct Funct Genet* 9:180–190.
- Shub D, Goodrich-Blair H, Eddy S. 1994. Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *Trends Biochem Sci* 19:402–404.
- Smith DR, Doucette-Stamm LA, Delonghery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lum W, Pothier B, Qiu D, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush D, Safer H, Patwell D, Prabhakar S, McDougall S, Shimer G, Goyal A, Pietrokovski S, Church GM, Daniels CJ, Mao JJ, Rice P, Nöling J, Reeve JN. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* strain H: Functional analysis and comparative genomics. *J Bacteriology* 179:7135–7155.
- Tanaka Hall TM, Porter JA, Young KE, Koonin EV, Beachy PA, Leahy DJ. 1997. Crystal structure of a hedgehog autoprocessing domain: Homology between hedgehog and self-splicing proteins. *Cell* 91:85–97.
- Telenti A, Southworth M, Daugelat S, Alcaide F, Jacobs WRJ, Perler FB. 1997. The *Mycobacterium xenopi* *gyrA* intein: Characterization of a minimal intein. *J Bacteriology* 179:6378–6382.
- Valentini RP, Brookhiser WT, Park J, Yang T, Briggs J, Dressler G, Holzman LB. 1997. Post-translational processing and renal expression of mouse Indian hedgehog. *J Biol Chem* 272:8466–8473.
- Xu MQ, Perler FB. 1996. The mechanism of protein splicing and its modulation by mutation. *EMBO J* 15:5146–5153.