Molecular Microbiology (2003) **50**(5), 1569–1577 doi:10.1046/j.1365-2958.2003.03825.x

Distribution of split DnaE inteins in cyanobacteria

Jonathan Caspi, Gil Amitai, Olga Belenkiy and Shmuel Pietrokovski*
Molecular Genetics Department, The Weizmann Institute of Science, Rehovot 76100, Israel.

Summary

Inteins are genetic elements found inside the coding regions of different host proteins and are translated in frame with them. The intein-encoded protein region is removed by an autocatalytic protein-splicing reaction that ligates the host protein flanks with a peptide bond. This reaction can also occur *in trans* with the intein and host protein split in two. After translation of the two genes, the two intein parts ligate their flanking protein parts to each other, producing the mature protein. Naturally split inteins are only known in the DNA polymerase III alpha subunit (polC or *dnaE* gene) of a few cyanobacteria. Analysing the phylogenetic distribution and probable genetic propagation mode of these split inteins, we conclude that they are genetically fixed in several large cyanobacterial lineages. To test our hypothesis, we sequenced parts of the *dnaE* genes from five diverse cyanobacteria and found all species to have the same type of split intein. Our results suggest the occurrence of a genetic rearrangement in the ancestor of a large division of cyanobacteria. This event fixed the *dnaE* gene in a unique two-genes one-protein configuration in the progenitor of many cyanobacteria. Our hypothesis, findings and the cloning procedure that we established allow the identification and acquisition of many naturally split inteins. Having a large and diverse repertoire of these unique inteins will enable studies of their distinct activity and enhance their use in biotechnology.

Introduction

Inteins are genetic elements present in protein coding regions. All the element codes for a protein that is translated together with the coding region of its host gene. The intein protein is removed from the host protein by a protein-splicing reaction that joins the intein flanks with a peptide bond. This reaction is autocatalytic, fully catalysed by the intein and the residue C-terminal to it, with no need for other proteins, ATP or such molecules (Paulus, 2000). It is typically a *cis* intramolecular reaction but can also occur when the N- and C-terminal parts of the intein are split and encoded on separate protein chains, each with its own flank. A *trans* protein-splicing reaction then ligates the flanks of the two intein parts (Shingledecker *et al.*, 1998; Southworth *et al.*, 1998). A split intein is naturally present in the *dnaE* genes of a few cyanobacteria (Gorbalenya, 1998; Kaneko *et al.*, 2001; Nakamura *et al.*, 2002). Inteins are active with various flanks, in heterologous organisms and *in vitro*. Biochemical studies of the protein-splicing mechanism led to the use of typical and split inteins in diverse biotechnology applications (Perler and Adam, 2000; Ozawa *et al.*, 2001; Mootz and Muir, 2002).

Inteins are present in a variety of protein genes from diverse bacteria and archaea and in several eukaryotes. However, intein distribution is extremely sporadic. Although inteins are widely distributed, present in the three domains of life, they are relatively rare, and about 160 are currently known (Perler, 2002). Moreover, their distribution is discontinuous and irregular, with even closely related species differing in intein presence at homologous integration sites (Liu, 2000; Pietrokovski, 2001). Consequently, intein presence is very difficult to predict.

Intein distribution is the result of two parallel processes. The primary process seems to be independent of gradual loss of intein elements from separate species during evolution (Pietrokovski, 2001). Inteins are not known to contribute any advantage to their host genes or species and are believed to be selfish genetic elements (Belfort *et al.*, 1995). Active selection against inteins seems to be relatively weak because of their apparent negligible disruption of their host genes, protein products and organisms. Intein removal from the genome requires a precise DNA excision event as they are inserted in highly conserved points of genes coding for essential proteins (Derbyshire and Belfort, 1998). Counteracting this slow extinction is horizontal transfer to specific integration points, e.g. homing (Belfort and Roberts, 1997). Most intein proteins include a homing endonuclease domain that can mediate insertion of their intein gene into homologous unoccupied intein integration points (Gimble and Thorner, 1992). Hence, homing activ-

Accepted 12 September, 2003. *For correspondence. E-mail pietro@bioinfo.weizmann.ac.il; Tel. (+972) 8 934 2747; Fax (+972) 8 934 4108.

ity of inteins can reinsert their gene into integration sites that they were lost from.

Cyanobacteria are a large and diverse group of bacteria. They can be clustered by their 16S rRNA sequences into seven monophyletic evolutionary groups (Honda *et al.*, 1999). Although the exact relation between the groups is not fully certain, each group of species is highly likely to have diverged from a distinct species. The recent genome sequencing of various cyanobacteria prompted us to examine the distribution and origin of the cyanobacterial split DnaE inteins.

Here, we show that split inteins of the *dnaE* gene are common in several groups of cyanobacteria. This gene arrangement seems to be fixed in at least three large and probably related groups that include scores of known species. We cloned split inteins from five diverse species of these groups and thus show how to obtain split inteins from many different species. The distribution of these genes allows us to reconstruct their evolution. We discuss necessary steps in the transition from contiguous to split inteins and whether this transition is likely to be common.

Results

Split DnaE inteins are present in several cyanobacteria

We first screened for the presence of split inteins by database searches. DNA polymerase III alpha subunit (PolC or DnaE protein) was found to be encoded on two genes in several cyanobacteria – *Synechocystis* species PCC 6803 (Ssp PCC6803) (Kaneko *et al.*, 1996), *Synechococcus* species PCC 7002 (Ssp PCC7002) (Yu *et al.*, 2002), *Nostoc* species PCC 7120 (Nsp PCC7120, previously called *Anabaena* species PCC 7120) (Kaneko *et al.*, 2001), *Nostoc punctiforme* (Npu) (Meeks *et al.*, 2001), *Therrmosynechococcus elongatus* BP-1 (Tel) (Nakamura *et al.*, 2002) and *Trichodesmium erythraeum* (Ter, http://genome.jgi–psf.org/draft_microbes/trier/trier.home.html). The gene organization is identical in all cases. The *dnaE* coding sequences are split in the same highly conserved point into two genes. The 5′ part of the *dnaE* gene (*dnaE*1) is followed by a 5′ part of the split intein, and the *dnaE* gene 3′ part (*dnaE*2) is preceded by a 3′ part of the split intein (Fig. 1). The Ter *dnaE*1 gene also includes other inteins and group II introns (not shown). The mature DnaE

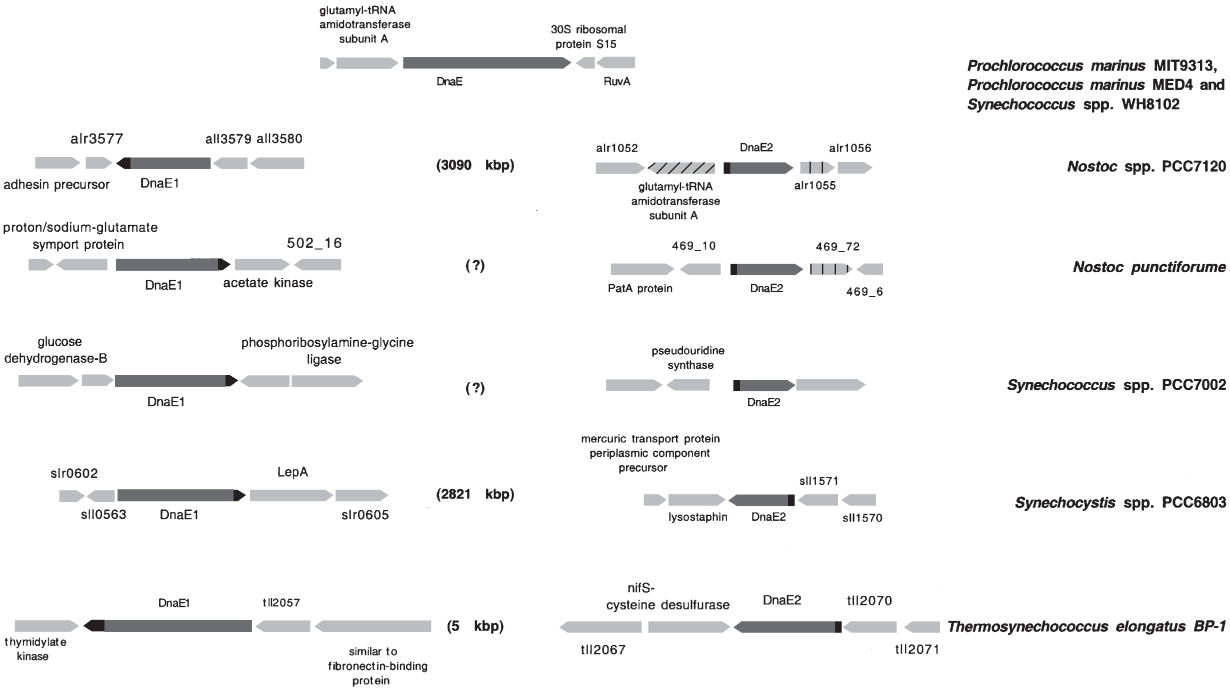

glutamyl-tRNA amidotransferase subunit A
30S ribosomal protein S15
DnaE
RuvA
Prochlorococcus marinus MIT9313, Prochlorococcus marinus MED4 and Synechococcus spp. WH8102
alr3577
all3579 all3580
adhesin precursor
DnaE1
(3090 kbp)
alr1052
DnaE2
alr1056
glutamyl-tRNA amidotransferase subunit A
alr1055
Nostoc spp. PCC7120
proton/sodium-glutamate symport protein
502_16
DnaE1
acetate kinase
(?)
469_10
469_72
PatA protein
DnaE2
469_6
Nostoc punctiforume
glucose dehydrogenase-B
phosphoribosylamine-glycine ligase
DnaE1
(?)
pseudouridine synthase
DnaE2
Synechococcus spp. PCC7002
slr0602
LepA
sll0563
DnaE1
slr0605
(2821 kbp)
mercuric transport protein periplasmic component precursor
sll1571
lysostaphin
DnaE2
sll1570
Synechocystis spp. PCC6803
DnaE1
tll2057
thymidylate kinase
similar to fibronectin-binding protein
(5 kbp)
nifS-cysteine desulfurase
DnaE2
tll2070
tll2067
tll2071
Thermosynechococcus elongatus BP-1

Fig. 1. Cyanobacterial *dnaE* gene loci. Each line shows the *dnaE* gene locus or the loci of *dnaE*1 and *dnaE*2 split genes from one species, except for the top locus that is identical in all three listed species. Gene protein-coding regions are shown as rectangles with an arrowhead at their 3′ ends. The *dnaE* genes are shown in dark grey with the split intein parts in black. Other homologous genes are indicated by similar patterns – there are only two such pairs, between *Nostoc* species PCC7120 and *Prochlorococcus marinus* MED4, and *Nostoc* species PCC7120 and *Nostoc punctiforme*. Gene names or functions are indicated where known. The distance between the split *dnaE* genes in each species is indicated where known.

protein in each species is assumed to be ligated by the split intein in a *trans* protein-splicing reaction from the separately translated DnaE1 and DnaE2 proteins (Wu *et al*., 1998; Perler, 1999).

DnaE genes were also found in three other species of cyanobacteria, *Prochlorococcus marinus* MED4 (Pma MED4) and MIT9313 (Pma MIT9313) (Hess *et al*., 2001) and *Synechococcus* species WH8102 (Ssp WH8102, http://genome.jgi-psf.org/finished_microbes/synw8/synw8.home.html). In all these species, the *dnaE* genes are contiguous, having no inteins, and are flanked by the same genes. These genes thus appear in the same genomic context (Fig. 1). Corresponding parts of known split *dnaE* genes are each flanked by different genes. Thus, split *dnaE* genes are in different genomic contexts (Fig. 1).

Distribution of split DnaE inteins

In addition to being integrated at the same point, the DnaE split intein amino acid sequences are more similar to each other then to those of other inteins (Fig. 2). The progenitor intein of all DnaE split inteins was probably a typical, contiguous intein. At some time point, this intein and its *dnaE* host gene were split by some genetic rearrangement event to form two genes (Perler, 1999).

Although all known split *dnaE* genes underwent further genomic rearrangements, evidenced by their differing genomic contexts, they retained the split intein parts. This indicates the stability of the split intein organization in dynamic genomes.

Five of the six above-mentioned known split *dnaE* genes are present in species from three of the seven distinct groups of cyanobacteria; *Thermosynechococcus elongatus* BP-1 (previously named *Synechococcus elongatus* Toray) is not placed in any of the seven groups by 16S rRNA analysis and is considered as part of an early diverged group of cyanobacteria (Honda *et al*., 1999). Cyanobacteria with known contiguous *dnaE* genes are all from a fourth group, and there is no public data on cyanobacterial *dnaE* genes from the other three defined groups (Fig. 2A).

Origin of split DnaE intein hypothesis

Split intein genes are extremely unlikely to be transferred by homing. Not only do both intein genes need to be copied, but they have no endonuclease domains. Thus,

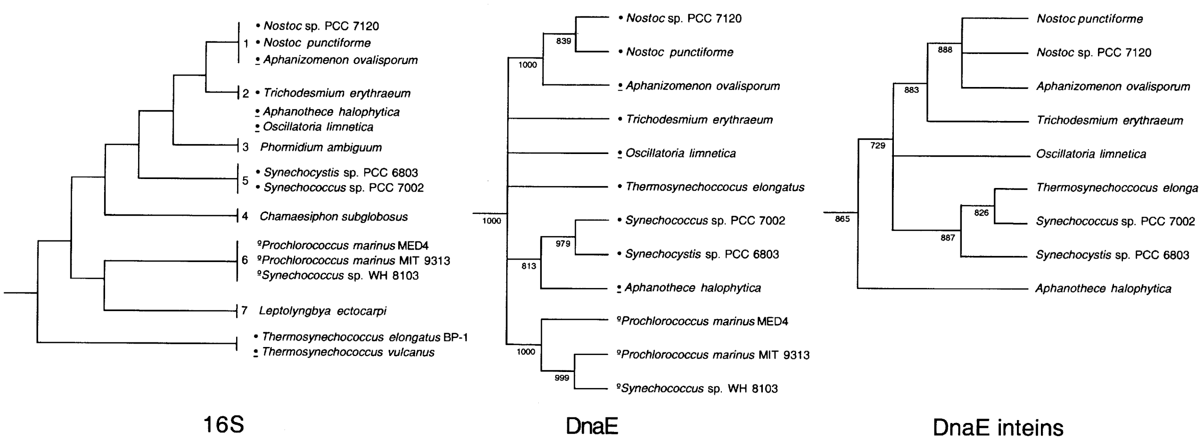

• Nostoc sp. PCC 7120
1 • Nostoc punctiforme
• Aphanizomenon ovalisporum
2 • Trichodesmium erythraeum
• Aphanothece halophytica
• Oscillatoria limnetica
3 Phormidium ambiguum
5 • Synechocystis sp. PCC 6803
• Synechococcus sp. PCC 7002
4 Chamaesiphon subglobosus
°Prochlorococcus marinus MED4
6 °Prochlorococcus marinus MIT 9313
°Synechococcus sp. WH 8103
7 Leptolyngbya ectocarpi
• Thermosynechococcus elongatus BP-1
• Thermosynechococcus vulcanus
16S
839
1000
• Nostoc sp. PCC 7120
• Nostoc punctiforme
• Aphanizomenon ovalisporum
• Trichodesmium erythraeum
• Oscillatoria limnetica
• Thermosynechoccocus elongatus
1000
979
813
• Synechococcus sp. PCC 7002
• Synechocystis sp. PCC 6803
• Aphanothece halophytica
1000
999
°Prochlorococcus marinus MED4
°Prochlorococcus marinus MIT 9313
°Synechococcus sp. WH 8103
DnaE
888
883
729
865
826
887
Nostoc punctiforme
Nostoc sp. PCC 7120
Aphanizomenon ovalisporum
Trichodesmium erythraeum
Oscillatoria limnetica
Thermosynechoccocus elongatus
Synechococcus sp. PCC 7002
Synechocystis sp. PCC 6803
Aphanothece halophytica
DnaE inteins

Fig. 2. Phylogenetic relations of cyanobacteria.
Left. A 16S rRNA-based dendogram of the seven monophyletic evolutionary groups of cyanobacteria and their inter-relations determined by Honda *et al.* (1999). Species with known DnaE sequences are marked by filled bullets for DnaE sequences split by inteins and by empty bullets for contiguous DnaE sequences. Underlined filled bullets denote split DnaE inteins identified and sequenced in this work (*Aphanizomenon*, *Aphanothece*, *Oscillatoria* and *Thermosynechococcus vulcanus*). The relations shown between the groups are the most probable ones but are not as certain as the group definitions (Honda *et al*., 1999). Listed for each group are species with known DnaE or DnaE intein sequences or a representative species chosen from the groups of Honda *et al.* (1999). The listed *Aphanothece* and *Oscillatoria* species 16S rRNA sequences were significantly most similar to each other (931/1000 bootstrap value). They were closest to group 2 species but could not be definitely clustered with it.
Centre. A dendogram calculated from conserved DnaE polymerase protein regions spanning 409 amino acids. The dendogram is rooted by the position of all cyanobacterial DnaE protein regions within a larger dendogram of DnaE proteins from various bacteria (not shown).
Right. A dendogram calculated from conserved split DnaE intein sequences spanning 132 amino acids (see Fig. 4A). The dendogram is rooted by the position of all split DnaE inteins within a larger dendogram of various inteins (not shown). Bootstrap confidence values for grouping of DnaE proteins and split inteins (values at the root) are from the larger dendograms. Bootstrap confidence values for DnaE polymerase and intein dendograms calculated from 1000 trials. Nodes below values of 700/1000 were collapsed. *Therrmosynechococcus vulcanus* sequences are not included in the analysis as only a small part upstream of its DnaE1 intein was determined. Nevertheless, the determined sequences of the intein and polymerase regions are very similar to *T. elongatus* BP-1, and the two species are expected to cluster together.

the distribution of known split DnaE inteins is probably the result of regular vertical transmission. We hypothesized that, as split DnaE inteins are present in species from three diverse cyanobacterial groups, they might be very common, or even invariably present, in other species from these groups and maybe also in other related groups (Fig. 2A).

Cloning DnaE genes from diverse cyanobacteria

To test our hypothesis, we set out to clone the DnaE genes from diverse cyanobacteria. Analysed species included *Aphanizomenon ovalisporum* (Aov), a freshwater cyanobacteria most similar by its 16S rRNA sequence to *Nostoc* species; *Microcystis* species (Vardi *et al*., 2002), a freshwater cyanobacteria belonging to group 5 (related to Ssp PCC6803 and Ssp PCC7002); *Aphanothece halophytica* (Aha) and *Oscillatoria limnetica* (Oli), unicellular and filamentous, respectively, facultative anaerobic photoautotrophic cyanobacteria that are most similar by their 16S rRNA sequences to group 2 (not shown); and *Therrmosynechococcus vulcanus* (Tvu), a thermophilic cyanobacteria species closely related to *T. elongatus* BP-1 (Nakamura *et al*., 2002).

Using sequential degenerate primer polymerase chain reaction (PCR) amplifications, we determined the sequences of conserved *dnaE* gene regions flanking one side of the known split intein integration point. We then amplified the region of the insertion site by single primer linear amplification and terminal transferase tailing (Rudi *et al*., 1999) (Fig. 3). The *dnaE* genes of all five tested species were found to be split in two and have a split intein at the same point as the other known cyanobacterial split DnaE inteins. We obtained complete sequences of the split inteins and the DnaE flanks from four of the species (Fig. 4A). Together with the previously publicly available sequences, there are now nine known complete split DnaE intein sequences.

Split DnaE intein sequence features

The length of the N′ split intein parts is between 123 and 101 amino acids, with the longest being from Ssp PCC6803 and the shortest from Aov. The length variability results from the C′ ends of these intein parts. In contrast, all eight known C′ split intein parts are 35 or 36 amino acids in length (Fig. 4A). The combined length of the Aov split intein is 137 amino acids, only three residues longer then the length of the shortest known intein, from the archaeon *Methanobacterium thermoautotrophicum* (Smith *et al*., 1997).

None of the split inteins has an endonuclease domain, like 18% of known inteins (http://bioinfo.weizmann.ac.il/~pietro/inteins). All nine split DnaE inteins have the six conserved sequence motifs that define the intein protein-splicing fold and active site (Duan *et al*., 1997; Pietrokovski, 1998). They are further conserved along their entire sequence, except for the C′ ends of the N′ parts, which differ from each other in sequence and length (Fig. 4A). The conserved minimal sequence regions of the N- and C-intein parts are those defined by computational and experimental analyses (Pietrokovski, 1998; Ghosh *et al*., 2001; Mootz and Muir, 2002).

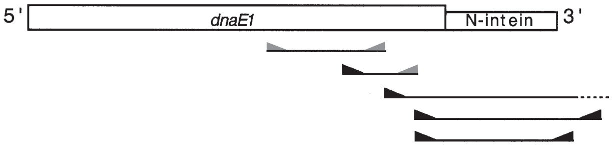


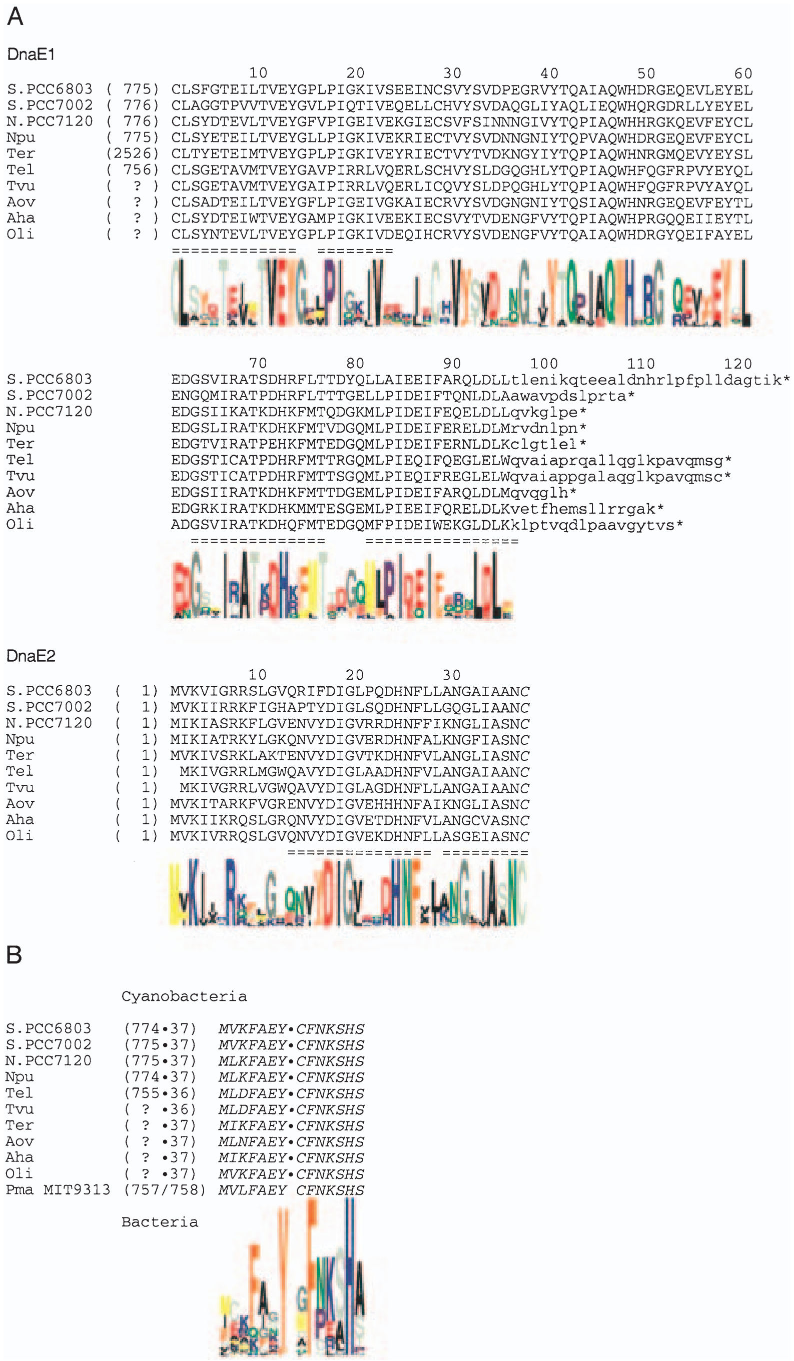
Fig. 3. Split inteins primer-walking amplification strategy. Sequence of a *dnaE*1, the *dnaE* N′ part, and its N′ intein part are shown as boxes. Amplification reactions are shown as lines beneath the sequence region amplified, with the primer regions shown as triangles. Degenerate primers are shown as grey triangles, and specific primers are shown as black triangles. Terminal transferase 3′ added tail is shown as a dotted line. The reaction order is from top to bottom. The first reaction amplified a region 5′ to the intein integration point using degenerate primers to two flanking conserved regions. The next reaction amplified the 3′ region of the previous reaction and a downstream region, using a specific 5′ primer and a degenerate 3′ primer. Typically, two or three such reactions sufficed to determine the sequence up to the insertion site. To amplify the less conserved N-intein region and its downstream 3′ untranslated region, a linear amplification using a single specific 5′ primer was followed by terminal transferase polycytosine tailing. Products from this reaction were then reamplified by a specific internal 5′ primer and a polyguanosine primer. Finally, the sequence was verified by amplifying from genomic DNA the intein part with upstream and downstream regions using specific primers. The C-intein part of the *dnaE*2 gene was cloned similarly with the amplification reactions proceeding in the reverse orientation – advancing from the *dnaE*2 central region towards its 5′ end.

Discussion

Split DnaE inteins are common in cyanobacteria

Split DnaE inteins were found to be common in cyanobacteria. They are present in all examined *dnaE* genes of species from three groups of cyanobacteria and an unassigned species (*Thermosynechococcus*) and were probably fixed in them by one common event. We suggest that most, if not all, cyanobacteria originating from the last common ancestor of species with split DnaE inteins also have this split intein. The apparent fixation of split DnaE inteins in a subdivision of cyanobacteria contrasts with the discontinuous distribution of other inteins, including those present in cyanobacteria (Table 1).

Split DnaE intein presence is a taxonomic trait that could be used in classifying cyanobacteria. We suggest it to be an ancient, highly persistent and vertically transmitted trait that separates cyanobacteria into two separate



A

DnaE1

```
                                10        20        30        40        50        60
S.PCC6803 ( 775) CLSFGTEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGEQEVLEYEL
S.PCC7002 ( 776) CLAGGTPVVTVEYGVLPIQTIVEQELLCHVYSVDAQGLIYAQLIEQWHQRGDRLLYEYEL
N.PCC7120 ( 776) CLSYDTEVLTVEYGFVPIGEIVEKGIECSVFSINNNGIVYTQPIAQWHHRGKQEVFEYCL
Npu       ( 775) CLSYETEILTVEYGLLPIGKIVEKRIECTVYSVDNNGNIYTQPVAQWHDRGEQEVFEYCL
Ter       (2526) CLTYETEIMTVEYGPLPIGKIVEYRIECTVYTVDKNGYIYTQPIAQWHNRGMQEVYEYSL
Tel       ( 756) CLSGETAVMTVEYGAVPIRRLVQERLSCHVYSLDGQGHLYTQPIAQWHFQGFRPVYEYQL
Tvu       (  ? ) CLSGETAVMTVEYGAIPIRRLVQERLICQVYSLDPQGHLYTQPIAQWHFQGFRPVYAYQL
Aov       (  ? ) CLSADTEILTVEYGFLPIGEIVGKAIECRVYSVDGNGNIYTQSIAQWHNRGEQEVFEYTL
Aha       (  ? ) CLSYDTEIWTVEYGAMPIGKIVEEKIECSVYTVDENGFVYTQPIAQWHPRGQQEIIEYTL
Oli       (  ? ) CLSYNTEVLTVEYGPLPIGKIVDEQIHCRVYSVDENGFVYTQAIAQWHDRGYQEIFAYEL
                 =============  ========

                                70        80        90        100       110       120
S.PCC6803        EDGSVIRATSDHRFLTTDYQLLAIEEIFARQLDLLtlenikqteealdnhrlpfplldagtik*
S.PCC7002        ENGQMIRATPDHRFLTTTGELLPIDEIFTQNLDLAawavpdslprta*
N.PCC7120        EDGSIIKATKDHKFMTQDGKMLPIDEIFEQELDLLqvkglpe*
Npu              EDGSLIRATKDHKFMTVDGQMLPIDEIFERELDLMrvdnlpn*
Ter              EDGTVIRATPEHKFMTEDGQMLPIDEIFERNLDLKclgtlel*
Tel              EDGSTICATPDHRFMTTRGQMLPIEQIFQEGLELWqvaiaprqallqglkpavqmsg*
Tvu              EDGSTICATPDHRFMTTSGQMLPIEQIFREGLELWqvaiappgalaqglkpavqmsc*
Aov              EDGSIIRATKDHKFMTTDGEMLPIDEIFARQLDLMqvqglh*
Aha              EDGRKIRATKDHKMMTESGEMLPIEEIFQRELDLKvetfhemsllrrgak*
Oli              ADGSVIRATKDHQFMTEDGQMFPIDEIWEKGLDLKklptvqdlpaavgytvs*
                   ==============    ================
```

DnaE2

```
                                10        20        30
S.PCC6803  (   1) MVKVIGRRSLGVQRIFDIGLPQDHNFLLANGAIAANC
S.PCC7002  (   1) MVKIIRRKFIGHAPTYDIGLSQDHNFLLGQGLIAANC
N.PCC7120  (   1) MIKIASRKFLGVENVYDIGVRRDHNFFIKNGLIASNC
Npu        (   1) MIKIATRKYLGKQNVYDIGVERDHNFALKNGFIASNC
Ter        (   1) MVKIVSRKLAKTENVYDIGVTKDHNFVLANGLIASNC
Tel        (   1)  MKIVGRRLMGWQAVYDIGLAADHNFVLANGAIAANC
Tvu        (   1)  MKIVGRRLVGWQAVYDIGLAGDHNFLLANGAIAANC
Aov        (   1) MVKITARKFVGRENVYDIGVEHHHNFAIKNGLIASNC
Aha        (   1) MVKIIKRQSLGRQNVYDIGVETDHNFVLANGCVASNC
Oli        (   1) MVKIVRRQSLGVQNVYDIGVEKDHNFLLASGEIASNC
                                 =============== =========
```

B

```
                Cyanobacteria

S.PCC6803       (774•37)   MVKFAEY•CFNKSHS
S.PCC7002       (775•37)   MVKFAEY•CFNKSHS
N.PCC7120       (775•37)   MLKFAEY•CFNKSHS
Npu             (774•37)   MLKFAEY•CFNKSHS
Tel             (755•36)   MLDFAEY•CFNKSHS
Tvu             ( ? •36)   MLDFAEY•CFNKSHS
Ter             ( ? •37)   MIKFAEY•CFNKSHS
Aov             ( ? •37)   MLNFAEY•CFNKSHS
Aha             ( ? •37)   MIKFAEY•CFNKSHS
Oli             ( ? •37)   MVKFAEY•CFNKSHS
Pma MIT9313     (757/758)  MVLFAEY CFNKSHS

                Bacteria
```

Fig. 4. Sequences of cyanobacteria DnaE split inteins.

A. DnaE1 (top) and DnaE2 (bottom) intein parts. The start position of each intein is shown in their host proteins, except for the DnaE1 sequences determined in this work where only the 3′ parts of the genes were determined. Asterisks indicate stop codons. Underlined regions are intein conserved sequence motifs (Pietrokovski, 1998). Lower case sequence regions in the C′ end of the DnaE1 intein parts are not conserved in these sequences and should not be considered aligned. Below the sequence alignment is a graphical representation of it (Henikoff *et al*., 1995).

B. DnaE split intein integration points. Known cyanobacterial N′ and C′ flanks of the intein parts are shown, from the DnaE1 and DnaE2 sequences, respectively, except for the *P. marinus* DnaE sequence that does not have an intein. Filled bullets indicate the presence of a split intein. Below the sequence alignment is a sequence logo of the corresponding position from 143 diverse bacterial DnaE sequences. Species abbreviations: S.PCC6803, *Synechocystis* species PCC 6803; S.PCC7002, *Synechococcus* species PCC 7002; N.PCC7120, *Nostoc* species PCC 7120; Npu, *Nostoc punctiforme*; Ter, *Trichodesmium erythraeum*; Tvu, *Therrmosynechococcus vulcanus*; Tel, *Therrmosynechococcus elongatus* BP-1; Aov, *Aphanizomenon ovalisporum*; Aha, *Aphanothece halophytica*; Oli, *Oscillatoria limnetica*; Pma MIT9313, *Prochlorococcus marinus* MIT9313.

clades. *Thermosynechococcus*, classified by its 16S rRNA gene as an early diverged genus of cyanobacteria (Honda *et al*., 1999), is found by the presence of a split DnaE intein to be part of a larger assembly of cyanobacterial groups. This could be investigated further by analysing the now available complete genome of *T. elongatus* BP-1 (Nakamura *et al*., 2002).

Split cyanobacterial *dnaE* genes are present in diverse loci, whereas all known cyanobacterial contiguous *dnaE* genes appear in stable genomic positions. The original genomic rearrangement that split the *dnaE* gene thus seems to have been followed by further genomic rearrangements. *Nostoc* species PCC 7120 and related cyanobacteria undergo developmentally regulated genome rearrangements (Golden *et al*., 1987; Carrasco and Golden, 1995). These are species in which split *dnaE* genes also occur. It is possible that some cyanobacteria are more tolerant to genome rearrangements. Such events produce split genes that are reactivated by regulated rearrangements. One rearrangement may have split

Table 1. Intein distribution in cyanobacteria.[a]

	Species						
Protein	Ssp Pcc6803	Nsp Pcc7120	Npu	Tel	Pma Mit9313	Pma Med4	Ssp Wh8102
DnaB helicase	+	+	+	–	–	–	–
DNA polymerase III alpha subunit	+	+	+	+	–	–	–
Unknown function, gene *rtcB*	\	–	+	\	\	\	\
Class II ribonucleotide reductase	–	+	–	–	–	–	–
DNA polymerase III tau subunit	+	–	–	–	–	–	–
DNA gyrase subunit B	+	–	–	–	–	–	–
Totals0	6	4	3	3	1	0	0

a. Species include fully and almost fully sequenced cyanobacteria. +, intein is present; –, intein is absent; \, species does not have the orthologue of the intein host protein.

a *dnaE* gene in its intein region. However, the two *dnaE* gene parts could reassemble at the protein level by their intein regions.

Consequences of continuous distribution of split DnaE inteins

There is no biological assay for intein activity. New intein genes can be identified by amplifying genes known to include inteins in some species from other related species (Davis *et al*., 1994; Fsihi *et al*., 1996; Sander *et al*., 1998; Saves *et al*., 2000; Lazarevic, 2001). However, all known inteins are randomly distributed, and only one case was found of closely related species with consistent intein distribution (Sander *et al*., 1998). Usually only a few of the examined species were found to contain inteins. DnaE split inteins seem to be an exception, being present in all cyanobacteria species that we examined that belong to certain well-defined groups. Other species from these groups can thus serve as a reliable source of split inteins.

The continuous distribution of DnaE split inteins in cyanobacterial groups has theoretical and applied implications for intein studies. We can now obtain inteins that co-evolve with their host species. These inteins can be used to study the evolutionary change rate of inteins, for example to decide whether inteins are of ancient or recent origin (Pietrokovski, 2001). They could also be a valuable source in theoretical and experimental studies of protein–protein interaction and intein catalytic activity (Gorbalenya, 1998; Martin *et al*., 2001). From a practical aspect, our ability to obtain an apparently very large number of different DnaE split inteins will be enhanced. For example, the Pasteur Culture Collection of Cyanobacteria (PCC) maintains about 475 axenic strains, many belonging to the groups in which we believe DnaE split inteins are fixed. Of particular interest will be split inteins from species living in extreme conditions such as high temperature or salinity. The protein-splicing activity of these split inteins may depend on the conditions in which their host species are active.

Origin and cause of split intein fixation

Our hypothesis is that split inteins are very difficult to lose. This is based on the idea that split inteins cannot be lost by one, or even two, simple DNA excision events. Cyanobacterial DnaE proteins are split by inteins in a highly conserved motif appearing in all known DnaE proteins (Fig. 4B). This strongly suggests that the two split DnaE host protein parts would have to be attached by a peptide bond at the intein integration point for the motif to adopt the fold, and have the activity, found in other DnaE proteins. Loss of one or both intein parts will leave the products of their gene flanks split with no mechanism to ligate. To lose the intein genes and retain a functional *dnaE* gene requires parallel precise loss of the intein parts followed directly by precise fusion of their flanks into one gene; this is a highly unlikely event.

An alternative method for removal of split intein genes is to acquire a surrogate gene to replace the product of the split genes. However, DnaE is an essential and tightly regulated gene forming the core of the bacterial replicative DNA polymerase. It interacts with four other protein subunits in the polymerase holoenzyme and with various cofactors during replication (Kelman and O'Donnell, 1995). Thus, a surrogate gene must be precisely adapted to the species to replace its DnaE gene.

Are split inteins common or unique?

There are more then 50 known intein alleles, but only the cyanobacterial *dnaE* intein allele is split. Nevertheless, these split inteins are common in cyanobacteria. If split inteins are common, at least in one group of species, why is there only one known type (allele) of them? The rarity of split intein organization probably results from the difficulty in switching to a functional two genes and two products state from a single gene and a single product one. Gene splitting would require acquiring a promoter and translation initiation signal for the downstream part of the split gene, adapting co-regulation for the two formed genes and evolving amino acid sequences to stabilize the two new ends of the intein parts. Gene splitting of the intein might also require adaptation for *trans*-splicing and high-affinity interaction between the two intein parts. Artificially created split inteins were found to have some affinity between their two parts and a low level of *trans*-splicing (Shingledecker *et al.*, 1998; Southworth *et al.*, 1998; Mootz and Muir, 2002). Thus, adapting the protein function of a newly split intein gene might be the minor difficulty compared with adapting proper transcription and translation control and protein stability. Once all these adaptations were complete, the unidirectional ratchet nature of the intein-splitting process would ensure the evolutionary persistence of the gene organization. However, still unknown are what initial circumstances led to the selection of the individual cyanobacterium in which the intein splitting occurred. Significant advantage/s likely to have accompanied this event to compensate for all necessary genetic adjustments are discussed above.

Experimental procedures

Analysed species

Aphanizomenon ovalisporum and *Microcystis* species genomic DNA was provided by A. Kaplan (Hebrew University, Jerusalem, Israel). *Aphanothece halophytica* and *Oscillatoria limnetica* genomic DNA was provided by E. Padan (Hebrew University, Jerusalem, Israel). *Therrmosynechococcus vulcanus* genomic DNA was provided by N. Adir (Technion, Haifa, Israel).

Data for previously sequenced cyanobacterial *dnaE* genes were obtained from the NCBI sequence databases (http://www.ncbi.nlm.nih.gov/Database) or Joint Genome Institute WWW site (http://www.jgi.doe.gov).

Sequence determination strategy and implementation

Sequence determination of the DnaE split inteins was done in two steps for each part of the DnaE gene. First, a conserved region flanking the N′ or C′ split intein part was amplified by degenerate primer PCRs. In some cases, a number of partially overlapping regions were amplified until the intein parts were approached. Next, linear (single primer) amplifications were done towards each split intein part using specific primers, designed from the previously amplified regions. The 3′ (intein) ends of the single-stranded amplification products were tailed by a homo-oligomer using terminal transferase. The tailed products were then PCR amplified by a primer complementary to the homo-oligomer tail and a second specific primer, nested to the first specific primer (Rudi *et al.*, 1999). This PCR amplification was repeated twice, and the products were cloned and sequenced or sequenced directly. To confirm the resulting sequence, the intein-containing region, together with its adjacent untranslated and *dnaE* flanks, was amplified from genomic DNA using specific primers and sequenced on both strands (Fig. 3).

PCR amplifications included ≈50 ng of genomic DNA, 25 pmol of each specific primer and 30–100 pmol of each degenerate primer, 2 nmol of dNTP mix, 2.5 units of *Taq* DNA polymerase (Sigma) and 5 μl of 10× *Taq* polymerase buffer (Sigma) in 50 μl reaction volumes. Linear amplifications were done in the same way except using ≈500 ng of genomic DNA and 12.5 pmol of one specific primer. Amplification schedules and temperatures are detailed in *Supplementary material* (Table S1). Primer sequences are listed in *Supplementary material* (Table S2).

The linear, single-stranded amplification products were purified using a PCR purification kit (Qiagen), as recommended by the manufacturer. Cytosine tailing was done in 20 μl reactions with 5 μl of the purified linear amplification products, 200 pmol of dCTP, 30 units of terminal deoxynucleotidyl transferase (TDT) (Fermentas) and 4 μl of 5× TDT buffer (Fermentas). Reaction mixtures were incubated at 37°C for 20 min and stopped by a 10 min incubation at 72°C.

PCR amplifications with the polyguanine primer (polyG) used 4 μl of the TDT reaction products as template with all the other components as listed above. Products were purified by a PCR purification kit (Qiagen) and sent directly for sequencing.

The confirmed sequence data have been submitted to the GenBank database under accession numbers AY209003–AY209008 and AY311409–AY311410.

Phylogenetic analyses

Phylogenetic analysis was done using programs from the PHYLIP package (Felsenstein, 1989), version 3.55. Trees were calculated using the SEQBOOT, PROTDIST and NEIGHBOR programs with 1000 bootstrap trials and default settings.

Acknowledgements

We thank Z. Kelman for technical advice and comments on the manuscript, A. Kaplan, G. Shalev-Alon, R. Schwarz, E. Padan, M. Bronstein and N. Adir for the gift of cyanobacterial DNA and genomic libraries, and the Joint Genome Institute and Penn State University for making preliminary sequence data publicly available. S. Pietrokovski holds the Ronson and Harris Career Development Chair.

Supplementary material

The following material is available from
http://www.blackwellpublishing.com/products/journals/suppmat/mmi/mmi3825/mmi3825sm.htm
Table S1. DNA amplification conditions.
Table S2. PCR primers.

References

Belfort, M., and Roberts, R.J. (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res* **25:** 3379–3388.

Belfort, M., Reaban, M.E., Coetzee, T., and Dalgaard, J.Z. (1995) Prokaryotic introns and inteins: a panoply of form and function. *J Bacteriol* **177:** 3897–3903.

Carrasco, C.D., and Golden, J.W. (1995) Two heterocyst-specific DNA rearrangements of nif operons in *Anabaena cylindrica* and *Nostoc* sp. strain Mac. *Microbiology* **141:** 2479–2487.

Davis, E.O., Thangaraj, H.S., Brooks, P.C., and Colston, M.J. (1994) Evidence of selection for protein introns in the recAs of pathogenic mycobacteria. *EMBO J* **13:** 699–703.

Derbyshire, V., and Belfort, M. (1998) Lightning strikes twice – intron-intein coincidence. *Proc Natl Acad Sci USA* **95:** 1356–1357.

Duan, X., Gimble, F.S., and Quiocho, F.A. (1997) Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell* **89:** 555–564.

Felsenstein, J. (1989) PHYLIP – phylogeny inference package, version 3.2. *Cladistics* **5:** 164–166.

Fsihi, H., Vincent, V., and Cole, S.T. (1996) Homing events in the gyrA gene of some mycobacteria. *Proc Natl Acad Sci USA* **93:** 3410–3415.

Ghosh, I., Sun, L., and Xu, M.Q. (2001) Zinc inhibition of protein trans-splicing and identification of regions essential for splicing and association of a split intein. *J Biol Chem* **276:** 24051–24058.

Gimble, F.S., and Thorner, J. (1992) Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. *Nature* **357:** 301–306.

Golden, J.W., Mulligan, M.E., and Haselkorn, R. (1987) Different recombination site specificity of two developmentally regulated genome rearrangements. *Nature* **327:** 526–529.

Gorbalenya, A.E. (1998) Non-canonical inteins. *Nucleic Acids Res* **26:** 1741–1748.

Henikoff, S., Henikoff, J.G., Alford, W.J., and Pietrokovski, S. (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* **163:** 17–26.

Hess, W.R., Rocap, G., Ting, C.S., Larimer, F., Stilwagen, S., Lamerdin, J., and Chisholm, S.W. (2001) The photosynthetic apparatus of *Prochlorococcus*: insights through comparative genomics. *Photosynth Res* **70:** 53–71.

Honda, D., Yokota, A., and Sugiyama, J. (1999) Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine Synechococcus strains. *J Mol Evol* **48:** 723–739.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., *et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3:** 109–136.

Kaneko, T., Nakamura, Y., Wolk, C.P., Kuritz, T., Sasamoto, S., Watanabe, A., *et al.* (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* **8:** 205–213.

Kelman, Z., and O'Donnell, M. (1995) DNA polymerase III holoenzyme: structure and function of a chromosomal replicating machine. *Annu Rev Biochem* **64:** 171–200.

Lazarevic, V. (2001) Ribonucleotide reductase genes of *Bacillus* prophages: a refuge to introns and intein coding sequences. *Nucleic Acids Res* **29:** 3212–3218.

Liu, X.Q. (2000) Protein-splicing intein: genetic mobility, origin, and evolution. *Annu Rev Genet* **34:** 61–76.

Martin, D.D., Xu, M.Q., and Evans, T.C., Jr (2001) Characterization of a naturally occurring trans-splicing intein from *Synechocystis* sp. PCC6803. *Biochemistry* **40:** 1393–1402.

Meeks, J.C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., *et al.* (2001) An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. *Photosynth Res* **70:** 85–106.

Mootz, H.D., and Muir, T.W. (2002) Protein splicing triggered by a small molecule. *J Am Chem Soc* **124:** 9044–9045.

Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., *et al.* (2002) Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res* **9:** 123–130.

Ozawa, T., Kaihara, A., Sato, M., Tachihara, K., and Umezawa, Y. (2001) Split luciferase as an optical probe for detecting protein–protein interactions in mammalian cells based on protein splicing. *Anal Chem* **73:** 2516–2521.

Paulus, H. (2000) Protein splicing and related forms of protein autoprocessing. *Annu Rev Biochem* **69:** 447–496.

Perler, F.B. (1999) A natural example of protein trans-splicing. *Trends Biochem Sci* **24:** 209–211.

Perler, F.B. (2002) InBase: the intein database. *Nucleic Acids Res* **30:** 383–384.

Perler, F.B., and Adam, E. (2000) Protein splicing and its applications. *Curr Opin Biotechnol* **11:** 377–383.

Pietrokovski, S. (1998) Modular organization of inteins and C-terminal autocatalytic domains. *Protein Sci* **7:** 64–71.

Pietrokovski, S. (2001) Intein spread and extinction in evolution. *Trends Genet* **17:** 465–472.

Rudi, K., Fossheim, T., and Jakobsen, K.S. (1999) Restriction cutting independent method for cloning genomic DNA segments outside the boundaries of known sequences. *Biotechniques* **27:** 1170–1177.

Sander, P., Alcaide, F., Richter, I., Frischkorn, K., Tortoli, E., Springer, B., *et al.* (1998) Inteins in mycobacterial GyrA are a taxonomic character. *Microbiology* **144:** 589–591.

Saves, I., Laneelle, M.A., Daffe, M., and Masson, J.M. (2000) Inteins invading mycobacterial RecA proteins. *FEBS Lett* **480:** 221–225.

Shingledecker, K., Jiang, S.-Q., and 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:** 187–195.

Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., *et al.*. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* strain ΔH: functional analysis and comparative genomics. *J Bacteriol* **179:** 7135–7155.

Southworth, M.W., Adam, E., Panne, D., Byer, R., Kautz, R., and Perler, F.B. (1998) Control of protein splicing by intein fragment reassembly. *EMBO J* **17:** 918–926.

Vardi, A., Schatz, D., Beeri, K., Motro, U., Sukenik, A., Levine, A., and Kaplan, A. (2002) Dinoflagellate-cyanobacterium communication may determine the composition of phytoplankton assemblage in a mesotrophic lake. *Curr Biol* **12:** 1767–1772.

Wu, H., Hu, Z., and Liu, X.Q. (1998) Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis sp.* PCC6803. *Proc Natl Acad Sci USA* **95:** 9226–9231.

Yu, Z., Li, T., Zhao, J., and Luo, J. (2002) PGAAS: a prokaryotic genome assembly assistant system. *Bioinformatics* **18:** 661–665.