Protein Science (1994), *3*:2340–2350. Cambridge University Press. Printed in the USA.

Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins

SHMUEL PIETROKOVSKI

Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel

(RECEIVED August 10, 1994; ACCEPTED October 6, 1994)

Abstract

Inteins (protein introns) are internal portions of protein sequences that are posttranslationally excised while the flanking regions are spliced together, making an additional protein product. Inteins have been found in a number of homologous genes in yeast, mycobacteria, and extreme thermophile archaebacteria. The inteins are probably multifunctional, autocatalyzing their own splicing, and some were also shown to be DNA endonucleases. The splice junction regions and two regions similar to homing endonucleases were thought to be the only common sequence features of inteins.

This work analyzed all published intein sequences with recently developed methods for detecting weak, conserved sequence features. The methods complemented each other in the identification and assessment of several patterns characterizing the intein sequences. New intein conserved features are discovered and the known ones are quantitatively described and localized. The general sequence description of all the known inteins is derived from the motifs and their relative positions. The intein sequence description is used to search the sequence databases for intein-like proteins. A sequence region in a mycobacterial open reading frame possessing all of the intein motifs and absent from sequences homologous to both of its flanking sequences is identified as an intein. A newly discovered putative intein in red algae chloroplasts is found not to contain the endonuclease motifs present in all other inteins. The yeast HO endonuclease is found to have an overall intein-like structure and a few viral polyprotein cleavage sites are found to be significantly similar to the inteins amino-end splice junction motif. The intein features described may serve for detection of intein sequences.

Keywords: database searches; dodecapeptide motif; endonucleases; LAGLI DADG; polyproteins; posttranslational processing; protein motifs; protein splicing; sequence analysis

Protein splicing is a recently discovered posttranslational process in which an internal segment from a precursor protein is excised and the flanking regions rejoined, thus creating two protein products (Kane et al., 1990). The process is apparently self-catalyzed by the internal region, termed intein (Xu et al., 1993; Perler et al., 1994). Inteins have been found in homologous proteins from phylogenetically diverse species: the 69-kDa subunit of vacuolar ATPase of the yeasts *Saccharomyces cerevisiae* (Kane et al., 1990) and *Candida tropicalis* (Gu et al., 1993), the recA proteins of the mycobacteria *Mycobacterium tuberculosis* (Davis et al., 1991, 1992) and *Mycobacterium leprae* (Davis et al., 1994), the DNA polymerases of the extreme thermophilic archaebacteria *Thermococcus litoralis* (Hodges et al., 1992; Perler et al., 1992) and *Pyrococcus* species strains GB-D (Xu et al., 1993) and KOD1 (T. Imanaka, GenBank entry D29671 submission, 1994), and the pps1 open reading frame of *Mycobacterium leprae* (this work). The only apparent common feature of the spliced proteins is the presence of a nucleotide-binding domain near or within their intein integration region (Neff, 1993).

Inteins are difficult to identify from sequence data because they are in the same reading frame as the spliced protein and only a few short conserved sequence features have been reported to characterize them. In addition, no convenient biological assay exists for protein splicing. If the sequence of an intein-less protein homolog to the spliced protein is known, the intein presence can be recognized by the similarity of its exteins (sequence regions flanking the inteins; Perler et al., 1994) to the homologous sequence (Neff, 1993). Such a scheme can be practically deployed by computational sequence analysis. However, this type of a search requires pairwise comparison of sequences and

Reprint requests to Shmuel Pietrokovski at his present address: Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, Washington 98104; e-mail: pietro@sparky.fhcrc.org.

Abbreviations: aa, amino acid(s); C′, carboxyl; cp, chloroplast; dod, dodecapeptide; mt, mitochondrial; N′, amino; open reading frame, ORF.

the time for such comparisons is proportional to the square of the sequence lengths. This analysis has to be performed for each sequence suspected of harboring an intein. The large number of known and potential protein sequences makes this approach time consuming. Of course, this approach requires intein-less homologs to identify inteins.

An alternative approach for identifying inteins is to search the sequence databases for protein sequences with conserved intein features. Such a search is not for a specific intein, but for any sequence containing their common features. Each database sequence is separately compared with each relatively short pattern. The time for such a search is linearly related to the length of the analyzed sequences. The approach is directed to search only for the conserved intein features that are probably relevant to their structure and/or function. No intein-less homologs are required for the method and its results can improve as we know more about the conserved features of inteins.

The currently known inteins were reported to have rather dissimilar sequences, apart from the splice junction regions and dod[1] motifs (Davis et al., 1992, 1994; Cooper & Stevens, 1993; Xu et al., 1993; Anraku & Hirata, 1994). The present study reports the identification of additional conserved sequence features and refinement of the known ones. Searching for sequences containing the intein-conserved sequence features is demonstrated to be a useful tool for identifying intein sequences. The study identifies a number of proteins that were found to have regions significantly similar to the intein sequence features. By examining the known functions of proteins with intein sequence features, possible roles for these features are suggested.

Results and discussion

Mycobacterium leprae pps1 intein

Sequence blocks (Henikoff & Henikoff, 1991) of the conserved dod elements and the C′ end were constructed from the *Sce VMA*, *Ctr VMA*, *Mtu recA*, *Mle recA*, *Tli pol* 1, *Tli pol* 2, and *Psp pol* 1 inteins.

A search of the GenBank sequence database found the *M. leprae* pps1 ORF (K. Robison, GenBank entry U00013 submission, 1993) to have a significant similarity ($p < 3.4 \times 10^{-8}$) to the sequence blocks. The pps1 sequences flanking the region with the intein-like features are homologous to sequences of ORFs from the probable plastid genome of *Plasmodium falciparum* (Williamson et al., 1994) and the chloroplast genomes of red algae (Kostrzewa & Zetsche, 1992, 1993; M. Reith, pers. comm.) (Fig. 1). The pps1 ORF was previously suspected to contain an intein (D.R. Smith, GenBank entry U00013 annotation, 1993) due to its similarity to the HO endonuclease (Russell et al., 1986) and *Tli pol* 2 intein (Perler et al., 1992). The proposed intein has a length of 386 aa (Fig. 2) and is found to have all of the recognized intein motifs (following section; Table 1).

The pps1 ORF is the second protein to be identified as containing an intein in *M. leprae*, after the recA protein (Davis et al., 1994). The function of the pps1 ORF and its plastid-encoded homologs is unknown. The intein was named *Mle pps1*, according to published nomenclature conventions, and added to the intein registry with the designation "theoretical" (Perler et al., 1994).

Identification of intein conserved regions

Encouraged by the usefulness of the recognized intein conserved features in identifying of the pps1 intein, a systematic search for all conserved sequence features was done on all 10 published intein sequences. Seven conserved regions were found in the intein sequences using the ASSET and MACAW programs (Table 1; Fig. 3). In addition to the previously reported splice junctions and two dod motifs (Davis et al., 1992, 1994; Hodges et al., 1992; Shub & Goodrich-Blair, 1992; Cooper & Stevens, 1993; Doolittle, 1993; Gimble & Thorner, 1993; Gu et al., 1993; Xu et al., 1993; Anraku & Hirata, 1994), three more conserved motifs were detected.

In addition to their sequence conservation, the locations of the motifs are also conserved relative to the intein ends and to each other (Fig. 4). Motif A starts at the aa preceding the N′ end of the inteins and motif G ends at the aa following the C′ end of the inteins. The C′ splice junction area is composed of two motifs (F and G) that are either consecutive or separated by one or two aa. The second dod motif (E) is preceded by motif D by six or nine aa. The distance between the two dod motifs (C and E) is also conserved (see below).

Motif A

Motif A defines the N′ end splice junction, but is only found in the inteins having a cysteine at their N′ end. The motif is not found in the *Tli pol* 1, *Tli pol* 2, *Psp pol* 1, and *Psp pol* 3 inteins, all of which have serine at their N′ end. Nevertheless, cysteine and serine are similar chemically and these residues at the intein's N′ end may have the same role in protein splicing process. If a motif describing the N′ end of all inteins does exist, the currently available number of sequences is too small for it to be detected. When more intein sequences are known, the N′ end area should be reexamined and a statistically meaningful motif might be found for all the inteins. Alternatively, inteins with serine N′ end might have a distinct motif analogous to motif A. At present only two distinct intein sequences with serine N′ end are known (see Sequences section below).

Motif B

One of the models for protein splicing suggests an N → O acyl shift of serine or threonine residues at the splice sites with the assistance of a histidine residue, and suggests the conserved histidine at the C′ splice junction as the necessary histidine (Hodges et al., 1992; Wallace, 1993). Substituting the C′ splice junction histidine in the Tfp1 protein allowed some splicing activity (Cooper et al., 1993), and it was suggested that another histidine residue in the intein might perform the catalytic function (Cooper & Stevens, 1993). The seventh residue in motif B is invariably a histidine and might fulfill the required function.

Motifs C and E

Motifs C and E define the dod elements. These motifs are characteristic of dod homing endonucleases found in organellar and nuclear genomes of many species, mostly in group I introns (Doolittle, 1993). In homing endonucleases, the elements

[1] These motifs are also known as decamers (Waring et al., 1982), P1/P2 (Michel et al., 1982), and LAGLI-DADG (Hensgens et al., 1983). However, the term dod (dodecapeptide) seems to better describe these motifs, whose conserved length is usually 12 aa.

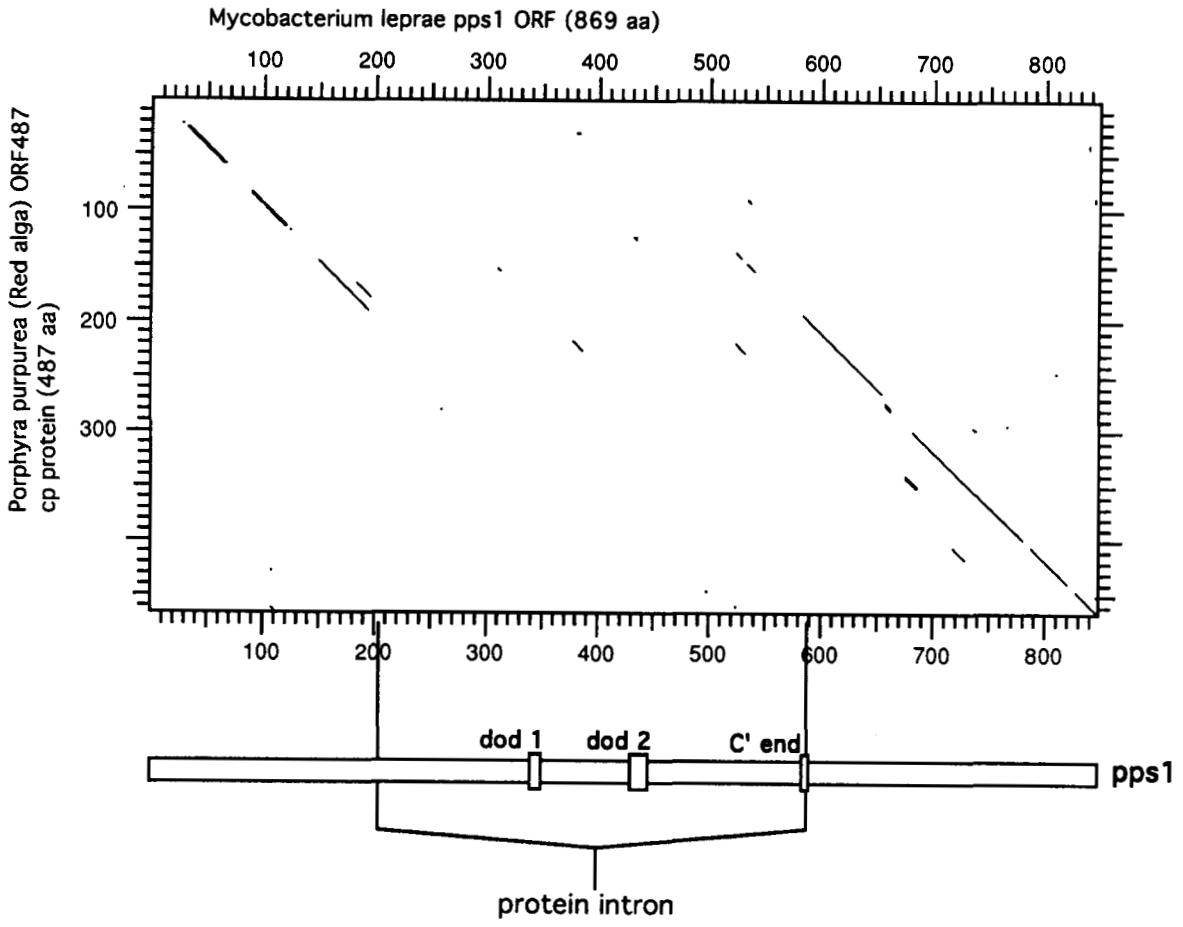

Fig. 1. Similarity of the *M. leprae* pps1 ORF to a red alga chloroplast ORF. The dot plot identifies 23-aa windows that have at least seven identities. The regions identified by the blocks are marked with the block names. The dot plot of the pps1 ORF and ORF 470 from *P. falciparum* (Williamson et al., 1994) give a very similar result. Data for pps1 from GenBank accession U00013, for ORF 487 from M. Reith (pers. comm.).

are typically 80–160 aa apart and 12 aa long (Lambowitz & Belfort, 1993; S. Pietrokovski, unpubl. results). The intein dod elements have similar lengths and distances between them. The lengths of the two intein dod motifs were found to be 9 and 14 aa, respectively, and their distances from each other 92–133 aa (Table 1). The intein dod elements are very similar to the homing endonuclease dod elements, yet some differences are apparent (Fig. 5). The most obvious resemblance is in the prevalence of glycines in the 4th and 10th positions of both elements and in the 8th position of the first dod element, and the presence of an acidic aa (aspartate or glutamate) in the 9th position of both elements and in the 7th position of the second element. The main differences between the intein dod elements and the *Podospora anserina* and *S. cerevisiae* mitochondrial dod elements are the absence of an acidic aa from the 7th position of the first intein element, the presence of an aromatic aa (phenylalanine, tryptophan, and tyrosine) in the first and last positions of the fungal dod elements, and the lack of conservation in the 1st, 11th, and 12th positions of the intein first dod element.

The differences between the intein and *P. anserina* and *S. cerevisiae* mt dod elements might be due to nonspecific sequence variation or to possible constraints in the intein sequences. The significance of these differences might be clarified by the determination of more intein sequences and the elucidation of the roles of the dod elements in the endonuclease function.

Sequences with intein motifs

The identified conserved sequence features from all 10 intein sequences were used to construct sequence blocks. The SwissProt and GenBank sequence databases were searched for sequences with regions similar to the blocks and having the same order and distances from one another as the blocks.

Porphyra purpurea chloroplast dnaB intein

The cp dnaB protein of the red alga *Porphyra purpurea* contains a 150-aa region that is not found in homologous dnaB proteins (M. Reith, pers. comm.). The region contains segments significantly similar to intein motifs B, F, and G (Fig. 6). The region starts with a cysteine, and its first 14 aa, together with 1 aa preceding it, make a significant motif with the corresponding positions of the N′ splice junction motif (not shown). The region is most likely an intein, but one that lacks the dod and D motifs.

```
         .        20         .        40         .        60         .        80
mtrtsettkspapelltqqqaidslgkygygwadsdvagasarrglsedvvrdisakkdepewmlqarlkalrvferkpm

         .       100         .       120         .       140         .       160
prwgsnldgidfdnikyfvrstekqaaswdelpedirntydrlgipdaekqrlvagvaaqyesevvyhqiradlkdqgvv

         .       180         .       200         .       220         .       240
fldtetglreypdifkqylgtvipagdnkfsalntavwsggCLTADARINVKGKGLVSIADVQPGDEVFGVNIGCELERG

         .       260         .       280         .       300         .       320
KVLAKVASGTKPVYEMHVAGRALEATGNHQFLVARRVEEGKRTRWTAVWAPLEEIESGEPIAVARVLPDDSGTIFFSESE

         .       340         .       360         .       380         .       400
LDIKNRTRQCLYFPCQNSVDLLWLLGLWLGDGHTAAPHKHMRQVAFSVPAGDPVHHTAIRVVSEQFGANVTVVNCGFIVS

         .       420         .       440         .       460         .       480
SKAFETWLAELGFSGDEKTKRLPAWIYSLPHEHQLALIGGLVDADGWTESSGATMSIAFASRELLEDVRQLAIGCGLYPD

         .       500         .       520         .       540         .       560
GRLVERTRSATCRDGRIVTSTSWRLRIQGSLDRVGTRTPGKRGKPVSNKGRRQRYVAAAGLNFSSLSTDTVGFARLKSKT

         .       580         .       600         .       620         .       640
LVGEKPTYDIQVVGLENFVANGIVAHNsfiyvppgvhvdiplqayfrintenmgqfertliiadtgsyvhyvegctapiy

         .       660         .       680         .       700         .       720
ksdslhsavveiivkpharvryttiqnwsnnvynlvtkrarvetgatmewidgnigskvtmkypavwmtgehakgevlsv

         .       740         .       760         .       780         .       800
afagegqhqdtgakmlhlasntssnivsksvargggrtsyrglvqvnkgahgsrssvkcdallvdtisrsdtypyvdire

         .       820         .       840         .       860
ddvtmgheatvskvsenqlfylmsrglaedeamamvvrgfvepiakelpmeyalelnrlielqmegavg
```

Fig. 2. Sequence of the intein in the *M. leprae* pps1 ORF. Numbers indicate positions in the pps1 ORF. The intein sequence (positions 202–587) is in upper case.

Unlike the sequences of all the inteins known so far, the putative dnaB intein does not include the dod motifs that also characterize the dod homing endonucleases. Several inteins are endonucleases (Perler et al., 1994), but this function was shown to be independent from the protein splicing activity in the *Tli pol* 2 intein (Hodges et al., 1992). Thus, it might be that the dod motifs are required for the endonuclease activity of both inteins and dod homing maturases, but are unnecessary for the protein splicing activity of inteins.

HO endonuclease

The HO endonuclease (Russell et al., 1986) is found to have significant similarities to six of the seven intein motifs (Fig. 7). This yeast protein is a site-specific double-stranded endonuclease that initiates the mating-type switch by cutting a specific locus. The HO endonuclease was previously recognized to be 33% identical to one intein sequence, the *Sce VMA* intein (Hirata et al., 1990; Kane et al., 1990). The similarity found in this work shows this endonuclease to resemble the intein family, not just one of its members.

Most notable are the position of the region similar to the N′ splice junction motif and the absence of the C′ splice junction motif. The N′ end of the HO endonuclease is similar to the N′ splice junction motif (motif A), with the initiation methionine corresponding to the intein's first aa. The first aa in the motif (corresponding to the last aa of the N′ extein) does not have a counterpart in the HO sequence. The only intein motif not found in the HO endonuclease is the C′ splice junction (motif G).

The HO protein is a site-specific DNA endonuclease and is not known to undergo protein splicing. However, four of the known inteins were found to be endonucleases (Perler et al., 1994). Hence, it can be assumed that at least some of the intein motifs found in the HO endonuclease are related to the endonuclease activity, whereas the missing motif G is not. The intein-similar region of the HO endonuclease ends at the start of the potential DNA-binding zinc finger domain of the protein (Rus-

Table 1. *Conserved sequence blocks of inteins*

Block A (N' end) [a] P=6.7*10-6 [b]

Intein [c]	sequence	position [d]
Sce VMA	GCFAKGTNVLMADGSIECIENIEVGNKVMG	0-29
Ctr VMA	GCFTKGTQVMMADGADKSIESIEVGDKVMG	0-29
Mtu recA	KCLAEGTRIFDPVTGTTHRIEDVVDGRKPI	0-29
Mle recA	GCMNYSTRVTLADGSTEKIGKIVNNKMDVR	0-29
Tli pol 1	-	-
Tli pol 2	-	-
Psp pol 1	-	-
Psp pol 2	RCHPADTKVVVKGKGIINISEVQEGDYVLG	0-29
Psp pol 3	-	-
Mle pps1	GCLTADARINVKGKGLVSIADVQPGDEVFG	0-29

Block B P=9.6*10-3

Sce VMA	FTCNATHEL	73- 81
Ctr VMA	FTVSADHKL	68- 76
Mtu recA	VWATPDHKV	67- 75
Mle recA	FAATPNHLI	73- 81
Tli pol 1	INITAGHSL	90- 98
Tli pol 2	IDVTEDHSL	82- 90
Psp pol 1	ITITEGHSL	90- 98
Psp pol 2	LKCTPNHKL	54- 62
Psp pol 3	IKITSGHSL	90- 98
Mle pps1	LEATGNHQF	62- 70

Block C (first dod) P=1.5*10-14

Sce VMA	LLGLWIGDG	211-219
Ctr VMA	LLGTWAGIG	202-210
Mtu recA	LLGYLIGDG	115-123
Mle recA	VLGSLMGDG	115-123
Tli pol 1	LLGYYVSEG	282-290
Tli pol 2	LVGLIVGDG	148-156
Psp pol 1	LLGYYVSEG	281-289
Psp pol 2	LAGILLAEG	118-126
Psp pol 3	LLGYYVSEG	281-289
Mle pps1	LLGLWLGDG	143-151

Block D P=1.0 [e]

Sce VMA	KNIPSFL	301-307
Ctr VMA	KSIPQHI	319-325
Mtu recA	KTIPNWF	195-201
Mle recA	-	-
Tli pol 1	KRIPSVI	359-365
Tli pol 2	RKIPEFM	228-234
Psp pol 1	KRVPEVI	358-364
Psp pol 2	-	-
Psp pol 3	KRIPEFV	358-364
Mle pps1	KRLPAWI	219-225

Block E (second dod) P=2.7*10-10

Sce VMA	TFLAGLIDSDGYVT	317-330
Ctr VMA	SLIAGLVDAAGNVE	335-348
Mtu recA	NLLFGLFESDGWVS	213-226
Mle recA	LVLAIWYMDDGSFT	209-222
Tli pol 1	SFLEAYFTGDGDIH	375-388
Tli pol 2	AFLRGLFSADGTVT	244-257
Psp pol 1	AFLEGYFIGDGDVH	374-387
Psp pol 2	SVLRGFFEGDGSVN	216-229
Psp pol 3	AFLEGYSSAMATST	374-387
Mle pps1	ALIGGLVDADGWIE	235-248

Block F P=1.3*10-10

Sce VMA	DYYGITLSDDSDHQFLLAN	430-448
Ctr VMA	NYYIGTLAEETDHQFLLSN	447-465
Mtu recA	RARTFDLEVEELHTLVAEG	417-435
Mle recA	SMNRFDIEVEGNHNYFVDG	342-360
Tli pol 1	EGYVYDLSVEDNENFLVGF	513-531
Tli pol 2	EGYVYDIEVEETHRFFANN	367-385
Psp pol 1	DGYVYDLSVDEDENFLAGF	512-530
Psp pol 2	EGKVYDLTLEGTPYYFANG	337-355
Psp pol 3	DGYVYDLSVEDNENFLVGF	511-529
Mle pps1	EKPTYDIQVVGLENFVANG	363-381

Block G (C' end) [f] P=6.0*10-8

Sce VMA	VVVHNC	450-455
Ctr VMA	ALVHNC	467-472
Mtu recA	VVVHNC	436-441
Mle recA	VMVHNS	361-366
Tli pol 1	LYAHNS	534-539
Tli pol 2	ILVHNT	386-391
Psp pol 1	LYAHNS	533-538
Psp pol 2	ILTHNS	356-361
Psp pol 3	VYAHNS	532-537
Mle pps1	IVAHNS	382-387

[a] Blocks correspond to the motifs discussed in the text and use the same names.

[b] *p*-Values calculated by the MACAW program, taking the whole length of the eight intein sequences (excluding the *Psp pol* 1 and 3 inteins, see Materials and sequences) as the searched sequence space.

[c] Intein names according to Perler et al. (1994).

[d] The first aa in block A is outside the intein sequences, being the last aa in the preceding exteins.

[e] A significant *p*-value for block D (2.3×10^{-4}) is found when only the regions between blocks C and E are taken as the searched sequence space.

[f] The last aa in block G is outside the intein sequences, being the first aa in the following exteins.

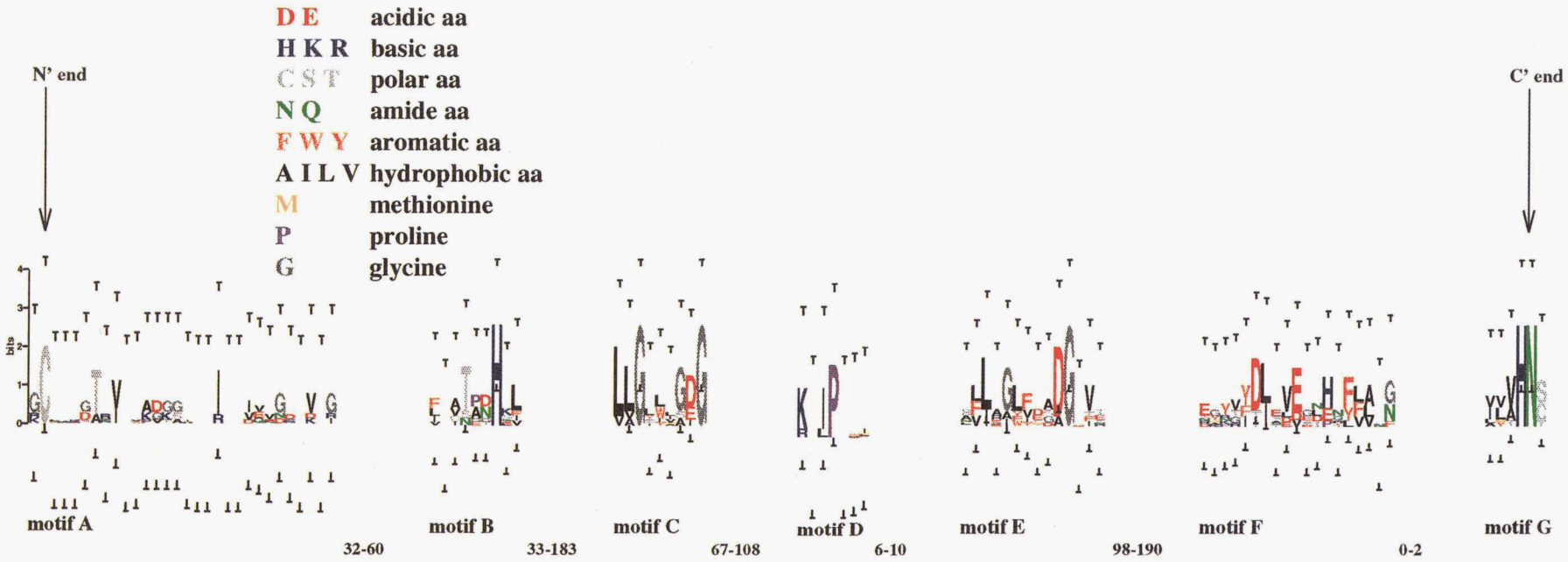

Fig. 3. Sequence logo of the intein-conserved motifs. The height of each aa in a position is proportional to its frequency there. The aa at each position are stacked according to their frequency, the most common being on top. The total height of each position is adjusted to signify the information content (conservation) of the aa at that position (Schneider & Stephens, 1990). The logo was constructed from all the intein sequences excluding the *Psp pol* 1 and 3 sequences (see Methods and sequences). The minimal and maximal number of intervening aa is indicated between the motifs names. Error bars at each position show ±1 standard deviation of its information content. The variance is due to the number of sequences used to calculate the information content (Schneider et al., 1986). Only the ends of the error bars are shown. Arrows mark the intein boundaries.

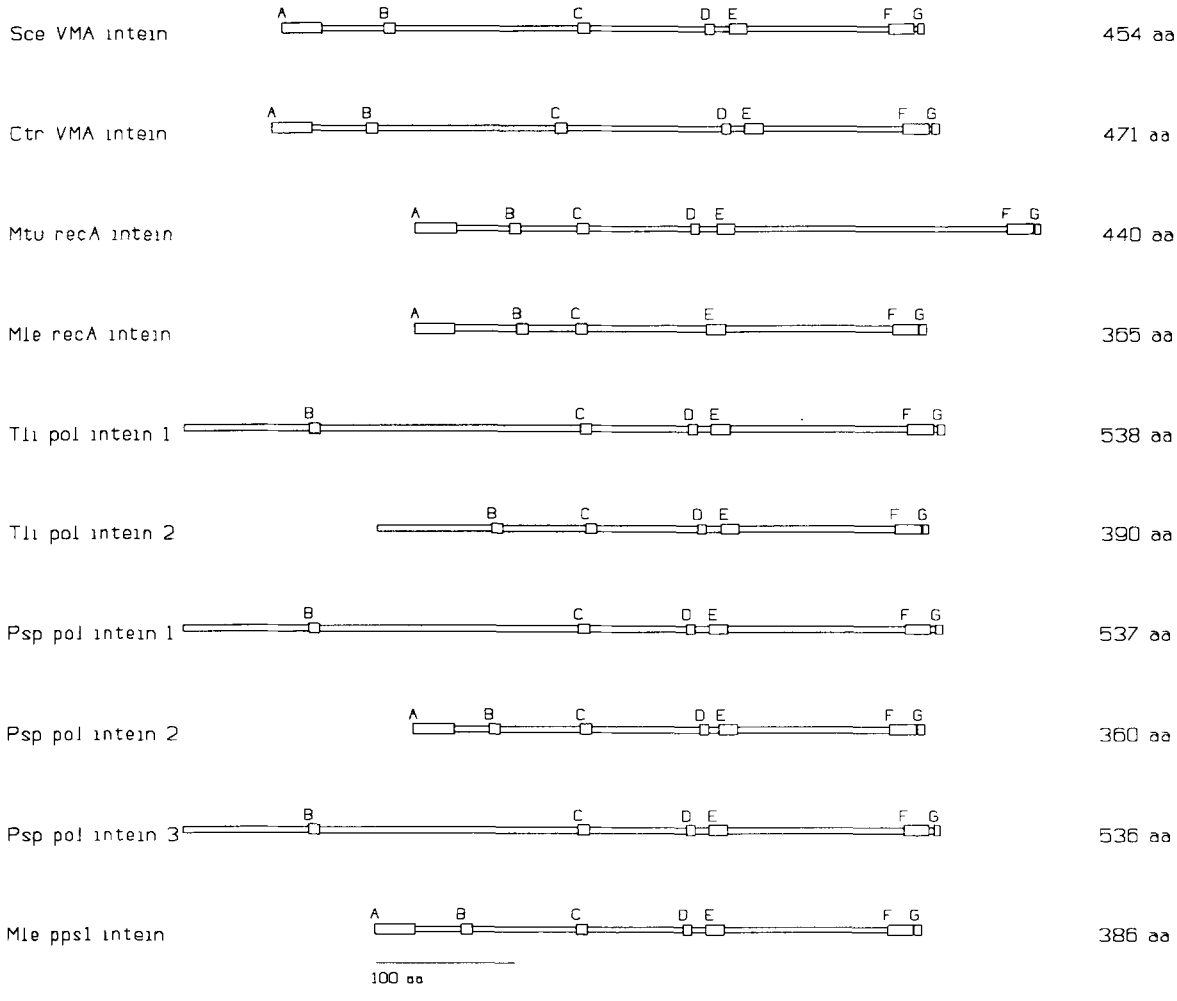

Fig. 4. Positions of the inteins' conserved motifs. A schematic diagram of the intein sequences aligned along the dod motifs (motifs C and E). The intein names and lengths are beside each sequence. Rectangles indicate motifs and are marked with their names.

sell et al., 1986). This may be further indication for the domain structure of the HO endonuclease.

The resemblance of the HO endonuclease and intein family indicates that the HO endonuclease may have originated from an intein sequence that lost its protein-splicing ability by changes in its splice junction regions. Alternatively, inteins may have originated from the HO endonuclease or perhaps both have a common origin. According to the HO origin hypothesis, the C′ splice junction was lately acquired by inteins. This scheme seems to be more complicated than the mere loss of this junction in the HO endonuclease implied in the common origin and the intein origin hypotheses.

The HO endonuclease seems to be related to the dod homing endonucleases both in sequence and function. The two contain the dod elements and mediate the homing of specific sequences into particular genomic locations by cutting that site (Strathern et al., 1982; Kostriken et al., 1983; Dujon, 1989). However, the HO endonuclease also includes intein-specific motifs, thus its resemblance to inteins cannot be due only to the similarity of both to the dod endonuclease family. In light of this study's finding, perhaps the HO endonuclease is in some way an intermediate between the intein and homing endonuclease families. It will be interesting to see if protein-splicing activity can be induced by replacing the HO endonuclease initiation methionine with cysteine, introducing the C′ splice junction motif at the appropriate position (1–4 aa from the end of motif F), and inserting the resulting sequence inside another protein sequence.

Polyproteins

Polyproteins undergo specific proteolytic processes similar to those of protein splicing (Wallace, 1993). For example, the final proteolytic cleavage of the picornaviruses polyprotein is apparently autocatalytic (Arnold et al., 1987) and some of the recently discovered prokaryotic polyproteins contain spacer regions that are removed in the polyprotein maturation (Thony-Meyer et al., 1992). The posttranslational cleavage of viral polyproteins occurs by a number of different processes involving autocatalysis and different types of viral and cellular proteases. Only some of these processes might be related to protein splicing.

Two independent cases of viral polyproteins, where sequences similar to the intein's N′ splice junction motif (motif A) were

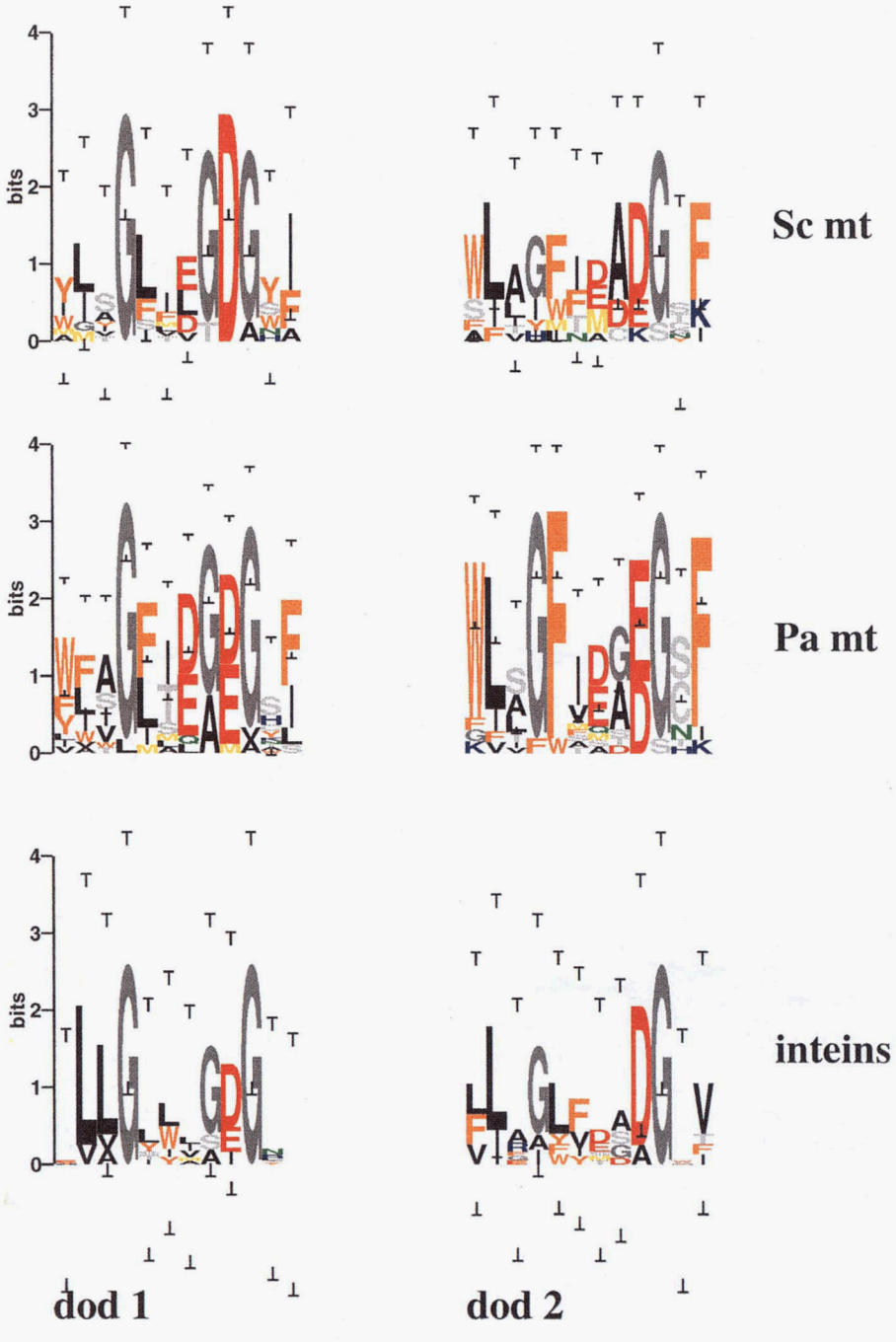

Fig. 5. Dod motifs in inteins and fungi. Sequence logos of dod motifs 1 and 2 in inteins and in the mt genomes of the ascomycetes *Podospora anserina* (Pa mt) and *S. cerevisiae* (Sc mt). Logo features as in Figure 3. The *P. anserina* dod logos were constructed from 18 nonhomologous dod maturase sequences from the mt genome of *P. anserina* (GenBank accession X55026)—LSU r1; COX1 i7a, i7b, i8, i10, i11, i12, i13, i15; COX2 i2; COB i1, i3(B); ND1 4a; ND4 i1; ND4L i1; ND5 i1, i2, i3. The *S. cerevisiae* dod logos were constructed from 10 nonhomologous dod maturase sequences from the mt genome of *S. cerevisiae* (GenBank accession M62622)—LSU rI; COX1 i3, i4, i5a, i5b; COB i2, i3; RF1; RF2; RF3. The intein dod logos constructed as in Figure 3.

adjacent to cleavage sites, are detected. The sequences are among the SwissProt entries with the highest similarity to this intein motif.

Three species of flaviviruses, from the Japanese encephalitis virus group, are found to have a sequence region significantly similar to the N′ splice junction motif three aa from the cleavage point of the major envelope protein E from the nonstructural protein NS1 (Fig. 8A).

The region immediately following the cleavage point between the VP4 and VP2 coat proteins in the human rhinovirus 1A polyprotein is found to be significantly similar to the N′ splice junction motif (Fig. 8B). This cleavage was already noted as being an intein-like processing event (Wallace, 1993).

The discovery of even a few polyprotein cleavage regions similar to the intein's N′ splice junction motif is probably not due to chance. Although the polyprotein sequences are approximately 1.8% of all SwissProt sequences, each contains only a few cleavage sites in sequences typically hundreds to thousands of aa long (E. Kolker, pers. comm.). Furthermore, the positions of the two motif-similar regions identified near the cleavage sites are the same, occurring three or five aa C′ to the cleavage point.

How common are inteins?

Four years after the discovery of the first intein (Hirata et al., 1990; Kane et al., 1990), 10 inteins have been reported. Although

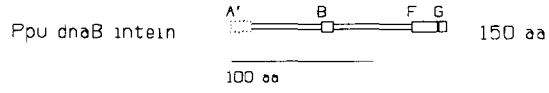

Fig. 6. Intein motifs in the *Porphyra purpurea* cp dnaB putative intein. Schematic sequence diagram as in Figure 4. Dotted box marked with A' indicate similarity to the N' half of motif A. The expectant value for finding the multiple block hits is 3.2×10^{-5} (Henikoff & Henikoff, 1994).

inteins have been found in the three kingdoms of living organisms, they are present in only a few specific genes. One explanation is that the inteins identified so far can move laterally across species more easily than they can move between nonhomologous genes. However, integration into specific sites in homologous genes, i.e., sequence homing, which is common in mobile introns (Dujon, 1989), cannot fully explain the known intein distribution because some of the host genes harbor different inteins at different locations (Perler et al., 1992; Xu et al., 1993; Davis et al., 1994; T. Imanaka, GenBank entry D29671 submission, 1994).

It may very well be that all the presently known inteins are only a distinct subset of a more diverse spectrum of inteins. As shown in this work, all the published intein sequences are of the same general sequence structure. However, the sequence structure of the *P. purpurea* cp dnaB putative intein illustrates that inteins with only part of this sequence structure do exist. Inteins might be found to be more varied and not so rare if only one could identify them better.

Validity of the observed motifs

The validity of the motifs presented in this work is supported by a number of sources. The use of different methods for identifying and assessing the studied motifs reduced the drawbacks of each individual method and enabled the detection of subtle motifs. Although difficult to quantify, the identification of the same motifs by independent methods attests to their authenticity.

Together with the conservation of sequence features in the intein proteins, there is some variability in part of the intein motifs, and not all the motifs were found in all of the inteins. Sequence variability could be the cause of natural polymorphism and could indicate nonconserved positions or be due to the small sample size. The small sample size from which the motifs were derived is also an obstacle in distinguishing genuine conserved regions from chance similarities.

It is important to note that the intein motifs presented in this work are most probably not the ultimate description of the conserved sequence structure of inteins. New intein sequences, better analysis methods and structural data, and other factors are likely to modify the motifs presented in the future.

Conclusion

This work's refinement of known intein motifs and the recognition of additional ones was made possible mainly due to the increasing number of available intein sequences and to new techniques for identifying and utilizing conserved sequence features (Henikoff & Henikoff, 1991; Lawrence et al., 1993; Neuwald & Green, 1994).

The known intein sequences are shown by this work to consist of a number of conserved features in otherwise dissimilar sequences. This implies some structural and/or functional selection for the intein motifs. The presence of similar sequence regions in different inteins is probably due to a common role of these regions rather than simply to a common origin. Inteins are not large proteins, with similar sizes, and at least some are multifunctional, hence it is likely that they all have the same overall structure.

The description of the intein motifs and the inferences about their roles presented in this work can aid the construction of modified and new polypeptides that can undergo protein splicing. Such polypeptides can help us understand the protein-splicing mechanism and may become powerful biotechnological tools.

Methods and sequences

Nomenclature

An alignment of equal length (ungapped) sequence segments is termed a block (Posfai et al., 1989; Smith et al., 1990). A conserved block describes a motif—a pattern occurring and characterizing a group of sequences.

Intein names and protein splicing nomenclature are according to Perler et al. (1994).

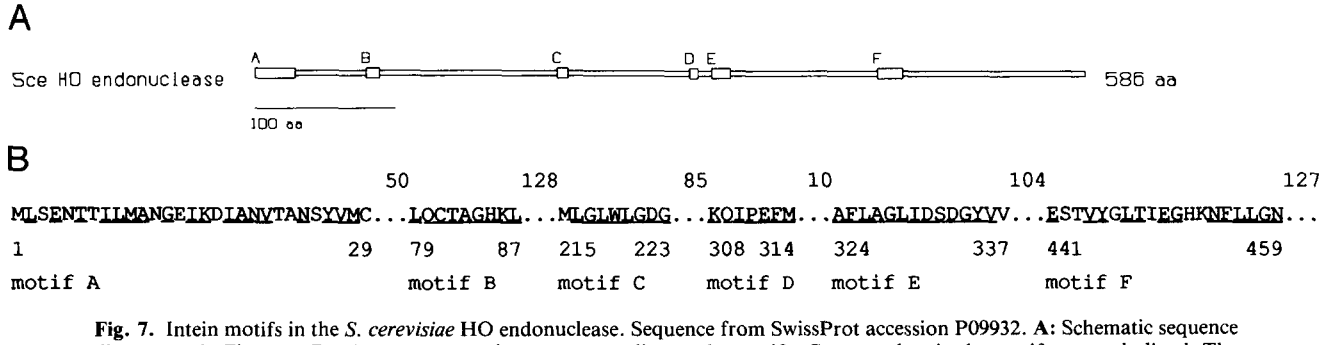

Fig. 7. Intein motifs in the *S. cerevisiae* HO endonuclease. Sequence from SwissProt accession P09932. **A:** Schematic sequence diagram as in Figure 4. **B:** The sequence regions corresponding to the motifs. Conserved aa in the motifs are underlined. The start and end positions of the regions are given below the line and the number of intervening aa and distance between the last region and the end are given above. The expectant value for finding the multiple block hits is 5.5×10^{-20} (Henikoff & Henikoff, 1994).

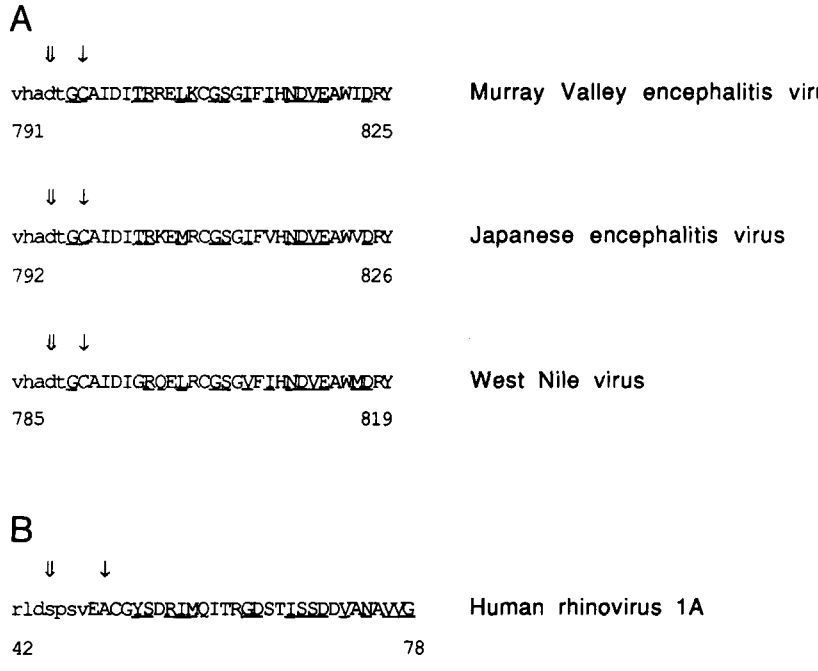

Fig. 8. Intein N′ splice junction motif in polyproteins. The sequence regions corresponding to motif A are in upper case and the preceding segments in lower case. Conserved aa in the motifs are underlined. The start and end positions of the sequences are given below the line. Double arrows mark the polyprotein cleavage points and single arrows mark the position corresponding to the inteins N′ splice junction. **A:** Flaviviruses. The positions of the scores of the regions with motif A in all the SwissProt scores (33,329 sequences) are: top 99.95% for the Murray Valley encephalitis virus; top 99.4% for the Japanese encephalitis virus; and top 98.9% for the West Nile encephalitis virus. SwissProt accessions P05769, P32886, and P06935, respectively. **B:** Human rhinovirus. The position of the motif score in all the SwissProt scores is in the top 99.3%. SwissProt accession P23008.

Sequences

The 10 published intein sequences *Sce VMA*, *Ctr VMA*, *Mtu recA*, *Mle recA*, *Tli pol* 1, *Tli pol* 2, *Psp pol* 1, *Psp pol* 2, *Psp pol* 3,[2] and *Mle pps1* were used in this work. The sequence sources (database accession codes) are *Sce VMA*, SwissProt P17255; *Ctr VMA*, GenBank M64984; *Mtu recA*, SwissProt P26345; *Mle recA*, GenBank X73822; *Tli pol* 1 and 2, SwissProt P30317; *Psp pol* 1, GenBank U00707; *Psp pol* 2 and 3, GenBank D29671; *Mle pps1*, GenBank U00013. The sequence databases used in this work are the SwissProt protein databank release 27 (Bairoch & Boeckmann, 1993) and the GenBank nucleotide database release 81, in finding the pps1 intein, and release 83 for all other searches (Benson et al., 1993).

All the intein sequences are not more than 34% identical, along the whole sequence, to each other except for the *Psp pol* 1, *Psp pol* 3, and *Tli pol* 1. These three inteins are highly similar to each other (59–64% identities along the whole sequences, not shown). Without proper weighting, these sequences would bias the search for conserved regions in the intein sequences and the calculation of their significance (Altschul et al., 1989). Consequently, the *Psp pol* 1 and 3 intein sequence were only included in the blocks calculation. In these calculations, the sequences are given branch-proportional weights (Thompson et al., 1994).

Computer programs

Conserved blocks were detected and evaluated with the ASSET (Neuwald & Green, 1994) and MACAW (Schuler et al., 1991) programs. The Gibbs sampling method (Lawrence et al., 1993) in the MACAW program was used for identifying the sequence blocks. The block limits were chosen to maximize the score using the BLOSUM62 comparison matrix (Henikoff & Henikoff, 1992).

Protein and nucleotide databases were searched with sequence blocks using the blimps program (Wallace & Henikoff, 1992). The blimps search results were assembled and evaluated with the MULTIMAT program (Henikoff, 1992).

Motifs were graphically displayed as sequence logos using the MAKELOGO program (Schneider & Stephens, 1990).

All the programs were obtained from the following computer sites by anonymous FTP: ncbi.nih.gov (ASSET, MACAW, MULTIMAT, and blimps) and ftp.ncifcrf.gov (MAKELOGO).

Motif determination

The MACAW program scores each position in an alignment using an aa similarity matrix (BLOSUM62). The total score of an alignment is the sum of all its position scores. The program calculates the chance probability for the appearance of an alignment score by a statistical formula using an extreme value distribution model of alignment scores (Karlin & Altschul, 1990; Altschul et al., 1994). The resulting p-value is independent from the ones calculated by the ASSET and Gibbs methods. These two methods were used to detect the inteins' motifs. They each employ different principles in their identification and evaluation of motifs (Neuwald & Green, 1994).

An alignment was considered a motif only if it was found significant (p-value $< 1 \times 10^{-2}$) and had the same relative position to other identified motifs in all the sequences. To avoid detecting regions shared by just a subgroup of the inteins, only motifs appearing in the majority of the sequences were considered. The motif boundaries were taken to maximize the motifs score (minimizing the p-value).

[2] The *Pyrococcus species* strain GB-D DNA polymerase intein was renamed *Psp pol* intein 1 and the two additional inteins discovered in *P. species* strain KOD1 DNA polymerase (T. Imanaka, GenBank entry D29671 submission, 1994) were named *Psp pol* intein 2 and *Psp pol* intein 3, according to the published nomenclature conventions (Perler et al., 1994).

Acknowledgments

I thank the Weizmann Institute's Biological Computing Division, BioInformatics Unit, and the Computing Center's UNIX group for help and advice; Steve Henikoff, Jorja Henikoff, William Alford, and Tom Schneider for providing and helping install and operate their software programs; Mike Reith for sharing sequence data prior to publication; Ed Trifonov for many insightful discussions; and Jacqui Beckmann, Philipp Bucher, Alex Girshovich, Eugene Kolker, Steve Henikoff, Jorja Henikoff, Amnon Horovitz, Eitan Rubin, and Edward Trifonov for critical reading of the manuscript. This work was supported by the Weizmann Institute Human Genome committee.

References

Altschul SF, Boguski MS, Gish W, Wootton JC. 1994. Issues in searching molecular sequence databases. *Nature Genet 6*:119–129.

Altschul SF, Carroll RJ, Lipman DJ. 1989. Weights for data related by a tree. *J Mol Biol 207*:647–653.

Anraku Y, Hirata R. 1994. Protozyme: Emerging evidence in nature. *J Biol Chem 115*:175–178.

Arnold E, Luo M, Vriend G, Rossmann MG, Palmenberg AC, Parks GD, Nicklin MJ, Wimmer E. 1987. Implications of the picornavirus capsid structure for polyprotein processing. *Proc Natl Acad Sci USA 84*:21–25.

Bairoch A, Boeckmann B. 1993. The SWISS-PROT protein sequence data bank: Recent developments. *Nucleic Acids Res 21*:3093–3096.

Benson D, Lipman DJ, Ostell J. 1993. GenBank. *Nucleic Acids Res 21*:2963–2965.

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*:2575–2583.

Cooper AA, Stevens TH. 1993. Protein splicing: Excision of intervening sequences at the protein level. *BioEssays 15*:667–674.

Davis EO, Jenner PJ, Brooks PC, Colston MJ, Sedgwick SG. 1992. Protein splicing in the maturation of *M. tuberculosis* recA protein: A mechanism for tolerating a novel class of intervening sequence. *Cell 71*:201–210.

Davis EO, Sedgwick SG, Colston MJ. 1991. Novel structure of the recA locus of *Mycobacterium tuberculosis* implies processing of gene product. *J Bacteriol 173*:5653–5662.

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.

Doolittle RF. 1993. The comings and goings of homing endonucleases and mobile introns. *Proc Natl Acad Sci USA 90*:5379–5381.

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

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

Gu HH, Xu J, Gallagher M, Dean GE. 1993. Peptide splicing in the vacuolar ATPase subunit A from *Candida tropicalis*. *J Biol Chem 268*:7372–7381.

Henikoff S. 1992. Detection of *Caenorhabditis* transposon homologs in diverse organisms. *New Biologist 4*:382–388.

Henikoff S, Henikoff JG. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res 19*:6565–6572.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA 89*:10915–10919.

Henikoff S, Henikoff JG. 1994. Protein family classification based on searching a database of blocks. *Genomics 19*:97–107.

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

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.

Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebl M, Stevens TH. 1990. Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H^+-adenosine triphosphatase. *Science 250*:651–657.

Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA 87*:2264–2268.

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

Kostrzewa M, Zetsche K. 1992. Large ATP synthase operon of the red alga *Antithamnion* spec. resembles the corresponding operon in Cyanobacteria. *J Mol Biol 227*:961–970.

Kostrzewa M, Zetsche K. 1993. Organization of plastid-encoded ATPase genes and flanking regions including homologues of infB and tsf in the thermophilic red alga *Galdieria sulphuraria*. *Plant Mol Biol 23*:67–76.

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

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science 262*:208–214.

Michel F, Jacquier A, Dujon B. 1982. Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie 64*:867–881.

Neff NF. 1993. Protein splicing: Selfish genes invade cellular proteins. *Curr Opin Cell Biol 5*:971–976.

Neuwald AF, Green P. 1994. Detecting patterns in protein sequences. *J Mol Biol 239*:698–712.

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, Davis EO, Dean GE, Gimble FS, Jack WE, Neff N, Noren JN, Thorner J, Belfort M. 1994. Protein splicing elements: Inteins and exteins—A definition of terms and recommended nomenclature. *Nucleic Acids Res 22*:1125–1127.

Posfai J, Bhagwat AS, Posfai G, Roberts RJ. 1989. Predictive motifs derived from cytosine methyltransferases. *Nucleic Acids Res 17*:2421–2435.

Russell DW, Jensen R, Zoller MJ, Burke J, Errede B, Smith M, Herskowitz I. 1986. Structure of the *Saccharomyces cerevisiae* HO gene and analysis of its upstream regulatory region. *Mol Cell Biol 6*:4281–4294.

Schneider TD, Stephens RM. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res 18*:6097–6100.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol 188*:415–431.

Schuler GD, Altschul SF, Lipman DJ. 1991. A workbench for multiple alignment construction and analysis. *Proteins Struct Funct Genet 9*:180–190.

Shub DA, Goodrich-Blair H. 1992. Protein introns: A new home for endonucleases. *Cell 71*:183–186.

Smith HO, Annau TM, Chandrasegaran S. 1990. Finding sequence motifs in groups of functionally related proteins. *Proc Natl Acad Sci USA 87*:826–830.

Strathern JN, Klar AJ, Hicks JB, Abraham JA, Ivy JM, Nasmyth KA, McGill C. 1982. Homothallic switching of yeast mating type cassettes is initiated by a double-stranded cut in the MAT locus. *Cell 31*:183–192.

Thompson JD, Higgins DG, Gibson TJ. 1994. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *CABIOS 10*:19–29.

Thony-Meyer L, Bock A, Hennecke H. 1992. Prokaryotic polyprotein precursors. *FEBS Lett 307*:62–65.

Wallace CJA. 1993. The curious case of protein splicing: Mechanistic insights suggested by protein semisynthesis. *Protein Sci 2*:697–705.

Wallace JC, Henikoff S. 1992. PATMAT: A searching and extraction program for sequence, pattern and block queries and databases. *CABIOS 8*:249–254.

Waring RB, Davies RW, Scazzocchio C, Brown TA. 1982. Internal structure of a mitochondrial intron of *Aspergillus nidulans*. *Proc Natl Acad Sci USA 79*:6332–6336.

Williamson DH, Gardner MJ, Preiser P, Moore DJ, Rangachari K, Wilson RJM. 1994. The evolutionary origin of the 35 kb circular DNA of *Plasmodium falciparum*: New evidence supports a possible rhodophyte ancestry. *Mol Gen Genet 243*:249–252.

Xu MQ, Southworth MW, Mersha FB, Hornstra LJ, Perler FB. 1993. In vitro protein splicing of purified precursor and the identification of a branched intermediate. *Cell 75*:1371–1377.