

# INTEINS: Structure, Function, and Evolution

---

J. Peter Gogarten,<sup>1</sup> Alireza G. Senejani, Olga Zhaxybayeva,  
Lorraine Olendzenski, and Elena Hilario<sup>2</sup>

*Department of Molecular and Cell Biology, University of Connecticut, 75 North  
Eagleville Road, Storrs, Connecticut 06269-3044; e-mail: gogarten@uconn.edu;  
ali@carrot.mcb.uconn.edu; olga@carrot.mcb.uconn.edu;  
lorraine@carrot.mcb.uconn.edu; elena@carrot.mcb.uconn.edu*

**Key Words** homing endonuclease, splicing, protein introns, selfish genes, parasitic genes

■ **Abstract** Inteins are genetic elements that disrupt the coding sequence of genes. However, in contrast to introns, inteins are transcribed and translated together with their host protein. Inteins appear most frequently in Archaea, but they are found in organisms belonging to all three domains of life and in viral and phage proteins. Most inteins consist of two domains: One is involved in autocatalytic splicing, and the other is an endonuclease that is important in the spread of inteins. This review focuses on the evolution and technical application of inteins and only briefly summarizes recent advances in the study of the catalytic activities and structures of inteins. In particular, this review considers inteins as selfish or parasitic genetic elements, a point of view that explains many otherwise puzzling aspects of inteins.

## CONTENTS

INTRODUCTION .....	264
Definitions and History .....	264
Nomenclature .....	264
INTEIN DISTRIBUTION, TYPES, AND STRUCTURE .....	265
Distribution of Inteins .....	265
Position of Inteins within Host Proteins .....	266
Large and Mini-Inteins .....	267
Split Inteins .....	267
Conserved Motifs .....	269
Comparisons of Three-Dimensional Structures .....	270
MECHANISM OF PROTEIN SPLICING .....	270
INTEINS AS PARASITIC GENES .....	270
Endonucleases and Homing .....	270

---

<sup>1</sup>corresponding author.

<sup>2</sup>current address: HortResearch, 120 Mt. Albert Road, Private Bag 92 169, Mt. Albert, Auckland, New Zealand.

Selfish and Parasitic Genes .....	272
The Cyclic Reinvasion Model for Endonuclease Maintenance .....	272
How Did Inteins Originate? .....	274
Multiple Origins? .....	276
Breaking the Homing Cycle: Acquisition of Nonparasitic Functions .....	277
Why Are Inteins Located Where They Are? .....	279
APPLICATION OF INTEINS IN BIOTECHNOLOGY .....	280
CONCLUDING REMARKS .....	282

## INTRODUCTION

### Definitions and History

Inteins (internal proteins) are genetic elements similar to self-splicing introns; however, inteins are transcribed and translated together with their host protein. Only at the protein level do the inteins excise themselves from the host protein. The two portions of the host protein separated by the intein are called exteins (external proteins) (16, 25, 69). During the splicing process the intein is excised, the two exteins are joined by a peptide bond, and the host protein assumes its normal folding and function. The first intein was discovered in 1987 when the carrot and *Neurospora crassa* vacuolar ATPases were compared with a putative Ca<sup>2+</sup>-pumping ATPase. The latter had been isolated as a gene whose mutation made yeast resistant against the calmodulin antagonist trifluoperazine (91). The beginning and end of the encoded protein was very similar to the vacuolar ATPase subunits whose sequences had been submitted to the databanks at the same time. However, the central region of the putative calcium pump had no similarity to any known ATPase. Rather, this portion showed weak similarity to endonucleases. Anraku's lab (41) isolated the cDNA for the yeast vacuolar ATPase A-subunit and found the same sequence, including the central region, that had been earlier described as the trifluoperazine resistance gene (91). Surprisingly, denaturing polyacrylamide gel electrophoresis of the isolated protein demonstrated that the catalytic subunit of the functioning yeast V-ATPase had a molecular weight of only 70 kDa, as expected for a subunit without the insertion. Subsequently Kane et al. (48) showed that the insertion was still present in the mRNA, that the whole protein including the insertion was translated, and that the insertion spliced itself out of the protein during posttranslational processing.

### Nomenclature

Inteins are named after the organism and host protein in which they reside (68, 69). If there is more than one intein in the host protein, the different inteins are designated by numbers. For example, *Tfu* Pol-2 denotes the second intein in the DNA polymerase of *Thermococcus fumicolans*. Inteins that occupy a homologous site in the host protein in a different organism are called intein alleles. For example,

the third intein in the *Thermococcus aggregans* DNA polymerase, *Tag* Pol-3, is located in a position corresponding to the *Tfu* Pol-2 insertion site. These two inteins therefore are considered alleles (83).

Many inteins contain an endonuclease domain, and thus also receive a name following the conventions for naming homing endonucleases (2). The name of these endonucleases begins with PI (for protein insert) to denote them as an intein, followed by a three letter species indicator and a roman numeral specifying the different PI endonucleases present in an organism. Other prefixes used for endonucleases are I for intron and F for freestanding (2). For example, PI *MgaI* denotes the endonuclease activity of the intein in an ABC transporter from *Mycobacterium gastri* (*Mga* Pps1) (85); PI-*SceI* denotes endonuclease activity of the intein in the yeast vacuolar ATPase catalytic subunit, *Sce* VMA1; and I-*SceI* denotes the endonuclease in the 23S mitochondrial rRNA intron.

## INTEIN DISTRIBUTION, TYPES, AND STRUCTURE

### Distribution of Inteins

The intein database, InBase (68), provides information on all described inteins. Among other data it lists the inteins' sequences, conserved motifs, host organisms, and host proteins. More than 130 inteins are known in 34 different types of proteins (68, 76, 77). The inteins are between 134 and 608 amino acids long, and they are found in members of all three domains of life: Eukaryotes, Bacteria, and Archaea (Table 1) (Figure 1). Inteins are found in proteins with diverse functions, including metabolic enzymes, DNA and RNA polymerases, proteases, ribonucleotide reductases, and the vacuolar-type ATPase. However, enzymes involved in DNA replication and repair appear to dominate (55).

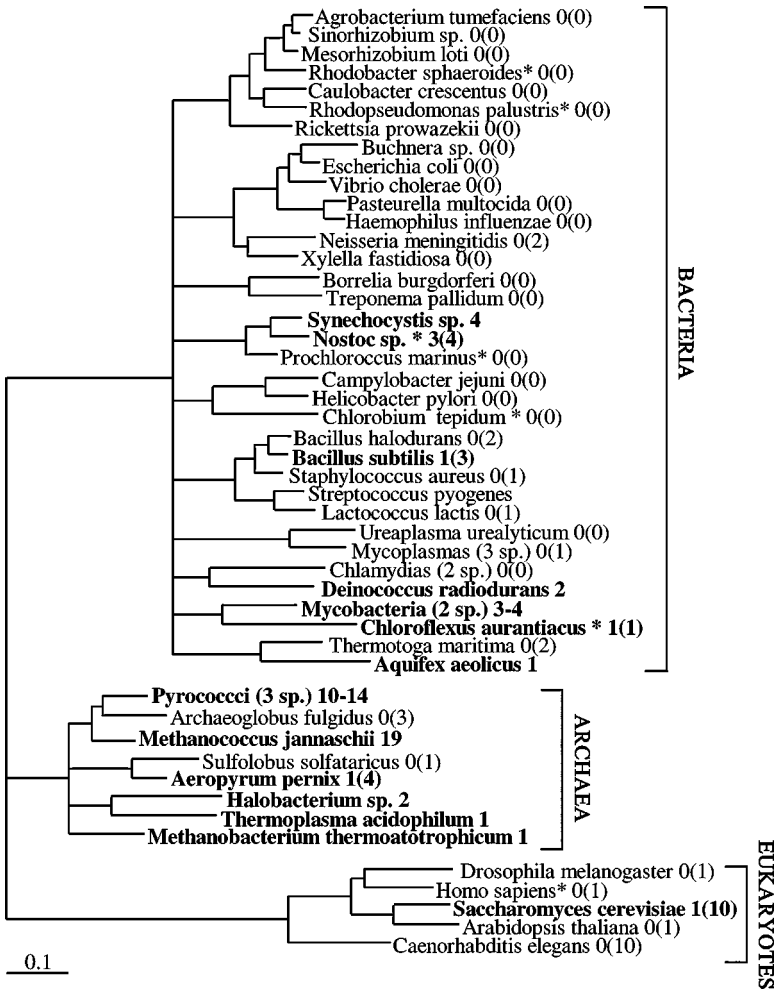
The ratio of an intein's size to that of its host protein varies widely. For instance, the size of the three inteins found in the *Methanococcus jannaschii* replication factor C (*Mja* RFC 1-3) are more than four times the size of their host protein, whereas the *Filobasidiella neoformans* pre-mRNA splicing factor (*Fne* pRP8) intein is less than one tenth of its host protein size (68).

**TABLE 1** Number of inteins reported for the three domains of life (68)

	Number of species	Number of inteins
Eukaryotes	7	7
Eubacteria	25	44
Archaea	16	79
Total	48	130

### Position of Inteins within Host Proteins

Within the host protein, inteins appear to prefer conserved regions, for example, nucleotide-binding domains (76). Figure 2 analyzes three families of conserved host proteins with respect to intein and intron insertion points. The ATPase catalytic subunits have two intein insertion sites: The vacuolar type ATPases of *Saccharomyces cerevisiae* and *Candida tropicalis* have an intein in location “a” (77); the archaeal ATPases of *Pyrococcus abyssi*, *P. furiosus*, *P. horikoshii*, *Thermoplasma acidophilum*, and *T. volcanium* have an intein in location “b” (Figure 2a). Replication factor C, which is less than 300 amino acids long, accommodates inteins in three different sites (Figure 2b). Each intein is twice the length of its host. CDC21 (cell division control protein 21) harbors six inteins: three in location “a” and three



in location “b.” All the inteins are inserted in the most conserved parts of the host protein. In contrast, the ATPase and replication factor C intron insertion sites do not appear to be restricted to conserved parts of the host protein (Figure 2*a,b*).

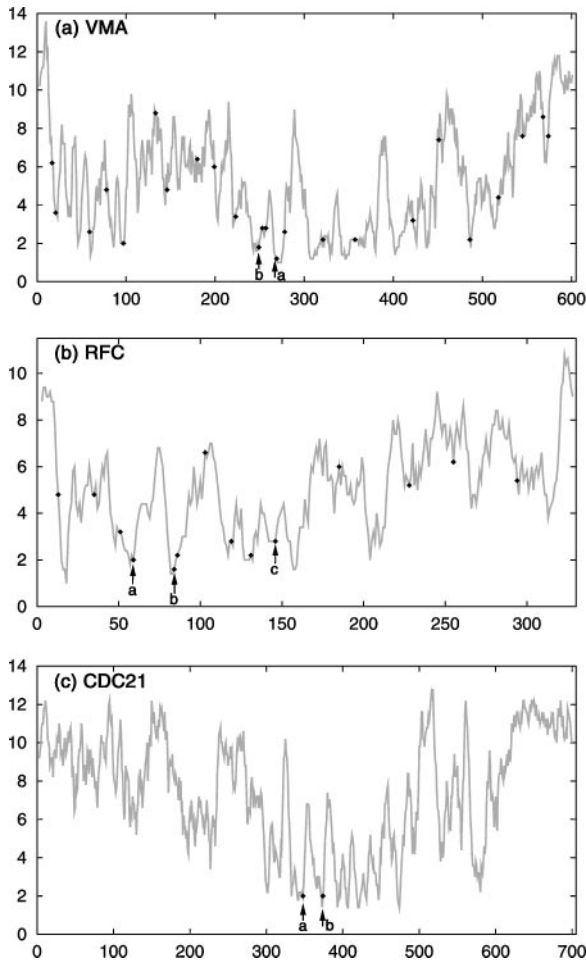
## Large and Mini-Inteins

Most reported inteins consist of two domains (55): a self-splicing domain and an endonuclease domain. These are known as large inteins. The N- and C-terminal regions of the large inteins contain the elements necessary for splicing. Some inteins, known as mini-inteins, consist of only the self-splicing domain. Deletion of the endonuclease domain from a large intein does not affect protein splicing (13, 22, 92). The endonuclease found in large inteins is thought to play a crucial role in the spread of inteins (see below). Of the 4 subfamilies of homing endonucleases, the LAGLIDADG family is the largest, with more than 150 members reported to date (17, 99). With two exceptions (see below), all endonucleases found in inteins belong to the LAGLIDADG family.

## Split Inteins

Two split inteins capable of protein trans-splicing were identified in DnaE, the catalytic subunit alpha of DNA polymerase III, in the cyanobacteria *Synechocystis* sp. strain PCC6803 and *Nostoc punctiforme* (28, 68, 104). The *Ssp* DnaE intein

**Figure 1** Distribution of inteins across the three domains of life. Only organisms whose genome sequences are complete or nearly complete are included. Organisms that harbor inteins are highlighted in bold. The distribution of inteins is often depicted showing all organisms that contain inteins, giving the impression that inteins are widely distributed. However, among those Bacteria and Eukaryotes whose genomes have been sequenced, only a minority harbors inteins. Genomes were screened for the presence of inteins using PSI-BLAST (1, 86) profiles of 114 known inteins (68). Genomes marked with \* are in progress and have not yet been annotated. Profiles were calculated using five iterations of PSI-BLAST, the known inteins as queries and the NCBI's nonredundant database as a target. Each of the 114 profiles was used to search the indicated genomes. The numbers of PSI-BLAST hits with E-value lower than  $10^{-5}$  are given in parentheses after the organism's name. If no number is given in parentheses, then the number reflects that given in InBase for this organism (68). The hits in the individual genomes were carefully examined for the presence or absence of inteins. Many of the hits were homing endonucleases of introns and self-splicing regulatory proteins. The number of identified inteins is given without parentheses. Multiple species of the same genus are represented by a single branch with number of species examined indicated. For those genera the number of inteins and the number of PSI-BLAST hits is given per species. The tree was calculated in TREE-PUZZLE 5.0 (95) using the HKY85 substitution model and an alignment of small subunit rRNA genes from the European ribosomal RNA database (100).



**Figure 2** Positions of inteins and introns along the coding sequence of the host protein. The graph represents the conservation of the sites in protein alignments of homologs of subunit A of the vacuolar/archaeal ATPase (*panel a*), replication factor C (*panel b*), and cell division control protein 21 (*panel c*). The abscissa indicates the amino acid position along the alignment, and the ordinate gives the number of different amino acids present in that position averaged over a window of size 5. Sites in the alignment where more than 50% of the sequences had a gap were excluded from the analysis; all other gaps were treated as twenty-first amino acid. The positions of inteins are indicated as *dots with arrows*. Positions of introns from *Drosophila melanogaster*, *Arabidopsis thaliana*, and *Schizosaccharomyces pombe* are indicated as *dots without arrows*. Inteins are situated in the most conserved parts of the proteins, whereas this does not hold true for intron locations in general.

was identified as a naturally occurring split mini-intein capable of protein trans-splicing. The intein and DnaE are encoded by two separate genes, *dnaE-n* and *dnaE-c*, located more than 700 kbp apart in the genome and on opposite DNA strands (104). *dnaE-n* encodes the amino-terminal part of the DNA polymerase subunit and the amino-terminal part of the intein; *dnaE-c* encodes the carboxy-terminal parts of the intein and of the DNA polymerase subunit. Splicing and cleavage activity of the split intein result in the functional DNA polymerase and have been described in detail in (57).

## Conserved Motifs

Comparative analysis of intein sequences reveals conserved motifs that can be used to identify and characterize inteins (68, 70, 72). Four conserved motifs, blocks A, B, F, and G, are found in all known inteins (Table 2). Blocks A and B are present at the intein N terminus, and blocks F and G are present at the C-terminal end of the intein. Pietrokovski (75) characterized two additional motifs (N2 and N4) that are located close to the N-terminal (Table 2). Inteins with an endonuclease domain have another four conserved motifs (blocks C, D, E, and H) (70). Studies using site-directed mutagenesis and comparative sequence and structure analyses (18, 22, 49, 70, 75) indicate that the N- and C-terminal motifs (blocks A, N2, B, N4, F, and G) are involved in protein splicing, whereas the endonuclease activity involves the central blocks C, D, E, and H (Table 2). The amino acid of the extein following the intein insertion site and block A at the N terminus of the intein contain residues chemically essential for splicing (15, 70, 105). Blocks C and E are the dodecapeptide motifs required for endonuclease activity (32, 42).

**TABLE 2** Comparison of names used for conserved intein motifs. The first column gives the abbreviation used in the older literature (68, 70, 72); the second column gives the names suggested in (75)

Perler et al. (69)	Pietrokovski (75)	Other names
A	N1	N-terminal splicing motif
—	N2	—
B	N3	—
—	N4	—
C	EN1	DOD, LAGLIDADG motif
D	EN2	—
E	EN3	DOD, LAGLIDADG motif
H	EN4	—
—	HNH	HNH endonuclease motif
F	C2	—
G	C1	C-terminal splicing motif

## Comparisons of Three-Dimensional Structures

The structure of the single, well-characterized intein in *S. cerevisiae* (*Sce* VMA1 or PI *Sce*I) (24, 34, 40, 43, 78) was determined by X-ray crystallography (24). The two-domain structure of PI *Sce*I is clearly visible (Figure 3). The self-splicing domain is similar in structure to the mini-intein in the *Mycobacterium xenopi* gyrase (*Mxe* GyrA) (51). As expected, PI-*Sce*I makes target sequence-specific contacts using residues from the endonuclease domain (93). However, a part of the other domain that is distant from the PI-*Sce*I cleavage site also contributes to the recognition of the target sequence (24, 43). These additional interactions were determined using photo-crosslinking and affinity cleavage (43).

Figure 3 compares these two intein structures to the structure of the autoprocessing domain of the hedgehog protein from *Drosophila melanogaster* (39). Hedgehog proteins undergo autocatalytic cleavage and esterification, and they play an important role in the development of multicellular animals (26, 54, 80). The splicing and autoprocessing domains are all-beta structures and contain two homologous subdomains related by a pseudo two-fold axis of symmetry [see (39) and below for a discussion of evolutionary implications].

## MECHANISM OF PROTEIN SPLICING

Substantial information about the chemical reactions involved in protein splicing is available (12, 15, 58, 66, 71, 79, 90, 94, 105) and was recently reviewed in (66). Briefly, protein splicing involves the following four steps (Figure 4):

**Step 1** The amino-terminal splice junction of the intein is activated by an N-O or N-S shift that leads to an ester or thioester intermediate. As a result of this rearrangement, the N-extein binds to the oxygen of a serine or to the sulfur of a cysteine residue at the amino-terminal splice junction.

**Step 2** Cleavage of the ester at the amino-terminal splice junction occurs through attack of a nucleophilic residue located at the carboxy-terminal splice junction. This transesterification results in a branched protein intermediate.

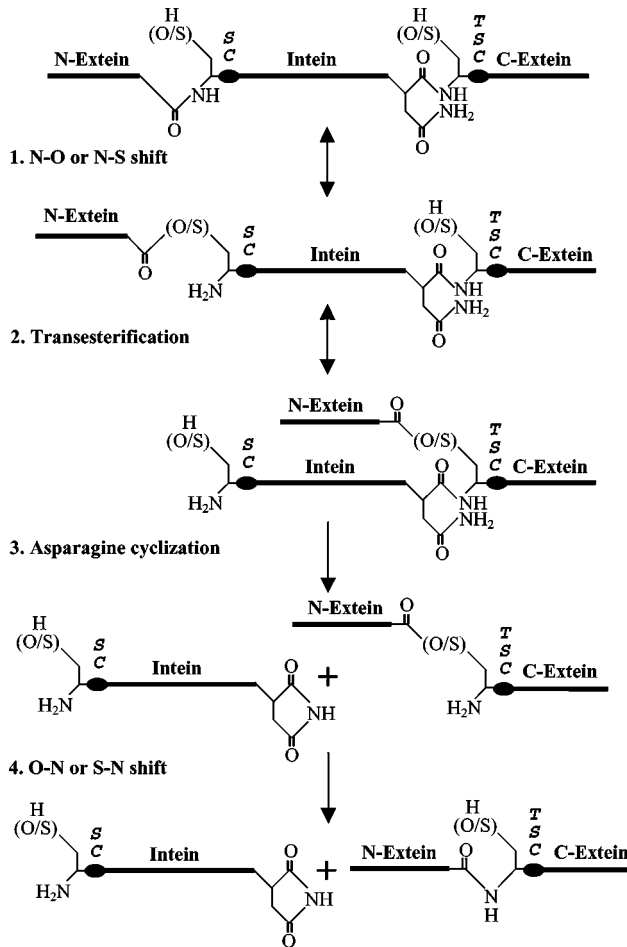
**Step 3** The cleavage proceeds through asparagine cyclization, which causes intein excision and splicing of the two exteins by an ester bond. Several inteins have glutamine rather than asparagine residues at their C-terminal end, suggesting that cleavage in these inteins might occur via an aminoglutarimide rather than an aminosuccinimide intermediate (66, 74).

**Step 4** A spontaneous rearrangement results in formation of a peptide bond between the two exteins.

## INTEINS AS PARASITIC GENES

### Endonucleases and Homing

Homing is the transfer of a parasitic genetic element to a cognate allele that lacks that element (9, 30, 33, 47). The result of homing is the duplication of the parasitic



**Figure 4** Splicing mechanism of inteins. Inteins splicing takes place in four reaction steps (for further discussion see text).

genetic element (47). Homing allows for super-Mendelian inheritance and guarantees the rapid spread of the genetic element in a population. Homing was first described for introns that harbor a homing endonuclease, but inteins appear to spread by the same mechanism (21, 30). When an allele that contains an intron or intein with a functional homing endonuclease coexists in the same cell as a gene that contains an allele without the parasitic element, homing can be initiated by the endonuclease, and the intein/intron-free allele can be converted into an intein/intron-containing allele.

The structure and function of homing endonucleases were recently reviewed (2, 47). Briefly, the homing endonucleases recognize sites of 14–40 residues and usually do not require a complete match with the target sequence (9, 47). The

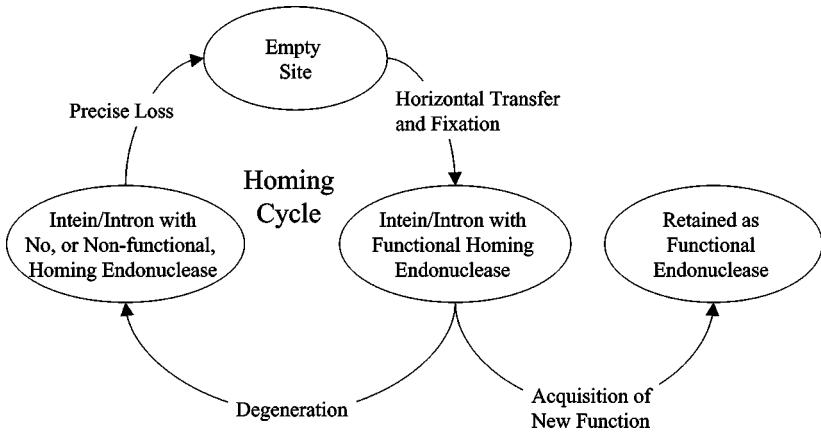
Scv VMA-1 intein recognizes a site of 31 bp (34, 40) residues, but only 9 of these residues are essential. The recognition site is so large that the homing endonuclease often cleaves a genome in a single site only. During the repair of the double-strand break, the gene encoding the homing endonuclease and surrounding sequences are copied into the cleavage site. This copying can be the result of legitimate or illegitimate recombination. The former case results in homing: The parasitic genetic element is copied into an intein/intron-free version of the allele. The latter might result in a new parasitic element if the endonuclease integrates into an existing intron or intein, or an existing parasitic element might be copied into a new target gene (2).

## Selfish and Parasitic Genes

Many aspects of inteins, in particular their evolution, properties, and distribution, are better understood if one views them as parasitic genetic elements. There is an ongoing debate about which genes should be labeled as selfish [e.g., (7)]. Dawkins considers all genes as selfish (20). His gene-centered view considers the organisms' cells and bodies as vessels constructed by the genes to ensure the gene's future existence and their propagation into the next generation. Although all genes are selfish, in most instances the gene's interests coincide with the interest of the organism: A gene increases its survival chances by increasing the fitness of the organism. In contrast, the genetic elements that utilize homing are selfish in an egoistic sense. The proliferation of a parasitic genetic element via homing will continue even if the presence of this element does not contribute to the fitness of the organisms carrying the affected allele. To clearly separate these egoistic genes from Dawkins' more benign selfishness (20), we denote the former as parasitic genetic elements.

## The Cyclic Reinvasion Model for Endonuclease Maintenance

Homing will lead to the rapid propagation of the genetic element that employs homing, and ultimately the allele that contains the element will be fixed in the population. At this point, however, there will be no more selection for functioning endonuclease activity. The endonuclease is only under selection for function during the super-Mendelian spreading phase; once fixed in a population, the endonuclease is expected to decay owing to random genetic drift or the deletion bias of the organism (59). Following this reasoning Goddard & Burt (35) studied the distribution and sequence of a self-splicing intron in the mitochondrial rRNA gene of different yeasts. They found that the intron frequently jumped between different yeast species and concluded that there is a cycle beginning with (a) the invasion of the empty site by an endonuclease-containing parasitic genetic element through horizontal transfer, homing, and fixation of the invaded allele in the population, followed first by (b) degeneration and loss of the endonuclease open reading frame and then by (c) loss of the parasitic genetic element through precise deletion, followed again by (a). The resulting cycle is depicted on the left side of Figure 5. This



**Figure 5** Homing cycle of parasitic genetic elements [modified from (35)].

homing cycle provides an explanation of how a functional homing endonuclease activity can be maintained over long periods of time through selection. Without re-invasion of a population, i.e., without the cycle operating, the parasitic genetic element would lose its endonuclease activity after the allele became fixed in the population. Not surprisingly, many inteins do not have a functioning endonuclease domain [see Figure 6 and (55, 68, 70, 75, 76)].

**HOMING, SEX, AND HORIZONTAL TRANSFER** In eukaryotes the intein-free and intein-containing alleles can be brought together through sex. Gimble & Thorner (33) reintroduced an intein-free version of the *vma-1* gene into *S. cerevisiae*. After meiosis all *vma-1* alleles contained the intein. Horizontal gene transfer is an important process that brings intein/intron-free copies together with those that contain the parasitic genetic element. Many inteins and introns with homing endonuclease were discovered in organellar genomes. For example, the Clp protease in *Chlamydomonas eugametos* chloroplasts and the DNA helicase B in red algal and chrytophyte plastids (23a, 73, 81) contain inteins; one of the most thorough studies on the population dynamics of an intron with homing endonuclease was performed on a group I intron in the large subunit RNA in yeast mitochondria (35). At least in the latter case the data clearly indicate frequent re-invasion events. It is remarkable that these parasitic genetic elements are transferred with a frequency sufficient to survive inside organellar DNA, which is usually considered to avoid recombination and therefore thought to be subject to Muller’s ratchet, i.e., the accumulation of slightly deleterious mutations in asexual populations (3, 82).

**ENDONUCLEASE LOSS AND HORIZONTAL TRANSFER** In the case of inteins there is insufficient data to estimate the relative times for each step of the homing cycle (Figure 5). For the intron in the large mitochondrial ribosomal subunit Goddard &

Burt (35) estimated the timescale of the homing cycle including the horizontal transmission to be  $10^6$ – $10^7$  years, and perhaps much faster. Cho et al. (10) estimated that the intron in the plant mitochondrial *coxI* genes was transferred between species over 1000 times during angiosperm evolution. The precise loss of an intein occurs less frequently than precise loss of introns. Similar endonuclease-free inteins are found in related species, suggesting the possibility that these inteins might be exclusively maintained by vertical inheritance over long periods of time (e.g., the small inteins in *vma-1* in the genus *Thermoplasma*) (Figure 6). However, transfer between divergent organisms is evident for several inteins with an endonuclease domain. The best-documented instance to date is the intein in the *dnaB* gene in *Rhodothermus marinus*. This gene is similar to the intein in *Synechocystis* sp. (56) and *N. punctiforme*. In *R. marinus* the intein has a different codon usage than the surrounding gene, and the intein sequences show higher similarity between *Synechocystis* sp. and *R. marinus* than the exteins. Taken together these findings indicate a recent invasion of the *dnaB* gene in *R. marinus* (56).

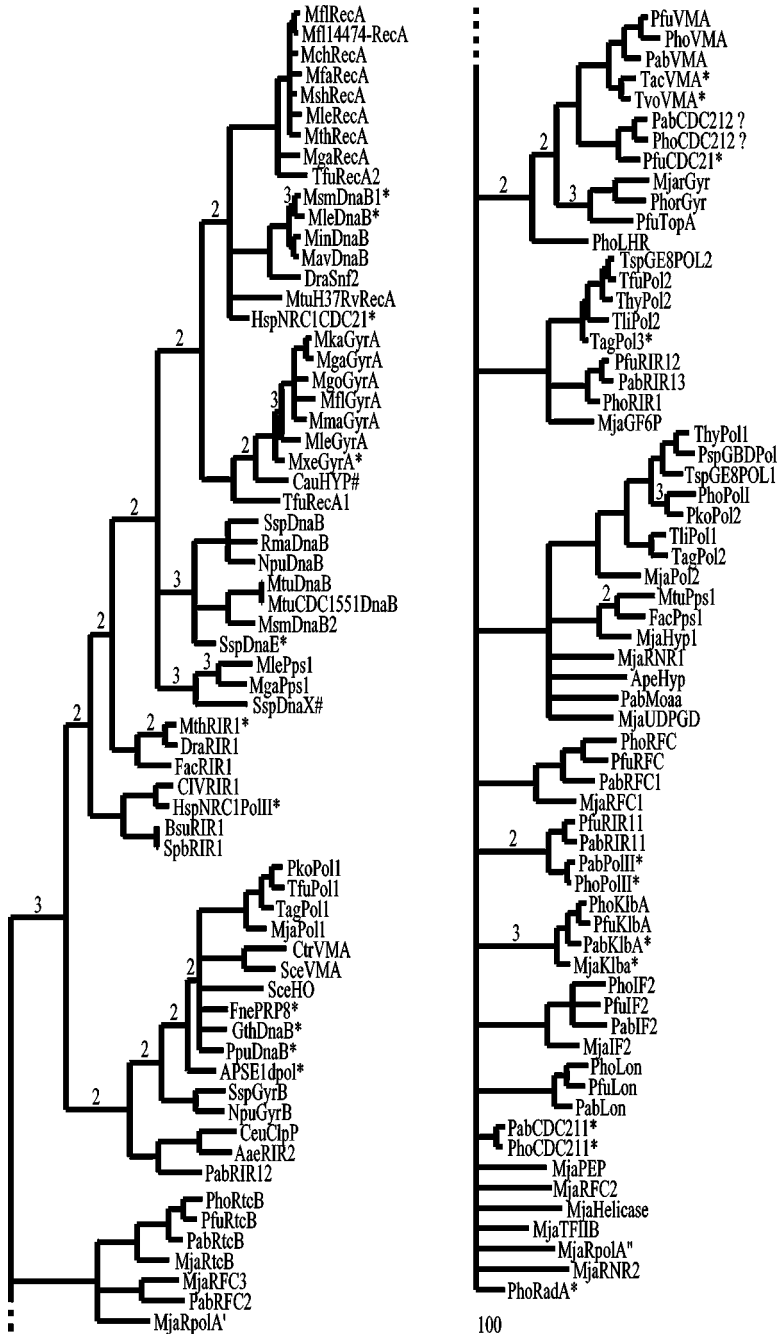
Many other instances of horizontal transfer are suggested by the sporadic distribution of intein alleles among divergent organisms. Similar inteins often occupy homologous sites in related proteins, but the exteins are not each other's closest relatives (56, 89). For example, the *Pyrococcus furiosus* DNA topoisomerase I and the *M. jannaschii* reverse gyrase exteins are structurally related and harbor an intein in corresponding sites, but the two enzymes have different functions (14). Intein homing has been observed under laboratory conditions (33); in contrast, the transfer of an intein to a species that did not previously harbor this intein or the invasion of a new, previously intein-free gene can at present only be inferred from comparative phylogenetic analyses (30). However, given that inteins are parasitic genetic elements, the alternative assumption of exclusive vertical inheritance with multiple parallel and convergent intein losses is untenable.

## How Did Inteins Originate?

The splicing domain of inteins is similar to the autoprocessing domains of regulatory proteins that undergo autocatalytic excision [see (52) and Figure 3]. It is not yet clear which came first: inteins as parasitic genetic elements (70) or the autocatalytic splicing domain in regulatory proteins (76). Liu (55) described a scenario

---

**Figure 6** Phylogenetic tree of 128 known inteins and *Sce* HO. Amino acid sequences were kindly provided by F.B. Perler, curator of InBase (68), and aligned using SAM (44). The tree was calculated using parsimony as implemented in PAUP\* v. 4.0beta8 (97). Numbers denote branches with Bremer decay indices (6) smaller than 4. Unlabeled branches do not decay even if three additional steps are allowed. The notation of names is as in InBase (68). “\*” indicates inteins without endonuclease domain; “#” denotes inteins in which a DOD family endonuclease is present in a different reading frame; and “?” specifies those inteins for which the presence of an endonuclease domain is questionable.



in which a domain with self-cleaving activity undergoes a duplication resulting in a protein that autocatalytically cleaves itself into three separate peptides. A loop exchange is postulated to have integrated the two self-cleaving units, producing an intein that splices rather than cleaves the protein. The homing endonucleases were likely to invade such an element at a later time. The latter hypothesis is supported by finding different types of endonucleases in inteins (17, 70, 75) and by finding the intein-typical LAGLIDADG endonucleases also in introns and as open reading frames without clear association with parasitic elements (2). The idea of homing endonucleases invading self-splicing introns and inteins after these elements originated is also supported by the following reasoning (21): If endonucleases are mobile in genomes, genetic elements that remove themselves from the gene product would constitute preferred integration sites because in these locations the functions encoded by the surrounding DNA would not be disrupted.

## Multiple Origins?

Inteins can be readily identified using position-specific iterated BLAST (PSI BLAST) or profile hidden Markov model (HMM) searches (1, 17, 72) (Figure 1). Protein space is so huge that it appears unlikely that protein folds and domains recognized by these approaches originated through convergent evolution (50); however, this possibility, albeit unlikely, cannot yet be ruled out. Both intein domains, the splicing and the endonuclease domain, probably had unique and independent origins. However, an intein that functions as an effective parasitic element only results from the combination of both domains. The homing cycle predicts that the endonuclease domain should be lost from the intein before the splicing domain; Gimble's (31) study of inteins in different yeasts illustrates that the endonuclease domain indeed has a tendency to decay, and several inteins without endonuclease domain or with an endonuclease that probably is nonfunctional have been described (68, 98). It is reasonable to assume that most of the mini-inteins that only contain the splicing domain evolved from inteins that had a functional endonuclease domain (30, 55). This view is supported by the finding that the mini-inteins do not form a coherent group in phylogenetic reconstructions as do the hedgehog proteins, but rather that inteins without an endonuclease are found for several different intein alleles (Figure 6) (18, 70).

Most inteins contain endonucleases of the LAGLIDADG type (also called DOD-type), and the relationship between splicing and endonuclease domains is similar in the different inteins (66–68, 70, 72, 75); however, an intein in the gyrase of two cyanobacteria (*Ssp* GyrB and *Npu* GyrB) contains an HNH-type endonuclease (17, 70, 75). (These endonucleases are named after the conserved amino acids sequence motifs “LAGLIDADG” and “HNH.”) The observations that the endonuclease is always inserted between the same splicing motifs argues for only a single or a few recombination events giving rise to inteins; however, only one region in the splicing domain might be successfully invaded by an endonuclease without disruption of the excision and splicing activity. The fact that two inteins (*Ssp* GyrB

and *Npu* GyrB) contain a different endonuclease type shows that the combination between the endonuclease and splicing domains occurred at least twice.

Little phylogenetic information is retained in the primary intein sequences (18, 70) (Figure 6). Inteин alleles group together in most instances. Inteins that occupy different integration sites do not group together reproducibly in different phylogenetic reconstructions. Whereas the comparison between intein, host protein, and assumed organismal phylogeny often provides evidence for the horizontal transfer of the intein as a parasitic genetic element (see above), this information is insufficient to assess the frequency of loss or gain of the endonuclease domain.

Pietrokovski (76) recently argued that the distribution of inteins among organisms from the three domains of life demonstrated their ancient character. Given the frequency of genetic exchange across even domain boundaries (23, 61), and the fact that inteins are parasitic genetic elements whose long-term survival as functional units depends on horizontal transfer, this conclusion seems unwarranted. To the contrary, the scarcity of inteins in organisms that link recombination to sex argues that the survival of inteins over evolutionary timescales indeed depends on frequent interspecies transfer. The diverse but extremely sporadic distribution of inteins (Figure 1) is probably due to their frequent transfer and does not necessarily imply an ancient origin.

## Breaking the Homing Cycle: Acquisition of Nonparasitic Functions

Inteins are not strongly counterselected because they do not interrupt the function of the host protein they invade. Their rate of loss through excision is low because they are located in those parts of the host protein that are most important for function and therefore are under most selection pressure. An imprecise deletion would disrupt the host protein's function and is strongly counterselected. A precise deletion of the intein restores the restriction site, and the homing cycle can start again. One way for a parasitic genetic element to escape the cycle of deletion and re-invasion is to acquire a function that provides a positive selective advantage to the host organism. Among several different positive contributions that have been suggested, the one most frequently mentioned is regulation of expression [see (21, 55, 76) for recent summaries].

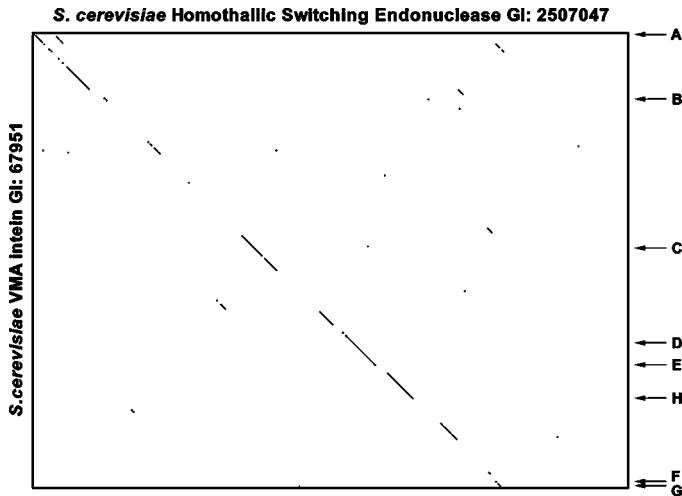
**REGULATION OF EXPRESSION AND DEVELOPMENT** The use of autocatalytic protein splicing or hydrolysis in regulation is a potential contribution of the splicing domain. Clearly, any obligatory posttranslational processing step provides opportunity for further regulation. The split inteins in the DNA polymerase of *N. punctiforme* and *Synechocystis* sp. (37, 68, 103) provide an example in which a parasitic genetic element became nearly essential for the survival of the organism. Without the intein the two exons that are encoded in different parts of the genome would not be joined together and the organism could not survive. The more elaborate posttranslational processing makes it even more difficult for the organisms to

delete the intein; however, even in this case it has not yet been demonstrated that the posttranslational splicing is actually utilized to fine-tune expression.

Paulus (66) reviewed several regulatory processes that utilize protein splicing or hydrolysis. In most instances it was not demonstrated that these processes are homologous to the splicing activity in inteins. The exceptions are the hedgehog proteins (26, 54). These proteins play an important role in animal development. During posttranslational processing the precursor is cleaved into the autoprocessing domain and the amino-terminal regulatory domain is esterified with cholesterol and secreted (80). The autoprocessing domain of the hedgehog proteins clearly is homologous to the splicing domain of inteins (52) (Figure 3). In phylogenetic analyses all hedgehog proteins group together, separate from the inteins, indicating that the transition between intein and hedgehog protein occurred only once (18). However, at present it is not clear if the ancestor was an intein that lost its endonuclease domain and whose protein-splicing domain was utilized for regulation (18), or if the common ancestor functioned in regulation and later was invaded by an endonuclease generating an intein that spread into different host proteins (76). The finding that the distribution of hedgehog proteins appears to be restricted to multicellular eukaryotes suggests the former, but given that inteins are parasitic genetic elements their wider distribution does not provide a strong argument for them being more ancient.

**PARASITES HELPING ONE ANOTHER** The homing endonuclease cleaves alleles that do not contain a parasitic genetic element that disrupts the recognition site. This feature might help phages and viruses that harbor an endonuclease in one of their genes to out-compete close relatives in mixed infections (21, 36). Support for this idea is provided by the *Bacillus subtilis* phages SPO1 and SP82. Both phages contain introns with homing endonucleases, but surprisingly both endonucleases prefer the heterologous DNA as a target. The SP82-encoded endonuclease is responsible for excluding the SPO1 intron and flanking genetic markers from the progeny of mixed infections (36).

**OTHER FUNCTIONS OF ENDONUCLEASES WITH COMPLEX RECOGNITION SITES** The most striking example of an intein acquiring a new function is that of the HO endonuclease in yeast. The role of this endonuclease is to initiate a gene conversion event that results in a switch of the mating type. The HO endonuclease catalyzes a double-strand break in the MAT locus, which initiates recombination with one of two unlinked loci that interconvert MAT $\alpha$  and MAT- $\alpha$  (53, 101). The HO endonuclease is homologous to the homing endonucleases of the LAGLIDADG-type (17, 69); in particular, the HO endonuclease is similar to the intein in the yeast *vma-1* gene. In phylogenetic analysis the HO endonuclease groups with the yeast's *vma-1* inteins (Figure 6) (77). Although it does not undergo autocatalytic splicing, the HO endonuclease contains self-splicing motifs (70) (Figure 7). The presence of a nonfunctioning self-splicing domain clearly indicates that the mating-type switching endonuclease evolved from an intein ancestor.



**Figure 7** Dot plot of *Saccharomyces cerevisiae* V-ATPase intein (*Sce* VMA) against the *Saccharomyces cerevisiae* HO endonuclease (*Sce* HO). Arrows indicate the location of the intein-specific motifs abbreviated A through G according to Perler (68, 70). *Sce* HO shows a high degree of similarity to *Sce* VMA. The similarity includes both the endonuclease and the self-splicing domains. The dot plot was generated using the Dotlet Applet (46).

Nishioka et al. (60) characterized two endonucleases that are part of the inteins in the DNA polymerase of *Pyrococcus kodakarensis* (a.k.a. *Thermococcus kodakarensis*). Both of these endonucleases have characteristics typically found in homing endonucleases. One of two enzymes (PI Pko2 or Tko Pol2) has a minimal recognition sequence of only 16 bp and cleaves the intein-free allele of the DNA polymerase. However, the enzyme also cleaves a site at the junction between intein and DNA encoding the host protein and digests chromosomal DNA from *P. kodakarensis*. The authors suggested that this endonuclease might play a role in chromosomal rearrangement similar to the role of HO endonuclease in yeast.

## Why Are Inteins Located Where They Are?

Inteins are found in conserved regions of conserved proteins (Figure 2). Many, but not all, of the host proteins are involved in nucleic acid metabolism and DNA replication. The following explanations for the selection of host proteins and integration sites have been proposed, but at present the discussion remains controversial (55, 76):

1. Many inteins are located in conserved regions because an intein in these regions cannot be easily deleted from the host protein. The regions are conserved because any mutation in this region leads to a nonfunctional protein,

and therefore changes are strongly counterselected. Any imprecise deletion of the intein will disrupt the function of the host protein and therefore will be counterselected.

2. Inteins survive in conserved regions of conserved proteins because only these regions are sufficiently conserved across species boundaries to allow the homing cycle to operate (Figure 5). In a less conserved region, populations could become immune to a parasitic genetic element through substitution of a site essential for endonuclease recognition. Although these substitutions have been observed in nature (84), presumably, they would be more frequent for an intein occupying a site that is not under strong selection.
3. Inteins would have a selective advantage if they occupied genes that are frequently transferred horizontally. Phage and viral genomes often encode their own enzymes used in DNA replication. An intein in these genes has a better chance of jumping across species boundaries and invading new populations (21, 74).
4. Homing endonucleases cleave genomic DNA. It is advantageous for the homing endonuclease to be expressed only when the machinery for DNA replication and repair is expressed as well. The homing process relies on the cell's machinery to repair the double-strand break using the intein-containing allele, and the host cell will be damaged if it is not prepared to repair double-strand breaks caused by the endonuclease (55).

## APPLICATION OF INTEINS IN BIOTECHNOLOGY

Inteins are fascinating tools from a biotechnological point of view. Recombinant protein purification has benefited most from the applications of protein splicing, but many other tools based on inteins are being developed (29).

Traditional protein purification techniques incorporate unique tags (e.g., histidine tags) into a gene to be expressed. These tags are used to bind the desired protein onto an affinity chromatography column and then release the desired protein by the addition of external agents (ions, proteases). These methods are often inconvenient because it is necessary to further purify the recombinant protein from contaminants, lowering the yield and compromising its stability, solubility, and activity.

By mutating either the N or C terminus of an intein gene, a desired gene can be ligated and expressed in frame to the intein tag (11, 12). After expression the fusion protein is extracted from the cell and purified utilizing properties of the intein. The recombinant protein is released from the mutant intein through a change in reducing conditions. This technique has been commercialized by New England Biolabs (IMPACT™). However, the release of the recombinant protein from the affinity column–intein complex can be slow, and the purified recombinant protein can be chemically modified by the reducing agent. Improvements have been made by engineering a naturally occurring intein to produce a mini-intein with compromised

activity. Through random mutation and selection a mutant mini-intein with pH-sensitive C-terminal cleavage was obtained and incorporated into a recombinant protein-affinity purification protocol (102). These inteins have been incorporated into a commercial purification kit (IMPACT-TWIN™, New England Biolabs) [for a review of current techniques see (5)]. The use of a green fluorescent protein mini-intein fusion system for protein expression (106) allows for the direct correlation between whole-cell fluorescence and protein yield. This greatly simplifies the tedious and difficult task of optimizing expression of fusion proteins.

An interesting application of inteins is the semisynthetic synthesis of cytotoxic proteins (27). Part of a cytotoxic protein is fused to an intein and expressed *in vivo*. Subsequent *in vitro* processing of the fusion protein results in cleavage of the intein, and addition of a synthetic peptide then yields the full-length and active cytotoxic proteins.

*In vitro* trans-splicing allows introducing nuclear magnetic resonance (NMR) labels into only a part of a large protein (62, 63). This approach promises to permit structural analysis of proteins over 50 kDa by NMR.

Daugelat & Jacobs (19) reported use of inteins in epitope mapping and antigen screening. In ORFTRAP, open reading frame fragments are selectively cloned into the *Mycobacterium tuberculosis recA* intein inserted in-frame into a kanamycin resistance gene. Only the clones containing a DNA fragment without stop codons and frameshifts (an open reading frame fragment) allow the correct intein splicing reaction and the kanamycin resistance phenotype to be expressed. Although this technique is in its preliminary stages, it could be further developed to overcome some of the initial drawbacks detected (e.g., only small inserts up to 425 bp can be cloned).

Peptide libraries are important tools in the search for ligand diversity in pharmacology. An intein-based method to biosynthesize backbone cyclic peptide libraries has been developed using the *Synechocystis* sp. PCC6803 *dnaE* split intein (28, 87, 88). Libraries of small, stable cyclic peptides were generated through site-directed mutagenesis. Libraries with  $10^7$ – $10^8$  primary transformants were generated for cyclic peptides with five variable residues and either one or four fixed residues.

Split inteins also form the basis for several studies of protein-protein interaction and active-site dissection. An optical probe to study protein-protein interactions has been made from the amino and carboxy termini of the split intein in *Synechocystis* sp. PCC6803 *dnaE*, fused to the amino and carboxy fragments of firefly luciferase. The luciferase fragments in turn are linked to the proteins of interest (64). Protein-protein interaction triggers folding and trans-splicing of the *dnaE* intein, thereby recovering the luciferase enzymatic activity. Phosphorylation cascades, integral membrane protein interactions, and high-throughput drug screening are some of the potential applications of this technique. The N and C terminus of the green fluorescent protein have also been used as a reporter system for protein-protein interactions, using an artificial split intein engineered from PI *SceI* (65), as well as the PI *PfuI* intein from *P. furiosus* (45).

The use of split inteins in transgenic plants might allow engineering of plants whose transgenes cannot be easily transferred to other species (8, 96). Using the *dnaE* intein from *Synechocystis* sp. PCC6803 coupled to two unlinked fragments of a herbicide resistance protein might result in nontransferable genes because each of the split exteins fused to the N or C terminus of a mini-intein could be inserted into unlinked regions of a genome or into two different cellular compartments.

## CONCLUDING REMARKS

Many features of inteins make sense when considered in an evolutionary context. However, a few puzzles remain controversial: How and how often did inteins originate? What is their relation to existing regulatory proteins? Why are some protein families preferred hosts? The endonuclease and the self-splicing domains of the large inteins are utilized in nature for nonparasitic purposes, and both domains are also being used in biotechnological applications. Many of these are still in their infancy, and future advancements will depend on a better understanding and utilization of the forces that shaped inteins in vivo.

## ACKNOWLEDGMENTS

Support in the authors' lab was provided through the NASA Exobiology Program and through the NASA Astrobiology Institute at Arizona State University.

**The Annual Review of Microbiology is online at <http://micro.annualreviews.org>**

## LITERATURE CITED

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–402
2. Belfort M, Roberts RJ. 1997. Homing endonucleases: keeping the house in order. *Nucleic Acids Res.* 25:3379–88
3. Bergstrom CT, Pritchard J. 1998. Germ-line bottlenecks and the evolutionary maintenance of mitochondrial genomes. *Genetics* 149:2135–46
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–42
5. Blaschke UK, Silberstein J, Muir TW. 2000. Protein engineering by expressed protein ligation. *Methods Enzymol.* 328: 478–96
6. Bremer K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42:795–803
7. Cavalier-Smith T. 2001. Obcells as proto-organisms: membrane heredity, lithophosphorylation, and the origins of the genetic code, the first cells, and photosynthesis. *J. Mol. Evol.* 53:555–95
8. Chen L, Pradhan S, Evans TC Jr. 2001. Herbicide resistance from a divided EP-SPS protein: the split *Synechocystis* DnaE intein as an in vivo affinity domain. *Gene* 263:39–48
9. Chevalier BS, Stoddard BL. 2001. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res.* 29: 3757–74
10. Cho Y, Qiu YL, Kuhlman P, Palmer JD.

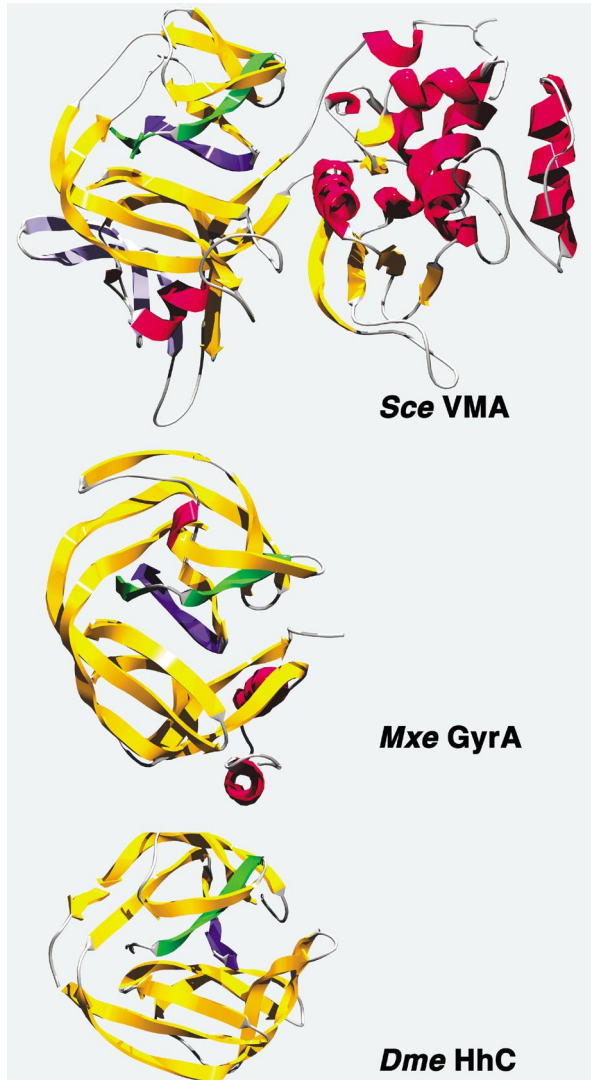
1998. Explosive invasion of plant mitochondria by a group I intron. *Proc. Natl. Acad. Sci. USA* 95:14244–49
11. Chong S, Mersha FB, Comb DG, Scott ME, Landry D, et al. 1997. Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. *Gene* 192:271–81
  12. Chong S, Montello GE, Zhang A, Cantor EJ, Liao W, et al. 1998. Utilizing the C-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step. *Nucleic Acids Res.* 26:5109–15
  13. Chong S, Xu MQ. 1997. Protein splicing of the *Saccharomyces cerevisiae* VMA intein without the endonuclease motifs. *J. Biol. Chem.* 272:15587–90
  14. Chute IC, Hu Z, Liu XQ. 1998. A topA intein in *Pyrococcus furiosus* and its relatedness to the r-gyr intein of *Methanococcus jannaschii*. *Gene* 210:85–92
  15. 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.* 12:2575–83
  16. Cooper AA, Stevens TH. 1995. Protein splicing: self-splicing of genetically mobile elements at the protein level. *Trends Biochem. Sci.* 20:351–56
  17. Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS. 1997. Statistical modeling and analysis of the LAGLI-DADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucleic Acids Res.* 25:4626–38
  18. 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
  19. Daugelat S, Jacobs WR Jr. 1999. The *Mycobacterium tuberculosis* recA intein can be used in an ORFTRAP to select for open reading frames. *Protein Sci.* 8:644–53
  20. Dawkins R. 1976. *The Selfish Gene*. Oxford, UK: Oxford Univ. Press
  21. Derbyshire V, Belfort M. 1998. Lightning strikes twice: intron-intein coincidence. *Proc. Natl. Acad. Sci. USA* 95:1356–57
  22. Derbyshire V, Wood DW, Wu W, Dansereau JT, 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* 94:11466–71
  23. Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–29
  - 23a. Douglas SE, Penny SL. 1999. The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J. Mol. Evol.* 48:236–44
  24. Duan X, Gimble FS, Quioco FA. 1997. Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell* 89:555–64
  25. Dujon B. 1989. Group I introns as mobile genetic elements: facts and mechanistic speculations—a review. *Gene* 82:91–114
  26. Echelard Y, Epstein DJ, St-Jacques B, Shen L, Mohler J, et al. 1993. Sonic hedgehog, a member of a family of putative signaling molecules, is implicated in the regulation of CNS polarity. *Cell* 75:1417–30
  27. Evans TC Jr, Benner J, Xu MQ. 1998. Semisynthesis of cytotoxic proteins using a modified protein splicing element. *Protein Sci.* 7:2256–64
  28. Evans TC Jr, Martin D, Kolly R, Panne D, Sun L, et al. 2000. Protein trans-splicing and cyclization by a naturally split intein from the *dnaE* gene of *Synechocystis* species PCC6803. *J. Biol. Chem.* 275:9091–94

29. Evans TC Jr, Xu MQ. 1999. Intein-mediated protein ligation: harnessing nature's escape artists. *Biopolymers* 51:333–42
30. Gimble FS. 2000. Invasion of a multitude of genetic niches by mobile endonuclease genes. *FEMS Microbiol. Lett.* 185:99–107
31. Gimble FS. 2001. Degeneration of a homing endonuclease and its target sequence in a wild yeast strain. *Nucleic Acids Res.* 29:4215–23
32. 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.* 270:5849–56
33. Gimble FS, Thorner J. 1992. Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. *Nature* 357:301–6
34. Gimble FS, Wang J. 1996. Substrate recognition and induced DNA distortion by the PI-SceI endonuclease, an enzyme generated by protein splicing. *J. Mol. Biol.* 263:163–80
35. Goddard MR, Burt A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci. USA* 96:13880–85
36. Goodrich-Blair H, Shub DA. 1996. Beyond homing: competition between intron endonucleases confers a selective advantage on flanking genetic markers. *Cell* 84:211–21
37. Gorbalenya AE. 1998. Non-canonical inteins. *Nucleic Acids Res.* 26:1741–48
38. Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18:2714–23
39. 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
40. He Z, Crist M, Yen H, Duan X, Quiocho FA, Gimble FS. 1998. Amino acid residues in both the protein splicing and endonuclease domains of the PI-SceI intein mediate DNA binding. *J. Biol. Chem.* 273:4607–15
41. Hirata R, Ohsumk 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.* 265:6726–33
42. Hodges RA, Perler FB, Noren CJ, Jack WE. 1992. Protein splicing removes intervening sequences in an archaea DNA polymerase. *Nucleic Acids Res.* 20:6153–57
43. Hu D, Crist M, Duan X, Quiocho FA, Gimble FS. 2000. Probing the structure of the PI-SceI-DNA complex by affinity cleavage and affinity photocross-linking. *J. Biol. Chem.* 275:2705–12
44. Hughey R, Karplus K, Krogh A. 2000. <http://www.cse.ucsc.edu/research/compbio/sam.html>
45. Iwai H, Lingel A, Pluckthun A. 2001. Cyclic green fluorescent protein produced in vivo using an artificially split PI-PfuI intein from *Pyrococcus furiosus*. *J. Biol. Chem.* 276:16548–54
46. Junier T, Pagni M. 1999. <http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html>
47. Jurica MS, Stoddard BL. 1999. Homing endonucleases: structure, function and evolution. *Cell Mol. Life Sci.* 55:1304–26
48. Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebel M, Stevens TH. 1990. Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* 250:651–57
49. 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.* 272:15668–74
50. Keefe AD, Szostak JW. 2001. Functional

- proteins from a random-sequence library. *Nature* 410:715–18
51. Klabunde T, Sharma S, Telenti A, Jacobs WR Jr, Sacchettini JC. 1998. Crystal structure of GyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing. *Nat. Struct. Biol.* 5:31–36
  52. Koonin EV. 1995. A protein splice-junction motif in hedgehog family proteins. *Trends Biochem. Sci.* 20:141–42
  53. Kostriken R, Strathern JN, Klar AJ, Hicks JB, Heffron F. 1983. A site-specific endonuclease essential for mating-type switching in *Saccharomyces cerevisiae*. *Cell* 35:167–74
  54. Lee JJ, Ekker SC, von Kessler DP, Porter JA, Sun BI, Beachy PA. 1994. Autoproteolysis in hedgehog protein biogenesis. *Science* 266:1528–37
  55. Liu XQ. 2000. Protein-splicing intein: genetic mobility, origin, and evolution. *Annu. Rev. Genet.* 34:61–76
  56. Liu XQ, Hu Z. 1997. A DnaB intein in *Rhodothermus marinus*: indication of recent intein homing across remotely related organisms. *Proc. Natl. Acad. Sci. USA* 94:7851–56
  57. Martin DD, Xu MQ, Evans TC Jr. 2001. Characterization of a naturally occurring trans-splicing intein from *Synechocystis* sp. PCC6803. *Biochemistry* 40:1393–402
  58. Mathys S, Evans TC, Chute IC, Wu H, Chong S, et al. 1999. Characterization of a self-splicing mini-intein and its conversion into autocatalytic N- and C-terminal cleavage elements: facile production of protein building blocks for protein ligation. *Gene* 231:1–13
  59. Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17:589–96
  60. Nishioka M, Fujiwara S, Takagi M, Imanaka T. 1998. Characterization of two intein homing endonucleases encoded in the DNA polymerase gene of *Pyrococcus kodakaraensis* strain KOD1. *Nucleic Acids Res.* 26:4409–12
  61. Olendzenski L, Liu L, Zhaxybayeva O, Murphey R, Shin DG, Gogarten JP. 2000. Horizontal transfer of archaeal genes into the Deinococcaceae: detection by molecular and computer-based approaches. *J. Mol. Evol.* 51:587–99
  62. Otomo T, Ito N, Kyogoku Y, Yamazaki T. 1999. NMR observation of selected segments in a larger protein: central-segment isotope labeling through intein-mediated ligation. *Biochemistry* 38:16040–44
  63. Otomo T, Teruya K, Uegaki K, Yamazaki T, Kyogoku Y. 1999. Improved segmental isotope labeling of proteins and application to a larger protein. *J. Biomol. NMR* 14:105–14
  64. Ozawa T, Kaihara A, Sato M, Tachihara K, Umezawa Y. 2001. Split luciferase as an optical probe for detecting protein-protein interactions in mammalian cells based on protein splicing. *Anal. Chem.* 73:2516–21
  65. Ozawa T, Nogami S, Sato M, Ohya Y, Umezawa Y. 2000. A fluorescent indicator for detecting protein-protein interactions in vivo based on protein splicing. *Anal. Chem.* 72:5151–57
  66. Paulus H. 2000. Protein splicing and related forms of protein autoprocessing. *Annu. Rev. Biochem.* 69:447–96
  67. Perler FB. 1999. InBase, the New England Biolabs InteIn Database. *Nucleic Acids Res.* 27:346–47
  68. Perler FB. 2000. InBase, the InteIn Database. *Nucleic Acids Res.* 28:344–45
  69. Perler FB, Davis EO, Dean GE, Gimble FS, Jack WE, et al. 1994. Protein splicing elements: inteins and exteins—a definition of terms and recommended nomenclature. *Nucleic Acids Res.* 22:1125–27
  70. Perler FB, Olsen GJ, Adam E. 1997. Compilation and analysis of intein sequences. *Nucleic Acids Res.* 25:1087–93
  71. Perler FB, Xu MQ, Paulus H. 1997. Protein splicing and autoproteolysis mechanisms. *Curr. Opin. Chem. Biol.* 1:292–99

72. Pietrokovski S. 1994. Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein Sci.* 3:2340–50
73. Pietrokovski S. 1996. A new intein in cyanobacteria and its significance for the spread of inteins. *Trends Genet.* 12:287–88
74. Pietrokovski S. 1998. Identification of a virus intein and a possible variation in the protein-splicing reaction. *Curr. Biol.* 8:R634–35
75. Pietrokovski S. 1998. Modular organization of inteins and C-terminal autocatalytic domains. *Protein Sci.* 7:64–71
76. Pietrokovski S. 2001. Intein spread and extinction in evolution. *Trends Genet.* 17:465–72
77. Pietrokovski S. 2001. *Inteins—Protein Introns*. <http://bioinfo.weizmann.ac.il/~pietro/inteins>
78. Pingoud V, Thole H, Christ F, Grindl W, Wende W, Pingoud A. 1999. Photocross-linking of the homing endonuclease PI-SceI to its recognition sequence. *J. Biol. Chem.* 274:10235–43
79. Poland BW, Xu MQ, Quioco FA. 2000. Structural insights into the protein splicing mechanism of PI-SceI. *J. Biol. Chem.* 275:16408–13
80. Porter JA, Young KE, Beachy PA. 1996. Cholesterol modification of hedgehog signaling proteins in animal development. *Science* 274:255–59
81. Reith ME, Munholland J. 1995. Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. *Plant Mol. Biol. Rep.* 13:333–35
82. Saccone C, Gissi C, Lanave C, Larizza A, Pesole G, Reyes A. 2000. Evolution of the mitochondrial genetic system: an overview. *Gene* 261:153–59
83. Saves I, Eleaume H, Dietrich J, Masson JM. 2000. The thy pol-2 intein of *Thermococcus hydrothermalis* is an isoschizomer of PI-TliI and PI-TfuII endonucleases. *Nucleic Acids Res.* 28:4391–96
84. Saves I, Ozanne V, Dietrich J, Masson JM. 2000. Inteins of *Thermococcus fumicolans* DNA polymerase are endonucleases with distinct enzymatic behaviors. *J. Biol. Chem.* 275:2335–41
85. Saves I, Westrelin F, Daffe M, Masson JM. 2001. Identification of the first eubacterial endonuclease coded by an intein allele in the pps1 gene of mycobacteria. *Nucleic Acids Res.* 29:4310–18
86. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, et al. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* 29:2994–3005
87. Scott CP, Abel-Santos E, Jones AD, Benkovic SJ. 2001. Structural requirements for the biosynthesis of backbone cyclic peptide libraries. *Chem. Biol.* 8:801–15
88. Scott CP, Abel-Santos E, Wall M, Wahnon DC, Benkovic SJ. 1999. Production of cyclic peptides and proteins in vivo. *Proc. Natl. Acad. Sci. USA* 96:13638–43
89. Senejani AG, Hilario E, Gogarten JP. 2001. The intein of the *Thermoplasma* A-ATPase A subunit: structure, evolution and expression in *E. coli*. *BMC Biochem.* 2:13
90. Shao Y, Paulus H. 1997. Protein splicing: estimation of the rate of O-N and S-N acyl rearrangements, the last step of the splicing process. *J. Pept. Res.* 50:193–98
91. Shih CK, Wagner R, Feinstein S, Kanik-Ennulat C, Neff N. 1988. A dominant trifluoperazine resistance gene from *Saccharomyces cerevisiae* has homology with FOF1 ATP synthase and confers calcium-sensitive growth. *Mol. Cell. Biol.* 8:3094–103
92. Shingledecker K, Jiang SQ, Paulus H. 1998. Molecular dissection of the *Mycobacterium tuberculosis* RecA intein: design of a minimal intein and of a trans-splicing system involving two intein fragments. *Gene* 207:187–95
93. Silva GH, Dalgaard JZ, Belfort M, Van Roey P. 1999. Crystal structure of the

- thermostable archaeal intron-encoded endonuclease I-DmoI. *J. Mol. Biol.* 286: 1123–36
94. Southworth MW, Amaya K, Evans TC, Xu MQ, Perler FB. 1999. Purification of proteins fused to either the amino or carboxy terminus of the *Mycobacterium xenopi* gyrase A intein. *Biotechniques* 27: 110–14, 116, 118–20
95. Strimmer K, von Haeseler A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13:964–69
96. Sun L, Ghosh I, Paulus H, Xu MQ. 2001. Protein trans-splicing to produce herbicide-resistant acetolactate synthase. *Appl. Environ. Microbiol.* 67:1025–29
97. Swofford D. 1998. *PAUP\* 4.0 beta version, Phylogenetic Analysis Using Parsimony (and Other Methods)*. Sunderland, MA: Sinauer
98. Telenti A, Southworth M, Alcaide F, Daugelet S, Jacobs WR Jr, Perler FB. 1997. The *Mycobacterium xenopi* GyrA protein splicing element: characterization of a minimal intein. *J. Bacteriol.* 179:6378–82
99. Turmel M, Otis C, Cote V, Lemieux C. 1997. Evolutionarily conserved and functionally important residues in the I-CeuI homing endonuclease. *Nucleic Acids Res.* 25:2610–19
100. Van de Peer Y, De Rijk P, Wuyts J, Winkelmans T, De Wachter R. 2000. The European small subunit ribosomal RNA database. *Nucleic Acids Res.* 28:175–76
101. Wang R, Jin Y, Norris D. 1997. Identification of a protein that binds to the HO endonuclease recognition sequence at the yeast mating type locus. *Mol. Cell. Biol.* 17:770–77
102. Wood DW, Wu W, Belfort G, Derbyshire V, Belfort M. 1999. A genetic system yields self-cleaving inteins for bioseparations. *Nat. Biotechnol.* 17:889–92
103. Wu H, Hu Z, Liu XQ. 1998. Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803. *Proc. Natl. Acad. Sci. USA* 95: 9226–31
104. Wu H, Xu MQ, Liu XQ. 1998. Protein trans-splicing and functional mini-inteins of a cyanobacterial dnaB intein. *Biochim. Biophys. Acta* 1387:422–32
105. Xu MQ, Perler FB. 1996. The mechanism of protein splicing and its modulation by mutation. *EMBO J.* 15:5146–53
106. Zhang A, Gonzalez SM, Cantor EJ, Chong S. 2001. Construction of a mini-intein fusion system to allow both direct monitoring of soluble protein expression and rapid purification of target proteins. *Gene* 275:241–52



**Figure 3** Comparison of intein and hedgehog protein structures. The structures of the *Sce* VMA, *Mxe* GyrA, and the autoprocessing domain of the hedgehog protein from *Drosophila melanogaster* have been determined by X-ray crystallography (24, 39, 51). The pdb-files were retrieved from the Protein Data Bank (4) and processed using the Swiss-PdbViewer (38). The orange arrows and red helices indicate  $\beta$ -sheets and  $\alpha$ -helices. The amino-terminal  $\beta$ -sheets are colored in green, and the carboxy-terminal  $\beta$ -sheets are shown as blue arrows. The endonuclease domain present in *Sce* VMA (top panel, right) forms a domain clearly distinct from the self-splicing domain (top panel, left). The part of the *Sce* VMA structure that is not part of the endonuclease domain, but partakes in DNA binding, is depicted in light blue.