Research article

The intein of the Thermoplasma A-ATPase A subunit: Structure, evolution and expression in E. coli

Alireza G Senejani1, Elena Hilario2 and J Peter Gogarten*3

Address: 1Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-3044, USA, 2Current address: HortResearch, 120 Mt Albert Road, Private Bag 92, 160 Mt Albert, Auckland, New Zealand and 3Department of Molecular and Cell Biology, University of Connecticut, 75 North Eagleville Rd. Storrs, CT 06269-3044, USA

E-mail: Alireza G Senejani - ali@carrot.mcb.uconn.edu; Elena Hilario - elena@carrot.mcb.uconn.edu; J Peter Gogarten* - gogarten@uconn.edu

*Corresponding author

Published: 14 November 2001 Received: 17 October 2001 Accepted: 14 November 2001

BMC Biochemistry 2001, 2:13

This article is available from: http://www.biomedcentral.com/1471-2091/2/13

© 2001 Senejani et al; licensee BioMed Central Ltd. Verbatim copying and redistribution of this article are permitted in any medium for any non-commercial purpose, provided this notice is preserved along with the article’s original URL. For commercial use, contact info@biomedcentral.com

Abstract

Background: Inteins are selfish genetic elements that excise themselves from the host protein during post translational processing, and religate the host protein with a peptide bond. In addition to this splicing activity, most reported inteins also contain an endonuclease domain that is important in intein propagation.

Results: The gene encoding the Thermoplasma acidophilum A-ATPase catalytic subunit A is the only one in the entire T. acidophilum genome that has been identified to contain an intein. This intein is inserted in the same position as the inteins found in the ATPase A-subunits encoding gene in Pyrococcus abyssi, P. furiosus and P. horikoshii and is found 20 amino acids upstream of the intein in the homologous vma-1 gene in Saccharomyces cerevisiae. In contrast to the other inteins in catalytic ATPase subunits, the T. acidophilum intein does not contain an endonuclease domain.

T. acidophilum has different codon usage frequencies as compared to Escherichia coli. Initially, the low abundance of rare tRNAs prevented expression of the T. acidophilum A-ATPase A subunit in E. coli. Using a strain of E. coli that expresses additional tRNAs for rare codons, the T. acidophilum A-ATPase A subunit was successfully expressed in E. coli.

Conclusions: Despite differences in pH and temperature between the E. coli and the T. acidophilum cytoplasms, the T. acidophilum intein retains efficient self-splicing activity when expressed in E. coli. The small intein in the Thermoplasma A-ATPase is closely related to the endonuclease containing intein in the Pyrococcus A-ATPase. Phylogenetic analyses suggest that this intein was horizontally transferred between Pyrococcus and Thermoplasma, and that the small intein has persisted in Thermoplasma apparently without homing.

Background

During the last decade several genes have been found to be interrupted by selfish genetic elements translated in frame with their host proteins. During post translational processing these elements excise themselves out of the host protein (see [1] and [2] for recent reviews). The sequences removed during splicing are called inteins
Inteins facilitate their excision out of the host protein without the help of any known host specific activity. This phenomenon, called protein splicing, was first discovered about a decade ago in the *Saccharomyces cerevisiae* V-ATPase catalytic subunit A [6,7]. Intein excision depends on the splicing domain of the intein and the first amino acid residue of the C-extein [8]. The inteins known to date are between 134 and 608 amino acids long, and they have been reported from all three domains of life: eukaryotes, eubacteria and archaea. Pietrokovski's webpage on inteins [http://blocks.fhcrc.org/~pietro/inteins/] currently lists more than 100 inteins in 34 different types of proteins [9]. The host proteins are diverse in function, including metabolic enzymes, DNA and RNA polymerases, gyrases, proteases, ribonucleotide reductases, and vacuolar and archaeal type ATPases. Common features suggested for these proteins are their expression during DNA replication [1] and their low substitution rate during evolution [9].

Most reported inteins are composed of two domains: one is responsible for protein splicing, and the other has endonuclease activity [10–13]. The function of the endonuclease is to spread the intein in intein-free homologs of the host protein. During this process, called homing, the gene encoding the intein-free homolog is cleaved by the endonuclease at or close to the intein integration site. During the repair of the cleaved gene, the intein is copied to the previously intein-free homolog. Gimble and Thorner [14] demonstrated intein homing in *Saccharomyces cerevisiae* using engineered V-ATPase genes from which the intein encoding portion had been previously removed. However, some inteins lack the endonuclease domain. Inteins without this domain perform autocatalytic splicing [15,16]. Homing endonucleases and the process of homing have been more intensively studied in self splicing introns [17], and the process is assumed to be similar for inteins.

*Thermoplasma acidophilum* is among sixteen archaea for which inteins have been reported to date (Intein database [http://www.neb.com/inteins/int_reg.html] [18]). Members of the genus *Thermoplasma* lack cell walls and possess a cytoskeleton. They live in hot and acidic environments, and are often found adhering to sulfur particles [19]. *T. acidophilum* grows optimally at 59°C and at an external pH between 1–2 [20]. A cytoplasmic pH of 5.5 has been measured indirectly [21].

Proton pumping ATPases/ATPsynthases are found in all groups of present day organisms [22]. The typical archaenal ATPsynthase is homologous to the eukaryotic vacuolar ATPase. Because of the high degree of sequence similarity the archaenal ATP synthase (A-type ATPase) is sometimes labeled as vacuolar or V-type ATPase. The archaenal and the vacuolar ATPase are both homologous to the bacterial F-ATPases, but the level of sequence similarity with the F-ATPases is much lower than between the V- and the A-ATPases. To date seven species have been identified to harbor inteins in their ATPase catalytic subunits. The first intein was discovered in the *vma-1* gene of *Saccharomyces* [7]. The yeast *Candida tropicalis* [23] possesses an intein in the same location. Inteins are also present in the A subunit of the A-type ATPases of *Thermoplasma acidophilum*, *T. volcanium*, *Pyrococcus abyssi*, *P. horikoshii* and *P. furiosus*.

Here we present data on the cloning and expression of the *T. acidophilum* A-ATPase A subunit in *E. coli*, and we discuss implications for the location, propagation and distribution of inteins among organisms.

**Results**

**Sequence analysis of the *T. acidophilum* intein**

The *T. acidophilum* intein was discovered while sequencing the catalytic subunit of the archaenal ATPase/ATPsynthase from *T. acidophilum* for systematic purposes [24]. More recently, the complete genome sequences of *T. acidophilum* [25] and *T. volcanii* [26] have been reported. In both instances, the catalytic subunit of the ATPsynthase is the only gene in the entire genome for which an intein was reported. Using PSI-BLAST [27] with different inteins as seeds, divergent inteins with less than ten percent sequence identity as compared to the query sequence are recovered (data not shown); however, using this approach we did not discover any additional intein or homing endonuclease encoding genes in *T. acidophilum*.

Multiple sequence alignments of diverse intein sequences identified eight motifs composed of moderately conserved residues [18,28]. The A-ATPase A subunit of the *Thermoplasma* and *Pyrococcus* intein multiple sequence alignment (with manual modification) is shown in figure 1. The *T. acidophilum* intein (173 amino acids long) is among the shortest inteins known, and the alignment with other inteins reveals the absence of sequences homologous to the typical endonuclease motifs. Only the motifs characteristic for the self splicing domain are present in the *Thermoplasma* intein (see figure 1).

The significance of the match between the *T. acidophilum* and the three pyrococcal ATPase inteins was assessed using PRSS at [http://fasta.bic.informatics.illinois.edu/ fasta/prss.htm] [29]. The P-value for this match, i.e. the probability of obtaining a match of this quality by chance alone, was calculated to be below 10⁻¹⁰. This indicates
that not only the exetins, but also the inteins themselves are recognizably homologous. In contrast, a sequence alignment of all known inteins shows intein sequences to be much more divergent (not shown). For example, the P-value for the comparison between the *T. acidophilum* and the *Saccharomyces cerevisiae* intein was 0.28, i.e. no significant similarity between the yeast and the *Thermoplasma* inteins was detectable using pairwise alignments only.

**Location of the intein within the host protein**

The vacuum and the archael ATPases are homologous to the bacterial and organellar F-type ATPases. The structure of the bovine mitochondrial F\(_1\)-ATPase has been determined by X-ray crystallography [30]. The intein insertion points in the yeasts’ V-ATPase and the *Thermoplasma* and *Pyrococcus* A-ATPases correspond to the catalytic site where the ATP binds and is hydrolyzed during the catalytic cycle. Figure 2 shows that the
inteins are located in the regions of these very conserved proteins [22] that have the lowest substitution rates.

Comparison of A-ATPase catalytic subunit and 16S rRNA phylogeny
The phylogenies of archaeal ATPase catalytic subunits and small subunit ribosomal RNAs are shown in Figure 3. Both phylogenies were calculated for a similar set of species. While the two phylogenies show some significant differences, neither of them groups the molecules from Thermoplasma with the Pyrococcus homologs. In both phylogenies several other Archaea, i.e., M. jannaschii, M. thermolithotrophicus, Thermococcus sp., Halobacterium sp., Methanosarcina and Methanobacterium thermoautotrophicus branch off between the two groups that carry inteins in their ATPase catalytic subunits. All of these intervening archaeal ATPases do not carry an intein.

Codon usage comparison
Codon usage varies among organisms. The production of tRNAs corresponds to the frequencies with which the different codons are present in their protein coding genes. The exact causes for tRNA regulation and codon usage are not completely understood; however, expression of foreign genes in an organism is often prevented by a different codon usage of the foreign gene [31]. Many Archaea have codon usage frequencies and tRNA compositions different from E. coli. The T. acidophilum A-ATPase A subunit has 763 codons. Codon AUA (Ile) is the most frequent (39/1000). In E. coli the same codon (AUA) is a rare codon present at a frequency of only 5.5
Figure 3
Comparison of vacuolar/archaeal ATPase A subunit and small subunit ribosomal RNA phylogenies. The phylogeny depicted in (A) was calculated from the extein portions of the genes only; (B) was calculated from small subunit ribosomal RNAs. The trees were calculated as unrooted, but are depicted as rooted using the eukaryotes as an outgroup to the Archaea (A), or the Crenarchaeota as an outgroup to the Euryarchaeota (B). Numbers give bootstrap values calculated from 100 bootstrapped samples analyzed using parsimony analysis. Values are only given for groups with more than 50% support. Asterisks denote branches that in the maximum likelihood evaluation were not at least 3 or 2 times larger, respectively, than their estimated standard error. All other branches were at least three times larger than their standard deviation. See experimental procedures for further details on the phylogenetic reconstruction methods used. Species whose A-subunit contains an intein are indicated by arrows.
The other two major differences in codon usage between the *T. acidophilum* A-ATPase A subunit and *E. coli* are AGG (Arg) and AGA (Arg) which are present in *T. acidophilum* A-ATPase A subunit at frequencies of 41 and 16 per 1000 respectively. In *E. coli* however, these codons are considered rare and occur with frequencies of 1.7 and 2.8 per 1000 respectively.

**Expression and intein splicing of *T. acidophilum* A-ATPase A subunit**

The gene encoding the *T. acidophilum* A-ATPase A subunit was cloned into the expression vector pET-11a (Stratagene) and transformed into *E. coli* Bl21(DE3) and *E. coli* Bl21-CodonPlus(DE3)-RIL strain for protein expression. When *E. coli* Bl21(DE3), a strain that did not express additional rare tRNAs, was transformed with the cloned *T. acidophilum* A-ATPase A subunit at frequencies of 41 and 16 per 1000 respectively. In *E. coli* however, these codons are considered rare and occur with frequencies of 1.7 and 2.8 per 1000 respectively.

SDS polyacrylamide gel electrophoresis of proteins from induced *E. coli*. Panel A, Lanes 1 and 3: *E. coli* Bl21-CodonPlus(DE3)-RIL strain transformed with empty pET-11a vector (negative control); lanes 2, 4, 5 and 6: *E. coli* Bl21-CodonPlus(DE3)-RIL strain transformed with the *Thermoplasma* A-ATPase cloned into pET-11a; lane 1-4: cells were induced for 4 hours; lane 5: cells were induced at 16°C for 16 hours; lane 6: cells were induced at 42°C for 2 hours; Sup.: supernatant. Panel B depicts the splicing process schematically. The unprocessed *T. acidophilum* A-ATPase A subunit has a calculated size of 85 kDalton. The expected molecular weight of the intein is 20 kDalton, and the religated host protein weighs approximately 65 kDalton.

16°C for 16 hours (Fig. 4A). During preparation for SDS gel electrophoreses the samples are heated to 72°C in the presence of DTT. With respect to temperature these conditions are more similar to the conditions in the *T. acidophilum* cytoplasm than the conditions in *E. coli*. Intein excision might occur only during this high temperature treatment. Therefore, proteins from induced *E. coli* were also separated under non-denaturing conditions with or without addition of DTT. The induced bands were excised from the non-denaturating gel and separated using denaturing SDS gel electrophoresis (Fig. 5). Both of the slower migrating bands visible after induction (A and B in Fig. 5), revealed only one major band corresponding to the spliced and religated A-subunit upon separation in an SDS denaturing gel. Presumably the slower migrating band was a dimer or higher aggregate of the A-subunit monomer. In none of these experiments did we find any indications for unprocessed or incompletely spliced inteins.

**Discussion**

The *Daucus carota*[32] and *Saccharomyces cerevisiae* (Alireza Senejani unpublished) catalytic V-ATPase subunits form inclusion bodies when expressed in *E. coli*. In contrast, the *T. acidophilum* subunit is expressed as a soluble protein in the *E. coli* cytoplasm. Despite the chemical and physical differences between the *E. coli* cy-
The T. acidophilum intein shows significant sequence similarity to the inteins found in the A-ATPase catalytic subunits of Pyrococcus. Moreover, these inteins are inserted into the same highly conserved sequence in the ATP binding site. This indicates that the inteins in Thermoplasma and Pyrococcus are homologous in the evolutionary sense, i.e., they are derived from a common ancestral gene. However, Pyrococcus and Thermoplasma are not considered closely related Archaea. Based on their 16S rRNA they are considered distantly related Euryarchaeotes (See Fig. 3b).

One explanation for the discrepancy between phylogenetic classification and distribution of the A-ATPase intein is horizontal gene transfer: the intein was not present in the last common ancestor of Thermoplasma and Pyrococcus, rather the intein invaded one of the lineages after their split and was more recently horizontally transferred to the other lineage. The horizontal transfer scenario is more parsimonious than the assumption of presence in the shared common ancestor because the latter requires several independently occurring losses of the intein together with its long-term persistence in the Pyrococcus and Thermoplasma lineages (cf. Fig. 3b). Horizontal transfer of whole genes and operons between divergent species is a frequent event [34–37]. Even house keeping genes are transferred between divergent species [37].

Two possibilities exist for the horizontal transfer scenario. Either the whole A-ATPase catalytic subunits was transferred, or the intein alone spread as a selfish genetic element. To discriminate between these two scenarios, we constructed the phylogeny of the host protein (Fig. 3a). The resulting phylogeny is in reasonable agreement with the ribosomal rRNA phylogeny, the main exception being the placement of Desulfurococcus. According to its 16S rRNA this organisms is clearly classified as a Crenarchoeote; however, its ATPase catalytic subunit groups with Thermococcus sp. The finding that the host protein itself does not group the genus Thermoplasma with the Pyrococci suggests that the intein alone was transferred between Thermoplasma and Pyrococcus, and that the sequence similarity between the Thermoplasma and Pyrococcus catalytic subunits was sufficient to allow homing into the same insertion site.

The dispersion of the intein as a selfish genetic element is consistent with the work of Goddard and Burt [38] on the persistence of an intron with homing endonuclease in yeast mitochondria. These authors studied the distribution of empty target sites, and introns with and without a functioning endonuclease gene among different yeasts. They concluded that the long term persistence of the intron depends on a cycle that begins with invasion of the empty target site by an intron with the help of a homing endonuclease encoded by an open reading frame within the intron. However, once the intron containing allele is fixed in the population, the endonuclease, which itself had been the reason for the rapid spread of the intron in the population, is no longer under selection, the endonuclease becomes non-functional and is lost, resulting in an intron without homing endonuclease activity. However, once the endonuclease is lost the inteon con-
taining allele becomes more likely to be replaced with an allele that has lost the intron altogether and the cycle of invasion and successive loss begins anew.

The process of intein homing is likely to depend on a similar cycle; however, the time intervals for loss of the intein are likely to be longer than the loss of the intron. The A-ATPase and the V-ATPase inteins are located in the most conserved part of the host gene. Any deletion of the intein from the gene itself needs to be precise, because any alteration of the amino acid sequence in the catalytic site is likely to result in a non-functioning enzyme. Comparative sequence analysis (Fig. 1) shows that the T. acidophilum and T. volcanium inteins do not contain an endonuclease domain. Our search of the genomes of these Archaea did not identify any endonucleases that might function in homing. While the failure to identify a homing endonuclease is not proof of absence, the presence of a homing endonuclease acting in trans would be unprecedented and has to be regarded as improbable. The cyclic reinvasion model for long term persistence of a selfish genetic element through homing [38] suggests that the small Thermoplasma intein evolved through reduction from a large endonuclease containing intein similar to the one found in the pyrococcal A-ATPase. Apparently, the small intein has been persisting in the Thermoplasma A-ATPase since the split between T. acidophilum and T. volcanium without the help of a homing endonuclease.

The cyclic reinvasion model also explains why the insertion site is in a region of very low substitution rates: The high degree of sequence similarity surrounding the integration point facilitates the intein transfer between different populations and species using the homing endonuclease. A more variable sequence surrounding the integration point would restrict homing to members of the same species, and would thus lower the chances for long term survival of the intein.

**Conclusion**

The small intein in the Thermoplasma A-ATPase is closely related to the endonuclease containing intein in the Pyrococcus A-ATPase. Phylogenies constructed with the host protein (A-ATPase catalytic subunit) and with 16S rDNA do not group these two organisms together, suggesting that the A-ATPase intein spread through horizontal gene transfer. The small intein has persisted in Thermoplasma apparently without homing. The T. acidophilum intein retains efficient self-splicing activity when expressed in E. coli. This activity does not depend on the physicochemical conditions in the T. acidophilum cytoplasm.

**Materials and Methods**

*Thermoplasma acidophilum* cultures were obtained from Denis Searcy at the University of Massachusetts in Amherst. Organisms were grown and their DNA isolated as described [39]. DNA manipulation, sequencing and cloning followed standard procedures as described in Sambrook *et al.* [40] and Ausubel *et al.* [41].

**Plasmid constructs**

The *T. acidophilum* A-ATPase A subunit encoding gene was amplified from genomic DNA using primers Ta-4 (ATGGATCTTCTCAACGAAGAGCAGTG) and Ta-5 (GAGGTGAACATATGGGAAAGATAATCATG). These primers match the coding sequence of the Thermoplasma acidophilum A-ATPase A subunit (gene identification number 9369337) and introduce restriction sites useful in subcloning. Initially, the PCR product was cloned into pCR 2.1 (TA cloning Vector, Invitrogen). After digestion with *NdeI* and *BamHI* the coding sequence was subeloned to the vector pET-11a (Stratagene) for gene expression experiments.

**Protein expression and determination**

*E. coli* Bl21(DE3) and Bl21-CodonPlus(DE3)-RIL strain (Stratagene) were used for protein expression. If not stated otherwise, transformed *E. coli* were grown overnight in a culture wheel at 37°C in Luria-Bertani broth with ampicillin (100 µg/mL). One mL of the broth was used to inoculate 10 mL of LB broth containing ampicillin and incubated at 37°C (200 rpm). After 2 hours, isopropyl thio-β-D-galactoside (IPTG) was added to a final concentration of 1 mM to induce gene expression and the cultures. Cells were harvested 4 hours after induction by centrifugation at 7000 rpm for 10 min and washed with TEP buffer (0.1 M Tris-HCl, pH 7.4, 0.01 M EDTA and 1 mM phenyl methyl sulfonyl fluoride). Cells were resuspended in 500 µL of TEP buffer and sonicated with a Braun-Sonic U with micro-tip for 4 minutes (0.5 duty cycle; power output approximately 120). The disrupted cells were centrifuged at 8000 rpm and the pellet was resuspended in 500 µL of fresh TEP buffer. Both the supernatant and the pellet were diluted by adding 6X sample buffer (10% SDS, 1.2 mg/mL bromphenol blue, 0.6 M DTT, 30% glycerol, 1 M Tris/HCl pH 6.8) and heated to 80°C for 15 minutes to denature the protein. Five to fifty microliters of this preparation were run in a 10% denaturing Tris-Tricine SDS polyacrylamide gel electrophoresis system as described by Schagger and von Jagow [42].

Non-denatured protein was prepared with the same procedure except that SDS was omitted from the buffers and that the samples were not heated. Proteins were electroeluted from the non-denaturing gel as describe here [24]. Gels were fixed with fixing solution (50% methanol, 10%
acetic acid) for 15 minutes, followed by staining in Coomassie staining solution (20% acetic acid, 0.025% Coomassie blue G-250) for one hour, followed by destaining in destaining solution (5% methanol, 10% acetic acid) [42].

**DNA sequencing and cloning**

DNA was sequenced using the ABI PRISM BigDye Terminator Cycle Sequencing (PE Applied Biosystems). Sequencing gels were run and processed in the Biotech Center (University of Connecticut)

**Codon usage**

The program Codon Usage Tabulated from GenBank (CUTG) at [http://www.kazusa.or.jp/codon/] [43] was used to calculate the codon usage of individual genes and genomes.

**Sequence retrieval, alignment and phylogenetic reconstruction**

The *Pyrococcus furiosus* A-ATPase sequence was retrieved via blastp from the unfinished genome using the web page [http://combdna.umbi.umd.edu/bags.html]. The aligned small subunit ribosomal RNA sequences were retrieved from the Ribosomal Database Project II [http://rdp.cme.msu.edu] [44]. Francine Perler (New England Biolabs) the curator of the intein database [http://www.neb.com/inteins/int_reg.html] kindly provided the sequences of all known inteins. All the other sequences were retrieved from the NCBI databank.

Sequences were aligned using CLUSTAL X 1.8 [45]. The number of substitutions per site in the aligned data set were calculated using a JAVA program written by Olga Zhaxybayeva (University of Connecticut). This program calculates and plots the number of substitutions in a sliding window of 10 aligned positions. The window is moved through the alignment one position at a time. For positions where less than 50% of the sequences have a gap, the average number of substitutions was calculated considering the gap as an additional character. Positions with gaps in more than 50% of the aligned sequences were skipped.

Phylogenies were reconstructed from amino acid sequences aligned using CLUSTAL X 1.8 [45]. The topologies of the depicted trees were calculated using neighbor joining as implemented in CLUSTAL X with correction for multiple substitutions. Branch lengths and their confidence intervals were calculated with TREE-PUZZLE 5.0 [46] using the JTT or the HKY model for substitution respectively, and assuming an among site rate variation described by a gamma distribution. Bootstrap samples were analyzed using parsimony as implemented in PAUP* 4.0 beta 8 [47] treating gaps as missing data. Each bootstrapped sample was analyzed using 10 different starting trees built through random addition, tree-branch-reconnection (TBR) branch swapping, and considering gaps as missing data. The following sequences were used for the 16S rRNA phylogeny (the sequences are available under these names from the Ribosomal Database Project II [http://rdp.cme.msu.edu/]: Sul.solfia4, Sul.acalda, Ap.pernixl, Me.thlitho, Me.janrrnA, Mb.tautot2, Pc.furiois2, Pc.abyssi, AP000001, AB016298, Tp.acidop, Arg.fulgid p, Hf.volcani, Hb.spCh2_2, Hb.salina2, Msr.mazei5, Msr.barke2, Dco.mobili.

**Acknowledgements**

The research was supported in part through the NASA Exobiology Program. The authors thank Torsten Schadtle for help with DNA sequencing. Ryan Murphy for constructing a *Thermoplasma acidophilum* genomic library, Olga Zhaxybayeva for providing computer programs, Lorraine Olendzenski for reading the manuscript and assistance with experimental procedures, and Denis Seary for help with culturing *Thermoplasma acidophilum*. A.S. thanks the Mannheim University of Applied Sciences and the University of Connecticut for support through fellowships.

**References**

1. Liu XQ: Protein-splicing intein: Genetic mobility, origin, and evolution. *Annu Rev Genet* 2000, 34:61-76
2. Paulus H: Protein splicing and related forms of protein auto-processing. *Annu Rev Biochem* 2000, 69:447-496
3. Dujon B: Group I introns as mobile genetic elements: facts and mechanistic speculations – a review. *Gene* 1989, 82:91-114
4. Perler FB, Davis EO, Dean GE, Gibble FS, Jack WE, Neff N, Noren CJ, Thorner J, Belfort M: Protein splicing elements: inteins and exteins – a definition of terms and recommended nomenclature. *Nucleic Acids Res* 1994, 22:1125-1127
5. Cooper AA, Stevens TH: Protein splicing: self-splicing of genetically mobile elements at the protein level. *Trends Biochem Sci* 1995, 20:351-356
6. Hirata R, Ohsumk Y, Nakano A, Kawasaki H, Suzuki K, Anraku Y: Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of Saccharomyces cerevisiae. *J Biol Chem* 1990, 265:6726-6733
7. Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebel M, Stevens TH: Protein splicing converts the yeast TFPI gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* 1990, 250:651-657
8. Chen L, Benner J, Perler FB: Protein splicing in the absence of an intein penultimate histidine. *J Biol Chem* 2000
9. Pietrokovskii S: Intein spread and extinction in evolution. *Trends Genet* 2001, 17:465-476
10. Pietrokovskii S: Modular organization of inteins and C-terminal autocatalytic domains. *Protein Sci* 1998, 7:64-71
11. Perler FB, Olsen GJ, Adam E: Compilation and analysis of intein sequences. *Nucleic Acids Res* 1997, 25:1087-1093
12. Dalgaard JZ, Moser MJ, Hughley R, Mian IS: Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J Comput Biol* 1997, 4:193-214
13. Duan X, Gibble FS, Quiocho FA: Crystal structure of PI-Scel, a homing endonuclease with protein splicing activity. *Cell* 1997, 89:555-564
14. Gibble FS, Thorner J: Homing of a DNA endonuclease gene by meiotic gene conversion in Saccharomyces cerevisiae. *Nature* 1992, 357:301-306
15. Chong S, Xu MQ: Protein splicing of the Saccharomyces cerevisiae VMA intein without the endonuclease motifs. *J Biol Chem* 1997, 272:15587-15590
16. Derbysheir V, Wood DW, Wu W, Dansereau JT, Dalgaard JZ, Belfort M: Genetic definition of a protein-splicing domain: functional mini-inteins support structure predictions and a model for intein evolution [published erratum appears in Proc Natl
