

Intein spread and extinction in evolution

Shmuel Pietrokovski

Inteins are selfish DNA elements found within coding regions. They are translated with their host protein, but then catalyze their own excision and the formation of a peptide bond between their flanking protein regions. Understanding what drives and selects inteins is relevant for assessing whether they have unidentified biological functions and whether they can invade and become established in new genes and organisms. Inteins are suggested to have been present and more common in the progenitors of eukaryotes and prokaryotes. In these cells, inteins had some beneficial function or had evolved from an unknown beneficial protein. Since then, this putative benefit has been lost and inteins are gradually becoming extinct. The proteins in which inteins are currently found are proposed to be proteins vital for the survival of the organism, where intein removal is most difficult.

Inteins are genetic elements present in protein-coding sequences. Following translation, the intein protein efficiently removes itself from the host protein in an autoproteolytic protein-splicing reaction that re-ligates its two flanks (Fig. 1). The intein protein-splicing domain is ~140 amino acids long and is sufficient for carrying out the protein-splicing reaction. Most inteins also include an endonuclease domain that is not involved in protein splicing. More than 100 inteins are known, occurring in eukaryotes, bacteria and archaea (<http://bioinfo.weizmann.ac.il/~pietro/inteins> and <http://www.neb.com/neb/inteins.html>). Although much progress is being made in understanding the biochemistry of protein splicing and intein structure^{2,3}, the genetics of inteins is less well understood⁴. Determining the driving forces of intein transmission and selection will tell us whether inteins are spreading or going extinct, where can they be found and, perhaps, identify whether they have, or had, functions we are not aware of.

Intein occurrence

Intein occurrence can be examined by looking at the points in which they are integrated, the proteins that host them and the species in which they occur. Inteins have been found integrated at 48 different positions in 33 different types of host protein, with some hosts having several integration points (Table 1). Inteins integrated in the same point in orthologous genes are termed intein alleles. Most intein alleles are more similar to each other than to other (nonallelic) inteins and, thus, seem to have diverged from a single gene ancestor⁵. Similarity between intein alleles is probably due both to vertical transmission during speciation events and to horizontal transmission between strains and species. Horizontal transmission of inteins is inferred from the cases where sequence

similarity between intein allele sequences is greater than that between their host proteins, and when codon usage and GC content of intein alleles is different from their hosts^{6,7}.

Nonallelic intein integration points have no amino acid sequence features in common, apart from a cysteine, serine or threonine residue immediately C-terminal to the intein (Fig. 2). The reactive sulfur or oxygen side-chain of these residues has a crucial role in the protein-splicing reaction. The protein structure of some intein integration points can be modeled, and no common structural features are seen for these either. Nevertheless, it is possible that before inteins are spliced out, the host proteins are folded differently to their mature forms⁸.

Although nonallelic intein integration points are not similar to each other, each allelic integration point is found in highly conserved sequence motifs⁹. Thus, for example, the three different nonallelic integration points present in archaeal family B DNA polymerases are each in different conserved motifs that are located next to the enzyme's active site⁸.

Intein host proteins are very diverse, including DNA and RNA polymerases, ribonucleotide reductases, H⁺ ATPases, proteases, metabolic enzymes and more. Some of these proteins have several points at which inteins can integrate, and there are cases where two or three inteins are found simultaneously on one protein¹⁰ (Table 1, <http://bioinfo.weizmann.ac.il/~pietro/inteins> and <http://www.neb.com/neb/inteins.html>). A common feature of intein host proteins is ancient origin; most seem to have been present in the ancestors of eukaryotes and prokaryotes (Table 1 and Fig. 3).

Liu suggested recently that there is a bias towards inteins being present in proteins involved in nucleic acid metabolism⁴. Three reasons are offered for this bias. First, these types of protein are frequently found on viruses and phage that might propagate inteins. Second, integration in such proteins ensures that the inteins are expressed during replication and repair of the host cell DNA. This aids the homing activity of the intein (see below). Third, expression during this period reduces the risk for the cell (and selection against the intein) by allowing efficient repair of possible nonspecific activity of the intein endonuclease domain.

These are reasonable explanations (and could be tested as suggested), but they do not address why many inteins are found in other types of proteins (Table 1). Also, it is not certain that there actually is the

Shmuel Pietrokovski
Molecular Genetics Dept,
The Weizmann Institute of
Science, Rehovot 76100,
Israel.
e-mail: pietro@bioinfo.weizmann.ac.il

Fig. 1. Protein splicing. A protein-coding gene containing an intron element (in color) is transcribed to RNA and translated into a protein. The precursor protein is matured by protein splicing, which removes the intron and re-ligates the protein flanks.

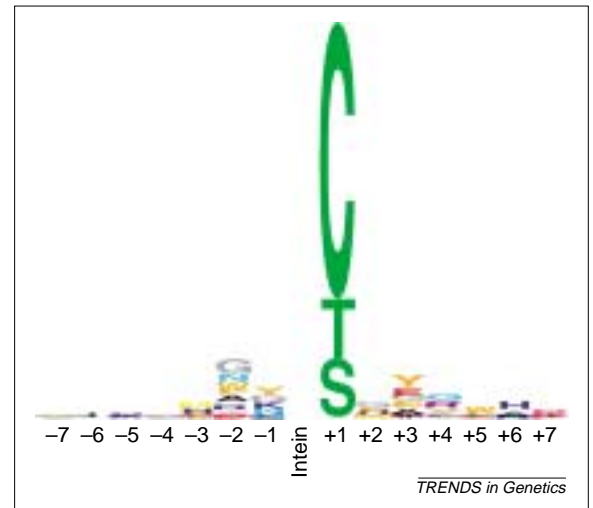
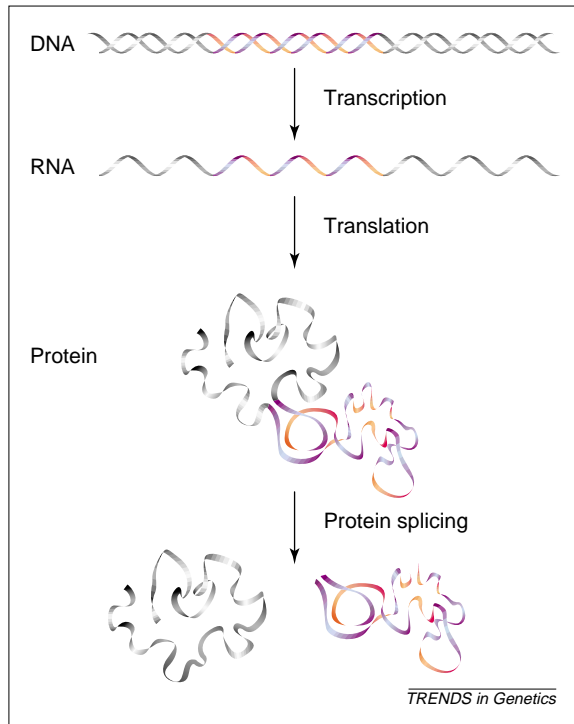


Fig. 2. Sequence conservation of intron integration points. The heights of amino acid residues and positions are proportional to their conservation. Amino acid residues of similar nature are colored the same: green, polar hydroxyl/sulfhydryl groups (cysteine, serine, threonine); cyan, polar amide groups (asparagine, glutamine); black, hydrophobic residues (alanine, valine, leucine, isoleucine); red, negatively charged (aspartate, glutamate); blue, positively charged (arginine, lysine, histidine); orange, aromatic (phenylalanine, tryptophan, tyrosine); yellow, methionine; purple, proline; gray, glycine. Ninety-eight integration points were aligned by their intron position and sequence weighing was used to correct for the uneven representation of different integration points⁴⁷.

proposed bias. In considering the number of inteins per protein type, inteins should be counted in fully sequenced genomes using the number of integration points or the total per genome; they should not be counted by the total number of all known inteins. This last number is biased by the presence of inteins in a few types of well-studied proteins, such as archaeal DNA polymerases. In addition, proteins involved in nucleic acid metabolism are crucial proteins that have a relatively large number of conserved regions from which inteins are very difficult to remove without affecting their host activity. So if the bias is true, it might mainly reflect the difficulty in removing inteins rather than a putative function of inteins in nucleic acid metabolising enzymes.

Various bacteria, archaea, chloroplasts and single-celled eukaryotes contain inteins (<http://bioinfo.weizmann.ac.il/~pietro/inteins> and <http://www.neb.com/neb/inteins.html>). Nevertheless, this far-reaching distribution is phylogenetically extremely disconnected (Table 2 and Fig. 4). Complete and almost-complete genomes of a nematode, insect, plant and mammal do not seem to contain inteins, and no intron has been identified in the available sequence data from any other multicellular organism. Inteins are either absent from or very rare in the current multicellular model organisms, or are beyond our current recognition skills.

Thus, inteins are sporadically found across diverse organisms, protein types and integration points, making them very difficult to predict in species and proteins that are not yet sequenced. This is important when we wish to find inteins for use in biotechnology¹¹, and, conversely, when a protein we plan to use or study is found to contain inteins¹⁰.

Intron structure

Sequence similarity between nonallelic inteins is extremely low (most can only be aligned across short sequence regions with low significance scores), but all inteins contain five or six common sequence motifs in the protein-splicing domain¹² that form their active site. The three intron structures^{2,13,14} that have been solved are very similar in their protein-splicing domains. Most inteins also have an endonuclease domain inserted in the protein-splicing domain¹⁵. The structures of two such inteins have an additional small DNA-binding domain^{2,13} (Fig. 5).

The DNA-binding domains probably increase the specificity of the endonuclease domains^{2,13}. Each domain has a novel DNA-binding fold. Considering the typical sizes of the protein-splicing and endonuclease domains (140–160 and 200–240 amino acids, respectively), it is possible that other new DNA-binding domains might be present in inteins longer than 350–400 amino acids.

Why are inteins retained?

Negligible effect on fitness

There is no evidence for an intron function that would be useful for their hosts. Replacement of a gene with its intron-containing homolog from a related organism showed no difference in the gene function^{16,17}, and the presence of inteins differs in related species and even strains of same species^{18–20}. Inteins are thus considered to be selfish genetic elements²¹. Because inteins (either the gene or its protein product) do not seem to contribute anything to their hosts, they constantly face the possibility of extinction through selection. Inteins are tolerated

Table 1. Intein host proteins

| | | No. intein integration points | Phylogenetic domains in which the family is found | | |
|--|--|-------------------------------|---|---------|-----------|
| | | | Bacteria | Archaea | Eukaryota |
| DNA replication, transcription and maintenance | | | | | |
| DNA polymerase catalytic subunits | Bacterial type I | 1 | + | - | + |
| | Bacterial type III | 1 | + | + | + |
| | Archaeal type B | 3 | + | - | - |
| | Archaeal PolIII DP2 subunit | 1 | - | + | - |
| RNA polymerase | Archaeal subunit A' | 1 | + | + | + |
| | Archaeal subunit A'' | 1 | + | + | + |
| Bacterial RecA and Archaeal RadA DNA-binding proteins | | 3 | + | + | + |
| Helicases | Bacterial and plastid DnaB helicases | 2 | + | - | + |
| | Archaeal large helicase-related helicase | 1 | + | + | + |
| | Bacterial SNF2 helicase | 1 | + | + | - |
| | Archaeal DEAD/DEAH-box helicase | 1 | + | + | + |
| Gyrases | Bacterial DNA gyrase subunit A | 1 | + | + | + |
| | Bacterial DNA gyrase subunit B | 1 | + | + | + |
| | Archaeal DNA reverse gyrase/topoisomerase domain | 1 | + | + | + |
| Archaeal replication factor C, 37 kD subunit and bacterial DNA pol III γ/τ subunit | | 4 | + | + | + |
| Metabolism and energy production | | | | | |
| Ribonucleotide reductase | Bacterial and eukaryotic aerobic class I α -subunit and archaeal class II | 3 | + | + | + |
| | Bacterial aerobic class I β -subunit | 1 | + | - | + |
| | Archaeal anaerobic class III | 2 | + | + | - |
| Archaeal glutamine fructose-6-phosphate transaminase | | 1 | + | + | + |
| Archaeal phosphoenol pyruvate synthase | | 1 | + | + | + |
| Archaeal UDP-glucose dehydrogenase | | 1 | + | + | + |
| Archaeal and eukaryotic vacuolar-type ATPase catalytic subunit | | 2 | + | + | + |
| Proteases | | | | | |
| Archaeal ATP-dependent protease LA (lon gene) | | 1 | + | + | + |
| Chloroplast Clp protease catalytic subunit (ClpP) | | 1 | + | - | + |
| Others | | | | | |
| Archaeal transcription factor IIB | | 1 | - | + | + |
| Archaeal translation initiation factor bIF-2 | | 1 | + | + | + |
| Archaeal cell division control protein 21 (CDC21) | | 2 | + | + | + |
| Archaeal virulence protein (KibA) | | 1 | + | + | - |
| Archaeal molybdenum cofactor biosynthesis protein A (MoaA) | | 1 | + | + | + |
| Archaeal uncharacterized protein | Family UPF0027, <i>E. coli</i> rtcB homologs | 1 | + | + | + |
| Bacterial uncharacterized protein | Family UPF0051, <i>E. coli</i> sufD homologs | 3 | + | + | + |
| Archaeal uncharacterized protein | <i>M. jannaschii</i> MJ0043 | 1 | + | + | - |
| Archaeal uncharacterized protein | <i>A. pernix</i> APE0745 | 1 | + | + | - |

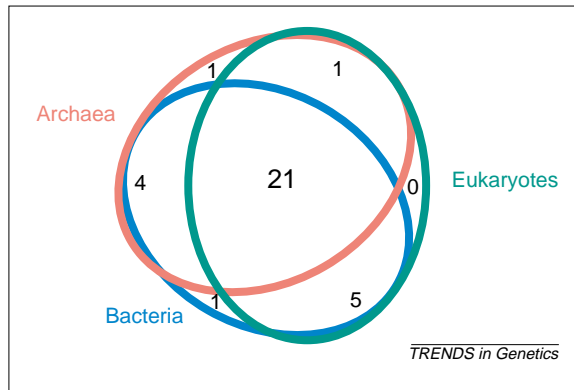
because they appear to have negligible impact on the expression of their host protein genes.

Integration in conserved sites

Intein genes are also not easy to get rid of because there is no specific host mechanism for their excision

and they are encoded in highly conserved protein-coding regions corresponding to active sites, ligand-binding sites, etc.²² If the intein gene is to be removed from the DNA, the excision must be precise⁹. Not only should the gene-reading frame be maintained, but if too much (or too little) is removed, then a deletion

Fig. 3. Phylogenetic distribution of intein hosts protein families in the three domains of life. The number of intein host protein families present in all the domains is shown where they all overlap and in the same way for families found in two or one of the domains. The families are listed in Table 1. Presence of homologs to intein protein hosts in different domains was determined by searching the NCBI nonredundant protein database with BLAST, using the intein host proteins as queries.



(or insertion) will occur in a crucial protein region. Both events will probably interfere with the proper function of the host protein and will be selected against. Occurrence of inteins in highly conserved protein motifs is probably selected for by the difficulty of removing them from these sites.

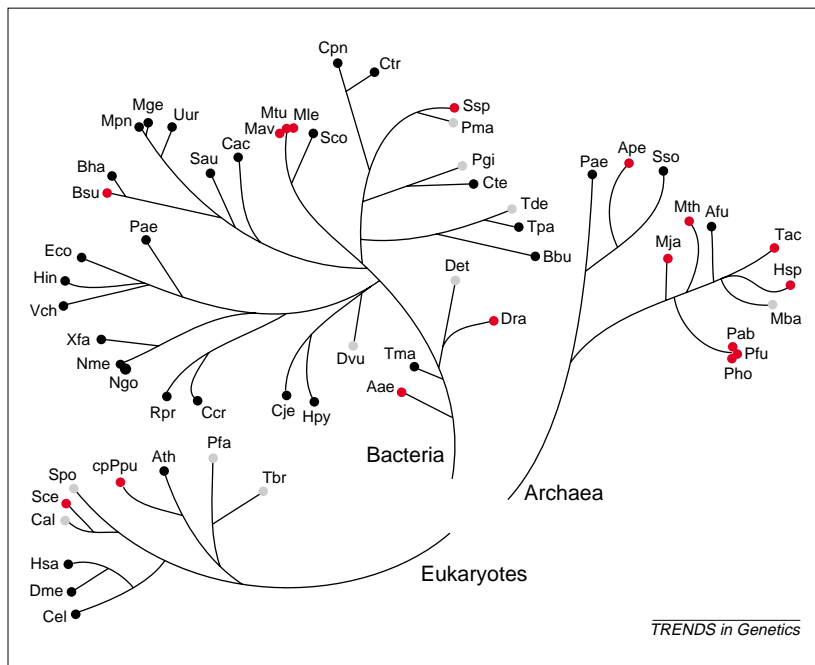


Fig. 4. Phylogenetic distribution of inteins in different species. A schematic tree representation of the relation between species whose whole genome is fully and nearly sequenced. Black, no inteins detected; red, the presence of inteins; gray, inteins are absent in a species whose genome is not fully sequenced. The tree is based on recently presented phylogenetic data^{48,49}. Bacteria: Aae, *Aquifex aeolicus*; Tma, *Thermotoga maritima*; Dvu, *Desulfovibrio vulgaris*; Hpy, *Helicobacter pylori*; Cje, *Campylobacter jejuni*; Ccr, *Caulobacter crescentus*; Rpr, *Rickettsia prowazekii*; Nme, *Neisseria meningitidis*; Ngo, *Neisseria gonorrhoeae*; Xfa, *Xylella fastidiosa*; Vch, *Vibrio cholerae*; Hin, *Haemophilus influenzae*; Eco, *Escherichia coli*; Pae, *Pseudomonas aeruginosa*; Bsu, *Bacillus subtilis*; Bha, *Bacillus halodurans*; Mpn, *Mycoplasma pneumoniae*; Mge, *Mycoplasma genitalium*; Uur, *Ureaplasma urealyticum*; Sau, *Staphylococcus aureus*; Cac, *Clostridium acetobutylicum*; Mtu, *Mycobacterium tuberculosis*; Mle, *Mycobacterium leprae*; Mav, *Mycobacterium avium*; Sco, *Streptomyces coelicolor*; Cpn, *Chlamydia pneumoniae*; Ctr, *Chlamydia trachomatis*; Ssp, *Synechocystis* sp. PCC6803; Pma, *Prochlorococcus marinus*; Pgi, *Porphyromonas gingivalis*; Cte, *Chlorobacterium tepidum*; Tde, *Treponema denticola*; Tpa, *Treponema pallidum*; Bbu, *Borrelia burgdorferi*; Det, *Dehalococcoides ethenogenes*; Dra, *Deinococcus radiodurans*. Archaea: Pae, *Pyrobaculum aerophilum*; Ape, *Aeropyrum pernix*; Sso, *Sulfolobus solfataricus*; Tac, *Thermoplasma acidophilum*; Hsp, *Halobacterium* sp. NRC-1; Pfu, *Pyrococcus furiosus*; Pho, *Pyrococcus horikoshii*; Pab, *Pyrococcus abyssi*; Afu, *Archaeoglobus fulgidus*; Mth, *Methanobacterium thermoautotrophicum*; Mja, *Methanococcus jannaschii*; Mba, *Methanosarcina barkeri*. Eukaryotes: Cel, *Caenorhabditis elegans*; Dme, *Drosophila melanogaster*; Hsa, *Homo sapiens*; Cal, *Candida albicans*; Sce, *Saccharomyces cerevisiae*; Spo, *Schizosaccharomyces pombe*; cpPpu, *Porphyra purpurea* chloroplast; Ath, *Arabidopsis thaliana*; Pfa, *Plasmodium falciparum*; Tbr, *Trypanosoma brucei*.

Intein-mediated homing

Although precise excisions of intein elements would be very rare, eventually they will occur, perhaps by gene conversion with an intein-less allele. To select against their loss, inteins possess another mechanism that enhances their survival. Most inteins have an endonuclease domain inside their protein-splicing domain. The endonuclease domains specifically cleave genes of intein-less alleles at the intein integration point. The endonuclease activity is highly specific, with very long, rare recognition sites that span intein integration points²³. Even one double-stranded break is lethal to cells, however, and must be repaired. Simple ligation of the cut would result in re-cleavage by the endonuclease domain. Repair of the cut with base changes that eliminate the endonuclease recognition site might change the amino acids coded in that site and would interfere with the function of the host protein.

Another way to repair the cut is by using the intein⁺ allele as a template. This would eliminate the endonuclease recognition site (now divided by the introduced intein element) and not interfere with the function of the new intein-containing protein. There is direct evidence for this 'homing process' in yeast¹⁸ and indirect evidence (ability of the intein endonuclease domain to cleave the intein integration sequence) from many other inteins (see intein entries in the REBASE database of endonucleases at <http://rebase.neb.com>).

Recent evidence from two intein integration points shows that these sites in intein-less alleles have point substitutions and, in one case, that this mutation makes the site immune to cleavage by the endonuclease domain of inteins integrated at this site in other proteins^{20,24}. This probably demonstrates one mechanism to avoid intein homing, although we still cannot estimate how frequently this occurs.

Another, complementary, mechanism to avoid homing was shown in a yeast mitochondrial group I intron²⁵. Initially, the intron is proposed to acquire an endonuclease domain that allows it to home into unoccupied integration points in the same and closely related species. When the point is occupied in all species, there is no strong selection for the endonuclease domain and it degenerates. This allows some species to lose the intron, and then there is a selective advantage for the intron to have an active endonuclease domain. Some intein endonuclease domains seem to have undergone inactivation^{26,27}, and this proposed mechanism could explain the sporadic distribution of inteins at each integration point.

Can some inteins benefit their hosts?

Another survival strategy, the one taken by most genes, is to promote the host organism's fitness. Although inteins are not known to benefit their hosts, it has been suggested that this does not exclude the possibility of some inteins evolving this capacity^{1,4,9,28}. An obvious way for an intein to aid its host organism

Table 2. Intein distribution in fully sequenced Archaea^a

| Protein hosts | Intein integration points ^b | Species ^c | | | | | | | | | | |
|--|--|----------------------|----|----|----|----|----|----|----|----|----|----|
| | | Mj | Ph | Pf | Pa | Mt | Af | Hs | Ta | Tv | Ap | Ss |
| Vacuolar-type ATPase, catalytic subunit | | - | + | + | + | - | - | - | + | + | - | - |
| Translation initiation factor bIF-2 | | + | + | + | + | - | - | - | - | - | - | - |
| KlB virulence protein | | + | + | + | + | - | - | - | - | - | - | - |
| Uncharacterized protein (<i>E. coli</i> rtcB homolog) | | + | + | + | + | - | - | - | - | - | - | - |
| Replication factor C, 37 kD subunit | a | + | + | + | + | - | - | - | - | - | - | - |
| | b | + | - | - | - | - | - | - | - | - | - | - |
| | c | + | - | - | + | - | - | - | - | - | - | - |
| DNA reverse gyrase/DNA topoisomerase I | | + | + | + | - | - | - | - | - | - | - | - |
| Archaeal type B DNA polymerase Pol I | a | + | - | - | - | - | - | - | - | - | - | - |
| | b | + | + | - | - | - | - | - | - | - | - | - |
| | c ^d | - | - | - | - | - | - | - | - | - | - | - |
| Archaeal DNA polymerase Pol II, DP2 subunit | | - | + | - | + | - | - | + | - | - | × | × |
| Cell division control protein 21 (CDC21) | a | - | + | - | + | - | - | + | - | - | - | - |
| | b | - | + | + | + | - | - | - | - | - | - | - |
| ATP-dependent protease LA (Ion) | | - | + | + | + | - | - | - | - | - | × | × |
| Class-II Ribonucleotide reductase | a | × | - | + | + | - | - | - | - | - | - | - |
| | b | × | + | + | + | + | - | - | - | - | - | - |
| | c | × | - | - | + | - | - | - | - | - | - | - |
| Class-III Ribonucleotide reductase | a | + | - | - | - | - | × | × | × | × | × | × |
| | b | + | - | - | - | - | × | × | × | × | × | × |
| DNA binding protein radA | | - | + | - | - | - | - | - | - | - | - | - |
| LHR – large helicase-related protein | | - | + | - | - | - | - | - | - | - | - | - |
| DEAD/DEAH-box helicase | | + | - | - | - | - | - | - | - | - | - | - |
| RNA polymerase subunit A' | | + | - | - | - | - | - | - | - | - | - | - |
| RNA polymerase subunit A'' | | + | - | - | - | - | - | - | - | - | - | - |
| Transcription factor IIB | | + | - | - | - | - | - | - | - | - | - | - |
| Phosphoenol pyruvate synthase | | + | - | - | - | - | - | - | - | - | - | - |
| UDP-glucose dehydrogenase | | + | - | - | - | - | - | - | × | × | × | - |
| Uncharacterized protein (Mja0043) | | + | - | - | - | - | × | × | × | × | × | × |
| Glutamine fructose-6-phosphate transaminase | | + | - | - | - | - | × | - | - | - | - | - |
| Molybdenum cofactor biosynthesis protein A | | - | - | - | + | - | - | - | - | - | - | - |
| Uncharacterized protein (Ape0745) | | - | - | - | - | - | - | - | - | - | + | - |
| Totals | 32 | 19 | 14 | 10 | 14 | 1 | 0 | 2 | 1 | 1 | 1 | 0 |

^a+, intein is present; -, intein is absent; ×, species does not have a homolog of the intein host protein.

^bWhen more than one integration point is present in archaeal species the points are labeled (a-c).

^cSpecies abbreviations are Mj: *Methanococcus jannaschii* DSM 2661, Ph: *Pyrococcus horikoshii* OT3, Pf: *Pyrococcus furiosus*, Pa: *Pyrococcus abyssi* GE5, Mt: *Methanobacterium thermoautotrophicum* ΔH, Af: *Archaeoglobus fulgidus* DSM 4304, Hs: *Halobacterium* sp. NRC-1, Ta: *Thermoplasma acidophilum* DSM 1728, Tv: *Thermoplasma volcanium* GSS1, Ap: *Aeropyrum pernix* K1, Ss: *Sulfolobus solfataricus* P2.

^dThis integration point is unoccupied in all of the species present in this table but is found in other archaeons.

^eSee <http://bioinfo.weizmann.ac.il/~pietro/inteins> and <http://www.neb.com/neb/inteins.html> for more details.

is to control its host protein. Because inteins are integrated in highly conserved protein regions, it is reasonable to assume that these proteins function differently (if they function at all) when the inteins are not excised⁸. This was shown experimentally for a few inteins^{29,30}. Controlling the function of a host protein will require an intein to excise itself in response to some signal. And inteins can, indeed, be engineered by single amino acid substitutions to respond to pH, temperature and thiol reagents^{31–33}.

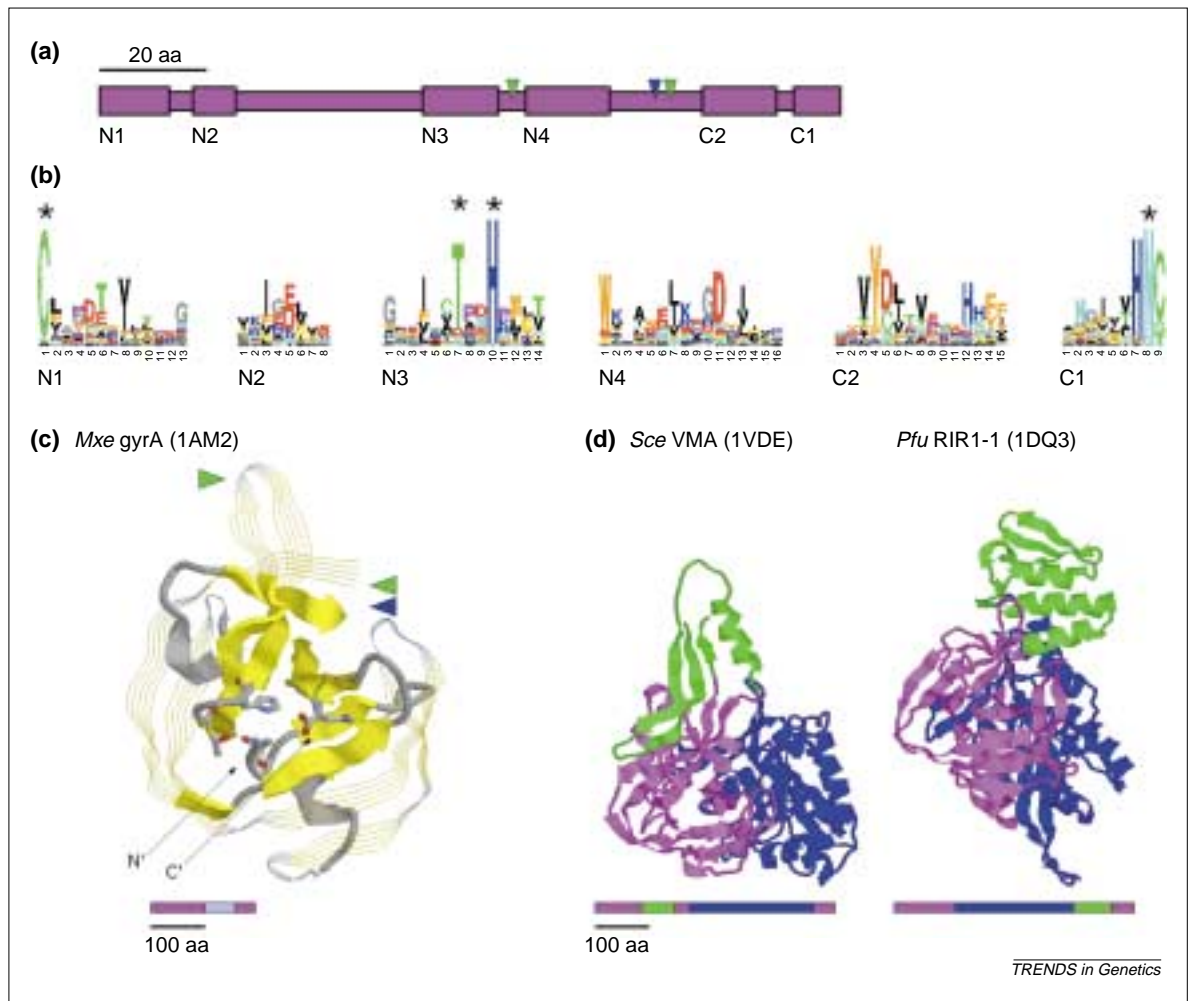
The closest evidence for an intein that might control its host protein is from the *Synechocystis* DnaE intein. This cyanobacterial intein is split, with

the N-terminal region coded downstream of the N-terminal coding region of DnaE gene and the C-terminal region coded upstream of the C-terminal coding region of DnaE gene³⁴. Protein *trans*-splicing by the split intein occurs in *E. coli* and *in vitro*^{35,36}. However, activity of this split intein has yet to be demonstrated in its native organism. It also remains to be shown that, in its native context, the *trans*-splicing has a regulatory role and is not constitutive.

The ancestral intein

Several lines of evidence indicate that prototypical inteins only had the protein-splicing domain and

Fig. 5. Intein structure. (a) Six conserved sequence motifs can be identified in intein protein-splicing domains¹². Endonuclease domains (blue arrowhead) are inserted in most inteins between the N- and C-terminal parts of the protein-splicing domain region. DNA-binding domains (green arrowheads) were found inserted in two points. (b) Sequence conservation of the motifs is shown as in Fig. 2. Residues forming the protein splicing active site are marked by asterisks. (c) Crystal structure of the protein-splicing domain of *M. xenopi* DNA gyrase subunit A intein¹⁴. This intein also has remnants of an endonuclease domain that are not shown. Regions of the six conserved sequence motifs are shown in solid and other regions as ribbons. The active site residues (Cys1, Thr72, His75 and Asn198) are shown with their side chains. Endonuclease and DNA-binding domains insertion sites are shown as in (a). Yellow, β strands. Sequence position of the domains is shown below; purple, protein-splicing domain; light purple, endonuclease domain remnant. (d) Crystal structures of inteins with endonuclease (purple) and DNA-binding domains (green). Purple, the protein-splicing domain. Below the structures are the domain positions in the sequences. The structures are shown with the same protein-splicing domain orientation as in (c).



not the endonuclease or DNA-binding domains. First, endonuclease and DNA-binding domains are not necessary for the protein-splicing reaction that is the minimal requirement for intein survival. About 15% of known inteins have no endonuclease domain, and in some other inteins the domain seems to have accumulated inactivating changes^{26,27}. Second, endonuclease domains were probably acquired more than once, because different types of homing-endonuclease domains and associated DNA-binding regions are found in inteins^{2,12,13,22}. Third, the protein-splicing domain of inteins is related by activity, structure and sequence to the C-terminal domain of the Hedgehog developmental proteins (the Hog domain)^{12,22,37}. The simplest explanation for all this evidence is that the protein-splicing and Hog domains have a common (homologous) progenitor that had no endonuclease domain^{12,37}.

Furthermore, inteins are in some respects similar to group I and group II introns. Like inteins, these introns catalyze their removal from conserved gene regions, but at the RNA level. These introns also frequently include homing endonuclease genes or domains that mediate intron homing in the same way that intein homing-endonuclease

domains do²³. Invasion of homing endonuclease into such introns has experimental evidence; an intron-encoded endonuclease can cleave its integration point in the intron as well as cleaving the intron integration point³⁸.

When did inteins appear?

Inteins are probably extremely ancient. Hedgehog proteins are found in metazoa, from invertebrates to vertebrates³⁹, and the Hog domain is also found in three other families of nematode proteins⁴⁰. This places the separation of inteins and Hedgehogs before the radiation of metazoa, because the Hog and protein-splicing domains originated from a common ancestor. As noted above, inteins appear in proteins present in organisms from all three domains of life (Fig. 3). The high sequence divergence of inteins also hints at very ancient evolutionary divergence. However, the rate of sequence change is difficult to approximate.

From the facts above, it would appear that inteins were present in the last common ancestor of bacteria, archaea and eukaryotes and certainly before the appearance of metazoa. An intein gave rise to the Hog domain, because the sequence variability of these domains is less than that of inteins^{12,22,37}.

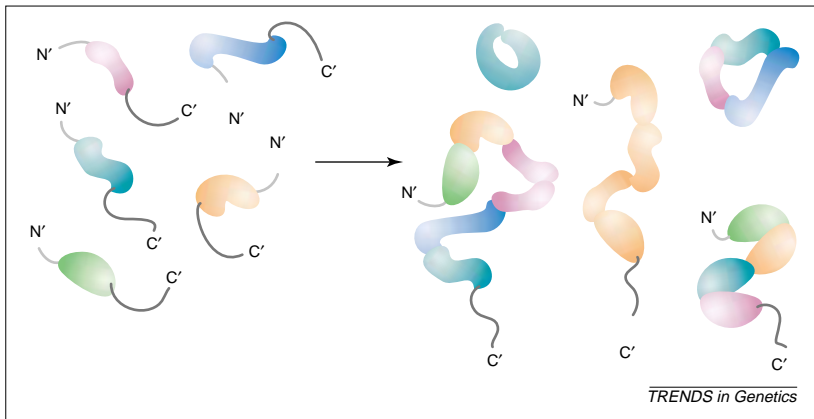


Fig. 6. Combinatorial *trans* protein-splicing. Protein segments, shown as colored ovals, with flanking N- and C-terminal intein parts, shown as black curved lines, can splice in various combinations. Only some precursors and a few of the possible combinations are shown, including cyclized proteins.

What disperses inteins?

Intein survival mechanisms lead to their sporadic dispersion. Homing allows inteins to expand by invading homologous integration points across species, and the selfish nature of inteins probably causes their loss from different species. Determining the relative importance of these two mechanisms would determine whether inteins are extremely ancient and are becoming extinct (on an evolutionary timescale), or whether they are colonizing new species and integration points, or both.

Homing certainly occurs within species, where strains exchange DNA¹⁸, and probably also occurs between species^{6,7}, where the transmission modes are unclear. However, to have a significant role in intein spread, horizontal transfer should be by transposition to new (nonhomologous) integration points. This requires a different mechanism from homing, because homing relies on sequence similarity between the donor and recipient alleles⁴¹.

Transposition events were observed for an endonuclease-containing group I intron and were suggested as a basis for the sporadic occurrence of these introns⁴². Inteins are similar to these introns in not having a specific transposition mechanism (such as that found in transposons). Nevertheless, transposition is simpler for introns than for inteins. After being spliced, introns are present as RNA molecules that can be reverse-transcribed and used to repair DNA double-stranded breaks. Unlike introns, the nucleic acid (DNA or RNA) coding for an intein is not separated from that of its host by any directed process we know. Thus, the intein will be transposed together with its host flanks. For this type of transposition to succeed, the intein must be active in its new location and the new host must function properly with the insertion of the intein flanks. Thus, inteins seem less likely to become established in new integration points than introns.

Another indication for the present minor role of transposition in the spread of inteins is the sequence similarity between inteins. The sequences of allelic

inteins are much more similar than those of nonallelic inteins, suggesting that the spread of inteins to new integration points, by whatever process, mostly occurred long ago. If transposition is not the major cause for the sporadic distribution of inteins, then not only are they extremely ancient, but there also used to be many more of them. Currently, only a few organisms have inteins, and in each of these organisms inteins are found in different integration points (Tables 1,2 and Fig. 3). I propose that formerly, organisms had inteins in all, or most, integration points found today.

Why would ancient organisms have many inteins?

The hypothesis developed here implies that inteins were more common in very early organisms. Gradually, or at a particular point, inteins began going extinct. This could have been the result of inteins losing some beneficial role, or the emergence of some process for the removal of their genes from the genome, or both. This putative early beneficial role could also have been the driving force for the emergence of inteins.

One possible function for primeval inteins is suggested by the ability of current inteins to carry out protein splicing in *trans*⁴³. *Trans*-splicing occurs in two inteins that are artificially split^{44,45} and in a naturally split intein identified in a cyanobacterium^{34,35}. Perler suggested that split inteins would mediate domain shuffling at the protein level and could also enhance it at the DNA level by providing conserved islands for recombination⁴³. Protein domains with either or both N- and C-terminal intein regions would allow the generation of multiple combinations of these domains, including tandem and separated repeats, as well as cyclized proteins (Fig. 6). Some of these products were found following expression of a protein flanked by a split intein³⁶. Combinatorial *trans*-splicing would allow a small genome to specify a large number of sequences and allow an accelerated selection of useful combinations.

Implications

Study of intein biochemistry identified a unique combination of self-catalyzed reactions¹. In addition to advancing our basic knowledge, the use of inteins as tools in biotechnology is rapidly expanding¹¹. It was recently even proposed that inteins could be used for gene therapy in humans⁴⁶. In this context, we should be aware that we still do not fully understand inteins, especially their transmission modes. *Mycobacterium tuberculosis*, *mycobacterium leprae* and *Candida tropicalis* are major intracellular human pathogens that carry several inteins and a virus from a family that infect insects, amphibia and fish also has an intein (<http://bioinfo.weizmann.ac.il/~pietro/inteins> and <http://www.neb.com/neb/inteins.html>). Research into the genetic invasion potential of inteins is desirable to assess the risk of intein-carrying pathogens and of using inteins for gene therapy.

Many intein host proteins are proteins vital for organism survival (Table 1), and many intein

integration points are recognized as functionally important positions²². However, there are some intein hosts with unknown function, and the role of many intein integration points is unidentified. Inteins can serve as natural indicators for critical proteins and important protein sites.

Currently most (or even all) known inteins are selfish genetic elements with no function that benefits their hosts. However, it is unlikely that proteins that can carry out such complex biochemical reaction would evolve without positive selection. Some beneficial function must have been present in

primeval inteins or their yet unknown progenitors. This function might be inactive in current inteins, needing some selection or specific context to emerge. The protein-splicing activity of inteins naturally lends itself to experiments where an intein is inserted into a selectable marker and the intein activity is tested after mutating it or under specific conditions³³. This conjectured 'molecular atavism' could uncover novel activities that will be useful in biotechnology and molecular biology. Perhaps of even more interest will be for us to find and study biological functions that were present in early cellular life.

Acknowledgements

This work was supported by MINERVA Foundation, Germany, and by the Crowne Human Genome Center of the Weizmann Institute.

References

- Paulus, H. (2000) Protein splicing and related forms of protein autoprocessing. *Annu. Rev. Biochem.* 69, 447–496
- Ichiyanagi, K. *et al.* (2000) Crystal structure of an archaeal intein-encoded homing endonuclease PI-PfuI. *J. Mol. Biol.* 300, 889–901
- Poland, B.W. *et al.* (2000) Structural insights into the protein splicing mechanism of PI-SceI. *J. Biol. Chem.* 275, 16408–16413
- Liu, X.Q. (2000) Protein-splicing intein: genetic mobility, origin, and evolution. *Annu. Rev. Genet.* 34, 61–76
- Perler, F.B. *et al.* (1997) Compilation and analysis of intein sequences. *Nucleic Acids Res.* 25, 1087–1093
- Liu, X.Q. and Hu, Z. (1997) A DnaB intein in *Rhodothermus marinus*: indication of recent intein homing across remotely related organisms. *Proc. Natl. Acad. Sci. U. S. A.* 94, 7851–7856
- Petrokovski, S. (1998) Identification of a virus intein and a possible variation in the protein-splicing reaction. *Curr. Biol.* 8, R634–R635
- Hashimoto, H. *et al.* (2001) Crystal structure of DNA polymerase from hyperthermophilic archaeon *Pyrococcus kodakaraensis* KOD1. *J. Mol. Biol.* 306, 469–477
- Derbyshire, V. and Belfort, M. (1998) Lightning strikes twice – intron–intein coincidence. *Proc. Natl. Acad. Sci. U. S. A.* 95, 1356–1357
- Niehaus, F. *et al.* (1997) Cloning and characterisation of a thermostable alpha-DNA polymerase from the hyperthermophilic archaeon *Thermococcus* sp. TY. *Gene* 204, 153–158
- Perler, F.B. and Adam, E. (2000) Protein splicing and its applications. *Curr. Opin. Biotechnol.* 11, 377–383
- Petrokovski, S. (1998) Modular organization of inteins and C-terminal autocatalytic domains. *Protein Sci.* 7, 64–71
- Duan, X. *et al.* (1997) Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell* 89, 555–564
- Klabunde, T. *et al.* (1998) Crystal structure of GyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing. *Nat. Struct. Biol.* 5, 31–37
- Dalgaard, J.Z. *et al.* (1997) Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucleic Acids Res.* 25, 4626–4638
- Frischkorn, K. *et al.* (1998) Investigation of mycobacterial recA function: protein introns in the RecA of pathogenic mycobacteria do not affect competency for homologous recombination. *Mol. Microbiol.* 29, 1203–1214
- Papavinasundaram, K.G. *et al.* (1998) Construction and complementation of a recA deletion mutant of *Mycobacterium smegmatis* reveals that the intein in *Mycobacterium tuberculosis* recA does not affect RecA function. *Mol. Microbiol.* 30, 525–534
- Gimble, F.S. and Thorner, J. (1992) Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. *Nature* 357, 301–306
- Maeder, D.L. *et al.* (1999) Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences. *Genetics* 152, 1299–1305
- Saves, I. *et al.* (2000) Inteins invading mycobacterial RecA proteins. *FEBS Lett.* 480, 221–225
- Belfort, M. *et al.* (1995) Prokaryotic introns and inteins: a panoply of form and function. *J. Bacteriol.* 177, 3897–3903
- Dalgaard, J.Z. *et al.* (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
- Belfort, M. and Roberts, R.J. (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res.* 25, 3379–3388
- Saves, I. *et al.* (2000) The thy pol-2 intein of *Thermococcus hydrothermalis* is an isoschizomer of PI-tIII and PI-TfIII endonucleases. *Nucleic Acids Res.* 28, 4391–4396
- Goddard, M.R. and Burt, A. (1999) Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci. U. S. A.* 96, 13880–13885
- Telenti, A. *et al.* (1997) The *Mycobacterium xenopi* GyrA protein splicing element: characterization of a minimal intein. *J. Bacteriol.* 179, 6378–6382
- Gimble, F.S. (2000) Invasion of a multitude of genetic niches by mobile endonuclease genes. *FEMS Microbiol. Lett.* 185, 99–107
- Petrokovski, S. (1996) A new intein in cyanobacteria and its significance for the spread of inteins. *Trends Genet.* 12, 287–288
- Davis, E.O. *et al.* (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
- Kawasaki, M. *et al.* (1997) Identification of three core regions essential for protein splicing of the yeast vma1 Protozyme. A random mutagenesis study of the entire vma1-derived endonuclease sequence. *J. Biol. Chem.* 272, 15668–15674
- Chong, S. *et al.* (1998) Modulation of protein splicing of the *Saccharomyces cerevisiae* vacuolar membrane ATPase intein. *J. Biol. Chem.* 273, 10567–10577
- Southworth, M.W. *et al.* (1999) Purification of proteins fused to either the amino or carboxy terminus of the *Mycobacterium xenopi* gyrase A intein. *Biotechniques* 27, 110–120
- Wood, D.W. *et al.* (1999) A genetic system yields self-cleaving inteins for bioseparations. *Nat. Biotechnol.* 17, 889–892
- Gorbalenya, A.E. (1998) Non-canonical inteins. *Nucleic Acids Res.* 26, 1741–1748
- Wu, H. *et al.* (1998) Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9226–9231
- Evans, T.C., Jr *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, 9091–9094
- Hall, T.M. *et al.* (1997) Crystal structure of a hedgehog autoprocessing domain: homology between hedgehog and self-splicing proteins. *Cell* 91, 85–97
- Loizos, N. *et al.* (1994) Evolution of mobile group I introns: recognition of intron sequences by an intron-encoded endonuclease. *Proc. Natl. Acad. Sci. U. S. A.* 91, 11983–11987
- Hammerschmidt, M. *et al.* (1997) The world according to hedgehog. *Trends Genet.* 13, 14–21
- Aspöck, G. *et al.* (1999) *Caenorhabditis elegans* has scores of hedgehog-related genes: sequence and expression analysis. *Genome Res.* 9, 909–923
- Belfort, M. and Perlman, P.S. (1995) Mechanisms of intron mobility. *J. Biol. Chem.* 270, 30237–30240
- Parker, M.M. *et al.* (1999) Intron homing with limited exon homology. Illegitimate double-strand-break repair in intron acquisition by phage T4. *Genetics* 153, 1513–1523
- Perler, F.B. (1999) A natural example of protein trans-splicing. *Trends Biochem. Sci.* 24, 209–211
- Mills, K.V. *et al.* (1998) Protein splicing in trans by purified N- and C-terminal fragments of the *Mycobacterium tuberculosis* RecA intein. *Proc. Natl. Acad. Sci. U. S. A.* 95, 3543–3548
- Southworth, M.W. *et al.* (1998) Control of protein splicing by intein fragment reassembly. *EMBO J.* 17, 918–926
- de Grey, A.D. (2000) Mitochondrial gene therapy: an arena for the biomedical use of inteins. *Trends Biotechnol.* 18, 394–399
- Henikoff, S. *et al.* (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* 163, GC 17–26
- Baldauf, S.L. *et al.* (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290, 972–977
- Nelson, K.E. *et al.* (2000) Status of genome projects for nonpathogenic bacteria and archaea. *Nat. Biotechnol.* 18, 1049–1054