Non-canonical inteins
Alexander E. Gorbalenya*

M. P. Chumakov Institute of Poliomyelitis and Viral Encephalitides, Russian Academy of Medical Sciences, 142782 Moscow Region, Russia and Department of Virology, Leiden University Medical Center, AZL P4-22, PO BOX 9600, 2300 RC Leiden, The Netherlands

Received November 25, 1997; Revised and Accepted February 6, 1998

ABSTRACT

Previous analyses have shown that inteins (protein splicing elements) employ two structural organizations: the ‘canonical’ Nintein-Dod-inteinC found in dozens of inteins and a ‘non-canonical’ Nintein-inteinC described in two inteins, where Nintein at the N-terminal and inteinC at the C-terminal are conserved domains involved in self-splicing and Dod is the Dod DNA endonuclease (DNase). In this study, four non-canonical inteins, each with unique structural features, have been identified using alignment-based Hidden Markov Models. A Nintein-inteinC intein, carrying an unprecedented replacement of the N-terminal catalytic Cys(Ser) by Ala, is described in a putative ATPase encoded by Methanococcus jannaschii. Three replicative proteins of Synnechocystis spp. contain inteins with the organizations: (i) Nintein – X – inteinC\textsuperscript{C}, where X is an uncharacterized domain and Dod DNase is located in an alternative open reading frame (ORF) being embedded between two novel CG and YK domains; (ii) Nintein-HN-inteinC, where HN stands for phase-like DNase from the EX\textsubscript{1}H-HX\textsubscript{3}H family; (iii) Nintein>|inteinC, where |> indicates that the intein domains are associated with a disrupted host protein encoded by two spatially separated ORFs. The expression of some of these newly identified inteins may affect the intein hosts. The variety of structural forms of inteins could have evolved through invasion of self-splicing proteases by different mobile DNases or the departure of mobile DNases from canonical inteins.

INTRODUCTION

Many genes are known to be interrupted with alien sequence(s), which are removed from gene products with concomitant religation of flanking sequences in the process called splicing. Intervening sequences removed during RNA and protein splicing are called introns and inteins, respectively; religated sequences are named exons or exteins, respectively (1,2). A number of parallels between inteins and introns, respectively; religated sequences removed during RNA and protein splicing are called inteins and are often (introns) autocatalytically (4,8,9). Certain genes encoding inteins and self-splicing inteins can move into an intein/intron-less allele of the host gene, a process called ‘intron/intein homing’ (6,10,11). The homing event is initiated by a site-specific DNA endonuclease (DNase) encoded by a self-splicing intron or associated with an intein (11,12), and some DNase genes were shown to be bona fide mobile elements invading intergenic regions and DNase-less copies of introns (13–17). Three large families of mobile DNases were identified which are associated with self-splicing introns and which are named after the conserved sequences LAGLI-DADG (Dod) (18–20), GIY-YIG (21) and EX\textsubscript{1}H-HX\textsubscript{3}H or H-N-H (22,23). Proteins belonging to these families are also encoded in a variable genetic context, although only the Dod-type DNases were so far identified in inteins. Eight conserved sequence blocks or motifs were recognized in all but two inteins, and two blocks (C and E or Dod1 and Dod2) are the hallmarks of the Dod protein family (24,25). (Hereafter, inteins with a full complement of the intein blocks are called ‘canonical’). Two non-canonical (‘minimal’) inteins lacking a region encompassing blocks C, D, E and H were also described (24,25) as well as similarities between inteins and the yeast mate-switching HO DNase (25,26) and the autoproteolytic domain (Hh-C) of signalling hedgehog (Hh) proteins (24,27) were documented.

intein-promoted protein splicing and gene mobility can functionally be separated (28). The N-terminal residue of inteins, which is always either Cys (Cys-inteins) or Ser (Ser-inteins) and part of block A, was shown to initiate splicing at the N-terminal border of inteins. Another nuclease, the N-terminal residue of the C-extein, which is always Ser, Cys or Thr in cooperation with the strictly conserved C-terminal Asn (block G) residue of the intein, mediate subsequent steps leading to intein excision and religation of inteins (29–32). Two conserved His residues in blocks B and G were implicated in protein autoprocessing as well (25,30,32) and they, together with the other catalytic residues, are spatially juxtaposed (33). The blocks C and E are not important for splicing but crucial for intein mobility (28,34). The recent X-ray structure of Saccharomyces cerevisiae VMA (PI-Scel) intein revealed that the Dod DNase blocks are organized with a separate structural domain which is looped out from the rest of the intein body forming a continuous β-sheet (33).

Previous studies have shown comparative sequence analysis to be the most powerful tool for the identification of inteins (24,25,35–38). In this study, four new non-canonical (putative) inteins, each with unique sequence features, were identified using Hidden Markov Models (HMMs) (39,40) trained on multiple sequence alignments of previously identified inteins. These findings are rationalized within a model describing inteins as molecular ensembles of proteolytic and, optionally, DNase domains of independent origin which may have diverse expression mechanisms affecting host proteins.

*All correspondence should be addressed to present address: Frederick Biomedical Supercomputer Center, SAIC/NCI-FCRDC, PO Box B, Building 430, Frederick, MD 21702-1201, USA. Tel: +1 301 846 1991; Fax: +1 301 846 5762; Email: gorbalen@ncifcrf.gov
RESULTS AND DISCUSSION

A M.jannaschii ATPase encodes a minimal intein which may be defective in splicing

The M.jannaschii genome was reported to encode 14 proteins containing 18 inteins (24,36). This analysis (see legend to Fig. 1) has identified an additional intein in M.jannaschii inserting after Gly404 within an MJ0781 ORF reported previously to be similar to a putative ATPase encoded by the plasmid RK2 kba gene (M.j jKlbA) (36). The intein-less version of M.j KlbA has been found to be the most closely related to five uncharacterized proteins other than the RK2 kba gene product (Fig. 2). Four of these proteins are of archael origin encoded by M.jannaschii, Methanococcus voltae and Sulfolobus solfataricus, while the fifth originates from green sulfur bacteria Chlorobium limicola. Like other inteins (24), the M.j KlbA intein disrupts a conserved domain (Fig. 2).

The M.j KlbA intein is small (168 aa) being the third ‘minimal’ Dom-less intein identified so far, the two others being Ppu DnaB and Mse GyrA inteins (24,25). An Asn/Ser splicing junction was identified at the intein/C-extein border but Ala replaces the catalytic Cys(Ser) nucleophile at the intein N-terminus (Fig. 1), the latter being a unique feature found only in the KlbA intein. The N-terminal Ala can not catalyze the N-terminal proteolytic cleavage (28,32) implying that the KlbA precursor is not likely to be spliced. The substitution of the catalytic Cyst(ser) by Ala might have been accepted in the course of co-evolution of the KlbA intein and its host to prevent intein excision after the host became dependent on the intein. The other explanations include: (i) trans-splicing of the KlbA intein with involvement of another intein; (ii) incorrect sequencing of the first codon of the KlbA intein.

Three new non-canonical inteins in Synechocystis

Three inteins with very unusual organizations were identified in the Synechocystis genome (55) in the Spi ORFs, sll1360 (DNA-polymerase IIIp1, Pol-IIIp1), sll2005 (DNA-gyrase B, GyrB) and srl0603 (DNA-polymerase IIIp1, Pol-IIIp1). The Synechocystis proteins hosting inteins are predicted to be components of the replicative apparatus and the insertions were found within the conserved areas after Gly129 (sll1360), Gly436 (sll2005) and Tyr774 (srl0603).

Three new inteins start with Cys. Nintein and inteinC domains as well as the C-terminal Asn/Cys(Ser) splice junction have been recognized in the sll1360 and sll2005 ORFs, although only a variant of the Nintein was identified in the srl0603 ORF (Fig. 1). A Dom domain was not identified in either of these ORFs by different methods (not shown), despite the Nintein and inteinC domains of sll1360 and sll2005 being connected by sequences long enough to encompass an additional domain (Fig. 1).

DNA-gyrase B subunit intein contains a DNase domain of the EX1H-HX3H family

Upon analysis of the GyrB sll2005 intein, a marginal similarity was initially detected between the middle part of this protein and an intron-encoded DNase from phage RB3, a variant of phage T4 (56). This DNase belongs to a subset (called hereafter HN) of the EX1H-HX3H DNase family (22,23; A.Gorbalenya, unpublished...
been involved in the transfer of the HN domain into the GyrB domain being identified in an intein. A phage might have
of inteins (Fig. 3 A and the GyrB scheme in Fig. 3 B). Although
Nintein, at a position occupied by the Dod domain in the majority
results) and subsequent analysis confirmed the region of the GyrB sll2005 intein to be the core domain conserved in viral HN proteins (Fig. 3A). In the GyrB sll2005, the HN domain is located not far downstream from the predicted C-terminal border of the Nintein, at a position occupied by the Dod domain in the majority of inteins (Fig. 3A and the GyrB scheme in Fig. 3B). Although a parallel has already been drawn between DNases from the EX1 H-HX 3 H and Dod families (22,23), this is the first case of an HN domain being identified in an intein. A phage might have been involved in the transfer of the HN domain into the GyrB

Figure 1. Nintein, Dod and intenC domains in the Sce HN DNase and 26 Hr proteins; (Dod), 68 Dod proteins including 40 intein-associated and 28 of non-intein origin; (inteinC), 46 inteinCs and four related domains of intronic origin (unpublished). The alignment is limited to eight conserved blocks in five protein groups separated by blank lines and including, from top to bottom: (i) a set of canonical inteins (Psp Pol-1, Mja TFIB and RpoL_A” and Mle RecA); (ii) two known minimal inteins (Mxe GyrA and Ppu DnaB); (iii) new inteins [Mja KlbA, Ssp GyrB, Pol-III”/ and Ppo GyrB] and Pol-III[Hx (sll1603 and sll1360x ORFs) and Pol-III[Hx (sll1603 and sll1572 ORFs); numbering according to the Cyanobase (55)]; (iv) two intronic Dods containing intenC domain (Pnn ND53 and Pnn COI2) along with new free-standing Dod-proteins from M.jannaschii [MJ1098, MJ0314, MJ0398 and U67501x, the latter is a new composite ORF containing two frameshifts and a termination codon and is encoded by nucleotides 69–729 in the complementary strand of the U67501 entry; sequencing numbers are shown at the right. GB, GenBank; TE, Trembl; Psp, unknown function; DnaB, DnaB replicative helicase; RpoL_A” , RNA polymerase subunit A”; TFIB, transcription factor IIB; COL, cytochrome oxidase; NADH, NADH ubiquinone reductase subunit b; M75 and HH, hedgehog proteins. Cel, Caenorhabditis elegans; Dme, Drosophila melanogaster; Sce, S.cerevisiae; Sp, Synchocystis sp.; Tli, Thermococcus littoralis. Colors: magenta, invariant residues; red and green, conserved and similar residues, respectively, in >55% positions in the original all-inclusive alignments. Groups of similar residues: &, I,L,V, M,F,Y,W; !, W,Y,F; $, D,E,N,Q; @, K,R,H; #, T,S,C; %, A,G. X, frameshift. The position of the
in the intein ORF sl1360 by blocks B and F (the Pol-III γ scheme in Fig. 3 B and see below). It was found that the Dod (Fig. 1). In line with this observation, the Dod domains of the minor other inteins (named minor and major groups, respectively) to together and separately from dozens of homologs associated with γ-differ strikingly (Fig. 3 C]). Each of the two groups also includes in the Dod2, only positions 156 (Leu dominancy) and 164 (Gly and major groups are different in many positions of blocks Dod1 and 34% identical residues belong to the active-site forming amino acids and include among others a CECGCCG run. In the sl1360 ORF, an almost perfect copy (69–74 aa) (53) of this hexapeptide flanks the CG domain from the C-terminus, indicating that the region upstream of the Dod domain may be composed of the two distantly related CG domains (the sl1360x scheme in Fig. 3 B). In the Ssp Pol-III γ, the conserved N-terminus of the first CG domain of the sl1360x ORF overlaps with the C-terminal border of the sl1360 Nintein and the conserved C-terminus of the YK domain overlaps with the N-terminus of the sl1360 inteinC (the Pol-III γ scheme along with the sl1360 and sl1360x ORFs translations in Fig. 3 B). This overlapping organization indicates a complex expression mechanism of the ORFs representing the sl1360x/sll1360 locus. The two ORFs might be expressed essentially independently, albeit in a coordinated manner, from two initiator AUG codons in different frames using one or two mRNAs. Alternatively, the "canonical" three-domain intein might be produced from the sl1360x/sll1360x mRNA by a ribosome slippage mechanism (58) which would allow two frameshifts, back/forward (~1/+1), with reading of the sl1360x domains in-frame and between the Nintein and inteinC domains (the Pol-III γ scheme in Fig. 3 B). Either of these mechanisms would host affect protein expression. [Very recently, splicing of the Ssp Pol-III γ intein has been demonstrated in Escherichia coli (58)].

Pol-III γ/τ Dod is flanked by two novel CG and YK domains within the ORF

The sl1360x Dod domain is flanked by additional domains on either side (the Pol-III/τ sl1360x scheme in Fig. 3 B). At the C-terminus, immediately downstream of the DNase domain, a small unique domain of ~35 aa (called YK domain after the two most conserved residues) is located. It is conserved in the proteins of the minor group (Fig. 3 D) indicating that the Dod and YK domains have co-evolved in these proteins.

Unlike the YK domain, a unique 119 aa sequence upstream of the sl1360x Dod domain (the sl1360x scheme in Fig. 3 B) is not conserved in the proteins of the minor group implying that it might have become associated with the DNase relatively recently. The N-terminal region of the sl1360x ORF product includes a Cys-rich 50 aa domain (named CG after the conserved residues) whose homolog was also detected in the HSPDCM4x ORF of the archael Halobacteria species sulfonating phage HSPD (57), the C-terminal half of the alignment placed between the Gyrb and sl1360x schemes in Fig. 3 B). In the two sequences, the majority of the 34% identical residues belong to the active-site forming amino acids and include among others a CECGCCG run. In the sl1360x ORF, an almost perfect copy (69–74 aa) (53) of this hexapeptide flanks the CG domain from the C-terminus, indicating that the region upstream of the Dod domain may be composed of the two distantly related CG domains (the sl1360x scheme in Fig. 3 B). Either of these mechanisms would host affect protein expression. [Very recently, splicing of the Ssp Pol-III γ intein has been demonstrated in Escherichia coli (58)].

Unrelated Pol-III γ/τ Gyrb DNase-containing ORFs are "linked" by a phage ORF through CG and HN domains

Not only the CG domain unites the sl1360x and HSPDCM4x ORFs, they also both include a DNase domain, although belonging to the two different families, Dod and HN, respectively. Within the HSPDCM4x ORF, the HN and CG domains are separated by only a short peptide with loop-forming potential (SDDLRSPEGPNP) (HSPDCM4x sequence beneath the Gyrb scheme in Fig. 3 B). Remarkably, the HN domain of the HSPDCM4x ORF revealed a strong sequence similarity to that of the Gyrb sl2205 intein. Thus, the phage ORF, through the HN and CG domains, "links" two unrelated Synechocystis DNase-containing ORFs (the Gyrb sl2205 and Pol-III/τ sl1360x schemes and the alignment between them in Fig. 3 B). The CG domain may cooperate with different DNase domains given the
Synechocystis DNA-gyrase B subunit and DNA-polymerase III γ/τ subunit inteins: domain organization and relatives. For designations see Figures 1 and 2 legends unless otherwise specified. (A) Alignment of the DNase domain of Ssp GyrB intein and a selected set of the viral HN DNases. An HMM trained on the depicted alignment of 16 viral HN proteins (22, 23; A. Gorbalenya, unpublished results) scored the sll2005 ORF (17.89 bits) on top of the other proteins, excluding the training set, in the Genpeptides-100. The coordinates of sequences within the proteins are listed at the right. sp82i, spO1i and phi-Ei: DNA-polymerase intron-encoded DNases from Bacillus subtilis bacteriophages SP82, spO1 and phi-E, respectively; spp1: B. subtilis phage SPP1 ORF36.1; LL-rlt: Lactococcus lactis bacteriophage rlt ORF41; LL-Hi: Lactobacillus lactis bacteriophage LL-H intron-encoded ORF168; RBNBi: E. coli phage RB3 intron-encoded T-even endonuclease III; Y38_BPT7: E. coli phage T7 ORF38; T5Ytr1 and T5Ytr2: E. coli phage T5 ORFs in the tRNA gene cluster; A87R, A354R and A422R: Paramecium bursaria Chlorella virus 1 ORFs. The latter three ORFs contain two copies of the HN domain (indicated with suffices n and c). Cons. family, residues conserved in the EX 1 H-HX 3 H family (unpublished). (B) Domain organization and sequence features of Ssp GyrB and Pol-III γ/τ inteins and an archael phage ORF. Two schemes depicting the intein/extein organization of Ssp GyrB and Pol-III γ/τ are shown. Between the schemes, alignments of HN domains (left; the EX 1 H-HX 3 H family residues: +, conserved and *, invariant) and CG domains (right; bar, the Cys-rich box) are given. Bold, residues conserved in any two sequences. The coordinates of sequences within the ORFs are indicated at the left. Ssp Pol-IIIγτ sll1360 and GyrB sll2005 matched a non-documented 202 aa ORF (named HSPDCM4x) encoded by the 3096–3701 nucleotides within the dcm4 locus (X80164) of the archael Halo bacterium salinarium phage εH. A TblastN-mediated search using the 119 aa sequence upstream of the sll1360x Dod reported a top match (P = 0.00013) between the 14–63 aa stretch and a region (105–154 aa) of the HSPDCM4x (CG domain). A reciprocal TblastN-mediated search confirmed this finding (top match, P = 0.0064). The sll2005 HN domain was among three top matches (P = 0.084) hit by the HSPDCM4x ORF using BlastP. No assignment was produced for the GyrB regions flanking the HN domain and enclosed between the Nintein and inteinC, and for the Pol-IIIγτ sll1360 region embedded between the Nintein and inteinC (not shown). Beneath the Ssp Pol-IIIγτ scheme, two regions of the Synechocystis genome (D90907) along with translation from the sll1360 and sll1360x ORFs are depicted. Italic bold, sequences of the Nintein and inteinC domains, and those conserved in the CG and YK domains (D) in colours to match the drawings. (C) Logos of the Dod1 (block C) and Dod2 (block E) of the minor and major groups of Dod proteins. The logos (S4) were produced using ClustalW-generated alignments (not shown) of the minor group (top line) and major group (bottom line) of the Dod domains. One standard deviation of the information content at each position is indicated. Colours: black, A,L,I,V,M; blue, F,Y,W; violate, K,R,H; red, D,E; yellow, N,Q; green, G; pink, P. (D) Multiple alignment of the YK domains. These domains were identified as part of the ChlastW-generated alignment of proteins of the Dod minor group (Fig. 1 and unpublished). Three conserved blocks of the YK domain were verified with the MACAW which guided a slight adjustment of the alignment within the least significant block 1 (P = 1.9e-8 using a searching space delimited by the Dod block H and the YK block 2). The distances to block H of the Dod domain (left) and to the end of the protein or, for the Mle RecA YK domain, to the inteinC domain (right) are indicated. Ama, Allomyces macrogynus; Ceu, Chlamydomonas eugametos; Cpa, Chlamydomonas pallidostigmatica; Csm, Chlamydomonas smithii; Chu, Chlamydomonas humicola; Kla, Kluyveromyces lactis; Kth, Kluyveromyces thermodenatus; Pca, Pichia canadensis; COR, apocytochrome b; Lar and Sor, large and small subunit ribosomal RNAs.

Figure 3. (Above and overleaf). Synechocystis DNA-gyrase B subunit and DNA-polymerase III γ/τ subunit inteins: domain organization and relatives. For designations see Figures 1 and 2 legends unless otherwise specified. (A) Alignment of the DNase domain of Ssp GyrB intein and a selected set of the viral HN DNases. An HMM trained on the depicted alignment of 16 viral HN proteins (22, 23; A. Gorbalenya, unpublished results) scored the sll2005 ORF (17.89 bits) on top of the other proteins, excluding the training set, in the Genpeptides-100. The coordinates of sequences within the proteins are listed at the right. sp82i, spO1i and phi-Ei: DNA-polymerase intron-encoded DNases from Bacillus subtilis bacteriophages SP82, spO1 and phi-E, respectively; spp1: B. subtilis phage SPP1 ORF36.1; LL-rlt: Lactococcus lactis bacteriophage rlt ORF41; LL-Hi: Lactobacillus lactis bacteriophage LL-H intron-encoded ORF168; RBNBi: E. coli phage RB3 intron-encoded T-even endonuclease III; Y38_BPT7: E. coli phage T7 ORF38; T5Ytr1 and T5Ytr2: E. coli phage T5 ORFs in the tRNA gene cluster; A87R, A354R and A422R: Paramecium bursaria Chlorella virus 1 ORFs. The latter three ORFs contain two copies of the HN domain (indicated with suffices n and c). Cons. family, residues conserved in the EX 1 H-HX 3 H family (unpublished). (B) Domain organization and sequence features of Ssp GyrB and Pol-III γ/τ inteins and an archael phage ORF. Two schemes depicting the intein/extein organization of Ssp GyrB and Pol-IIIγτ are shown. Between the schemes, alignments of HN domains (left; the EX 1 H-HX 3 H family residues: +, conserved and *, invariant) and CG domains (right; bar, the Cys-rich box) are given. Bold, residues conserved in any two sequences. The coordinates of sequences within the ORFs are indicated at the left. Ssp Pol-IIIγτ sll1360 and GyrB sll2005 matched a non-documented 202 aa ORF (named HSPDCM4x) encoded by the 3096–3701 nucleotides within the dcm4 locus (X80164) of the archael Halo bacterium salinarium phage εH. A TblastN-mediated search using the 119 aa sequence upstream of the sll1360x Dod reported a top match (P = 0.00013) between the 14–63 aa stretch and a region (105–154 aa) of the HSPDCM4x (CG domain). A reciprocal TblastN-mediated search confirmed this finding (top match, P = 0.0064). The sll2005 HN domain was among three top matches (P = 0.084) hit by the HSPDCM4x ORF using BlastP. No assignment was produced for the GyrB regions flanking the HN domain and enclosed between the Nintein and inteinC, and for the Pol-IIIγτ sll1360 region embedded between the Nintein and inteinC (not shown). Beneath the Ssp Pol-IIIγτ scheme, two regions of the Synechocystis genome (D90907) along with translation from the sll1360 and sll1360x ORFs are depicted. Italic bold, sequences of the Nintein and inteinC domains, and those conserved in the CG and YK domains (D) in colours to match the drawings. (C) Logos of the Dod1 (block C) and Dod2 (block E) of the minor and major groups of Dod proteins. The logos (S4) were produced using ClustalW-generated alignments (not shown) of the minor group (top line) and major group (bottom line) of the Dod domains. One standard deviation of the information content at each position is indicated. Colours: black, A,L,I,V,M; blue, F,Y,W; violate, K,R,H; red, D,E; yellow, N,Q; green, G; pink, P. (D) Multiple alignment of the YK domains. These domains were identified as part of the ClustalW-generated alignment of proteins of the Dod minor group (Fig. 1 and unpublished). Three conserved blocks of the YK domain were verified with the MACAW which guided a slight adjustment of the alignment within the least significant block 1 (P = 1.9e-8 using a searching space delimited by the Dod block H and the YK block 2). The distances to block H of the Dod domain (left) and to the end of the protein or, for the Mle RecA YK domain, to the inteinC domain (right) are indicated. Ama, Allomyces macrogynus; Ceu, Chlamydomonas eugametos; Cpa, Chlamydomonas pallidostigmatica; Csm, Chlamydomonas smithii; Chu, Chlamydomonas humicola; Kla, Kluyveromyces lactis; Kth, Kluyveromyces thermodenatus; Pca, Pichia canadensis; COR, apocytochrome b; Lar and Sor, large and small subunit ribosomal RNAs.
structural organizations of Pol-III\(\alpha\) and HSPDCM4x ORFs. Some of the conserved Cys/His/Glu residues of CG domain might be involved in the metal-mediated binding of other protein(s) or DNA/RNA and, through this activity, the CG could assist the DNase. The CG and DNase domains might cooperate in cis being expressed on the same protein, e.g. the HSPDCM4x or the Pol-III\(\gamma\) sll1360x ORF products, or in trans, when expressed on different proteins of the same organism, e.g. the sll2005 and sll1360x intein-associated products in *Synechocystis*.

**DNA-polymerase III\(\alpha\) Nintein and inteinC are associated with the disrupted host protein encoded by two spatially separated ORFs**

The Nintein domain identified within the Pol-III\(\alpha\) sll0603 ORF consists of ~95 aa residues (Fig. 1) and comprises the largest part of the 123 aa C-terminal sequence which is fused with the 774 aa N-extein but, surprisingly (24), not flanked by the C-extein (Fig. 4). Instead, a putative C-extein part of Pol-III\(\alpha\) is encoded within the 479 aa sll1572 ORF which is located far distant from sll0603 in the *Synechocystis* genome (55). The 1–774 aa portion of the sll0603 protein combined with the 58–479 aa portion of the sll1572 are perfectly aligned with Pol-III\(\alpha\) of other origins (Fig. 4 and data not shown). Accordingly, the N-terminal portion of the sll1572 (1–57 aa) has features found in inteinCs including the Asn/Cys splice junction as well as residues conserved in blocks F and G although only a marginal score was given to it by an HMM trained on the other inteins (Figs 1 and 4). Collectively, the above results strongly suggest that the DNase-less intein is associated with the *Ssp* Pol-III\(\alpha\) and consists of the Nintein and inteinC domains encoded by two spatially separated ORFs.

Assuming that an active form of *Ssp* Pol-III\(\alpha\) includes domains present in the N-extein and C-extein, the active and essential-for-replication enzyme must reconstitute from the expressed parts. If the reconstitution proceeds with the religation of the exteins, this would imply that the disrupted intein is splicing-competent and involved in the process. Such a role in the extein expression would secure intein survival. Alternatively, the *Ssp* Pol-III\(\gamma\) might function as a non-covalently bound complex of the exteins with the intein being splicing-defective, although possibly proteolytically active and involved in formation of the complex.

Although it is presently unclear how the *Ssp* Pol-III\(\alpha\) gene became disrupted, an intein DNase might have been involved. The sll1572/slr0603 intein may have originally consisted of the canonical three domains and resided within a non-disrupted Pol-III\(\alpha\) encoded by a *Synechocystis* ancestor. Subsequently, intein DNase might have initiated a relocation of either the N-extein or the C-extein along with a part of the intein. In line with this hypothesis, homologs of one of the possible derivatives of such the event, a Dod-inteinC combination, were recognized in proteins encoded by self-splicing introns (Fig. 1, *Pan* ND5i1 and COH1). However, in the *Ssp* Pol-III\(\alpha\), no vestiges of a DNase domain were found in the vicinity of either sll0603 Nintein or sll1572 inteinC domains (not shown) that would be compatible with a scenario in which the DNase departed for a new destination after splitting the gene.

**CONCLUSION**

With new non-canonical inteins described in this paper, the number of known intein organizations has expanded from two to five and evidence has been presented that, in addition to the Dod proteins, the DNase of another family (HN) can also be an integral part of an intein. The variety of structural forms of inteins could have evolved through invasion of self-splicing proteases by different mobile DNases or the departure of mobile DNases from canonical inteins. These DNases could also infect self-splicing introns and intergenic regions in cellular and viral genomes [see also (6,33)].
An intein lacking DNase activity is bound to be eliminated unless the host benefits from its presence [the loss hypothesis (24)]. Since four currently known DNase-free inteins [two reported previously (24,25) and two described in this paper] are in the minority among inteins, they may represent only a fraction of the original set of DNase-free inteins that have survived by providing a selective advantage to their hosts. The presence of the N-terminal Ala in the MjøKlbA intein and the existence of the disrupted intein associated with Ssp Pol-IIIα can be rationalized within the framework of this model. Inteins may assist their hosts in different ways, for instance, through participating in controlling host gene expression, as was discussed above for the vitally important Pol-IIIα and Pol-IIIαγ of Synechocystis. The cell viability may therefore be under the control of inteins that would depend on one another to survive and might interact to improve fitness.

NOTE ADDED IN PROOF

During reviewing of the original and modified versions of this paper I have become aware that some of the findings described above have most recently been reported also by others (62–64).

ACKNOWLEDGEMENTS

My thanks to Willy Span for encouragement and facilities, Gerard Canters for office space, Tom Schneider and Dave Gillespie for helping to install the logo package, Willem Luytjes and Fred Wassenaar for assisting with figures and Caroline Brown for reading and correcting this manuscript. This work was supported by the Netherlands Organization for Scientific Research (NWO) and the Russian Fund for Basic Research (RFBR).

REFERENCES