Nucleic Acids Research, 1998, Vol. 26, No. 7 **1741–1748**

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-terminus and inteinC at the C-terminus 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 *Synechocystis* spp. contain inteins with the organizations: (i) $\frac{\text{Nintein} - \text{X} - \text{inteinC}}{\text{Dod}}$, 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 phage-like DNase from the EX_1H-HX_3H 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 a subset of introns, known as self-splicing introns, were noticed (3–7). Splicing of these molecules proceeds exclusively (inteins) or often (introns) autocatalytically (4,8,9). Certain genes encoding inteins and self-splicing introns 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_1H-HX_3H 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 nucleophile, 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-*Sce*I) 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

METHODS, DATABASES AND APPROACH

Global multiple amino acid sequence alignments were generated with the assistance of the ClustalW(1.5) package using different parameter settings (41). Blocks conserved in some multiple alignments were verified and assessed using the MACAW suit (42). Some multiple alignments were used as templates for building larger alignments in the constrained simulated annealing regime within the HMMER(1.8) package (43). This package was used to generate HMMs (39,40) trained on multiple sequence alignments employing the Blossum50 scoring table (44,45) as prior and the maximum discrimination method (46). HMMs were used to search Swissprot-34, SP-Trembl-2, Genpeptide-100 or homemade protein databases with the programs employing the Smith–Waterman (47) and Needleman–Wunsch (48) algorithms. Similarity of a sequence with a model was assessed in bites of information (bits) and matches of >16 (when searching Swissprot-34) and 17 bits (SP-Trembl-2/Genpeptide-100) were considered to be significant (39). Some multiple sequence alignments were converted into the Swise-generated (49) weighted profiles corrected for a bias sequence representation (50) and were used to search databases. Searches of Non-redundant Databases (NRD) with single sequences were performed with a family of Blast programs (51,52). Repeats in proteins were identified using the SAPS program (53). Logos of multiple alignments were generated using the Delila program package (54) and a script of Pietrokovski available through ftp://bioinformatics.weizmann.ac.il/pub/software/logoaid/ . Alignments are presented using the BOXSHADE program downloaded through ftp://ulrec3.unil.ch/pub/boxshade/ .

Building an all-inclusive multiple alignment of inteins was initially explored using different inteins identified previously (24). Typically, a procedure started from a ClustalW-made alignment of closely related inteins and, using results of database searches, an alignment of inteins as large as possible with the intein blocks (24,25) being correctly aligned, was generated in an iterative manner. Further extention of the alignment proceeded using three non-overlapping parts of inteins derived from the N-terminal (Nintein), middle (Dod) and C-terminal (inteinC) regions, respectively. The Nintein encompasses blocks A and B, Dod domain blocks C, D, E and H and inteinC blocks F and G. The C-terminal border of the Nintein and N-terminal border of the inteinC were determined from comparisons of a new 'minimal' intein recognized in the *Methanococcus jannaschii* ORF MJ078 with other inteins (see legend to Fig. 1) and were set not far downstream from block B and upstream from block F, respectively. The boundaries of the Dod domain were limited by the N-terminal border of block C and C-terminal border of block H. Divergent sequences of variable size in different inteins between the Nintein and Dod, and the Dod and inteinC domains were excluded from further analyses. Alignments of the three parts of inteins were extended and completed in an iterative manner. Also, to search databases, some HMMs were trained on multiple alignments of artificial minimal inteins from which the part delimited by the Nintein and inteinC domains was removed and the flanking sequences were religated. As a result of this study, four 'non-canonical' inteins and non-intein relatives of them were identified and are described below.

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 klbA gene (*Mja* KlbA) (36). The intein-less version of *Mja* KlbA has been found to be the most closely related to five uncharacterized proteins other than the RK2 klbA 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 *Mja* KlbA intein disrupts a conserved domain (Fig. 2).

The *Mja* KlbA intein is small (168 aa) being the third 'minimal' Dod-less intein identified so far, the two others being *Ppu* DnaB and *Mxe* GyrA inteins (24,25). An Asn/Ser splice 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 Cys(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 *Ssp* ORFs, sll1360 (DNA-polymerase IIIγ/τ, Pol-IIIγ/τ), sll2005 (DNA-gyrase B, GyrB) and slr0603 (DNA-polymerase IIIα, Pol-IIIα). The *Synechocystis* proteins hosting inteins are predicted to be components of the replicative apparatus and the insertions were found within the conserved areas after Glu129 (sll1360), Gly436 (sll2005) and Tyr774 (slr0603).

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 slr0603 ORF (Fig. 1). A Dod 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 EX_1H-HX_3H 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 EX_1H-HX_3H DNase family (22,23; A.Gorbalenya, unpublished

Figure 1. Nintein, Dod and intenC domains in new non-canonical inteins. Shown is an excerpt from three ClustalW-made multiple alignments encompassing domains: (Nintein), 44 Ninteins and its homologs from the *Sce* HO DNase and 26 Hh 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* TFIIB and Rpol_A″ and *Mle* RecA); (ii) two known minimal inteins (*Mxe* GyrA and *Ppu* DnaB); (iii) new inteins [*Mja* KlbA, *Ssp* GyrB, Pol-IIIγ/τ (sll1360 and sll1360x ORFs) and Pol-IIIα (slr0603 and sll1572 ORFs); numbering according to the Cyanobase (55)]; (iv) two intronic Dods containing inteinC domain (*Pan* ND5i3 and *Pan* COIi2) 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; numbering according to the *M.jannaschii* database (36)] and (v) two representatives of the Hh proteins (*Cel* M75 and *Dme* HH). The blocks may contain more columns than shown. In the Dod domains, the two other blocks (D and H) are also conserved (24,25) but not shown. Not all proteins have a full complement of the conserved blocks (see also text). New inteins and related proteins were identified as follows. The *Mja* KlbA intein matched 10 previously described inteins of archael origin using Blast [probability of finding them by chance in a database of the same size (*P*) =3.2e^{-18} –9.0e^{-5}]. A subsequent analysis of these matches identified two regions, Nintein and inteinC, conserved in *Mja* KlbA and 11 Ser-inteins. These regions are separated by only 4 aa in the KlbA intein or a Dod domain in the Ser-inteins. Two blocks conserved less than those previously described (24,25), B′ in the Nintein and F′ in the intein C, flank the (DNase) insertion from the N-terminal and C-terminal sides, respectively. The MACAW confirmed (*P* <1.0e^{-20}) reliability of these blocks for the KlbA and eight Ser-inteins, but in some other inteins, only a tentative assignment was possible which should be confirmed in future (unpublished). The Nintein and inteinC domains were also identified in three *Synechocystis* proteins. The *Ssp* Pol-IIIγ/τ sll1360 and GyrB sll2005 ORFs were hit with 40.78 and 31.33 bits, respectively, by an HMM trained on a multiple alignment of 17 previously recognized inteins. The *Ssp* Pol-IIIα two ORFs, slr0603 and sll1572, were scored with 29.62 and 13.40 bits, respectively, by an HMM trained on a multiple alignment of 36 artificial Nintein-inteinC proteins. This HMM also matched (17.49–22.81 bits) an ~100 aa region in the most N-terminal part of the five Hh-C from *Caenorhabditis elegans* (60) and subsequent searches matched other 21 Hh including those shown. Sixty-four of the 68 Dod domains were classified with the minor or major groups. The minor group was identified as follows. The 323 aa *Ssp* Pol-IIIγ/τ sll1360x ORF, nucleotides 10962–11936 in the complementary strand of the D90907 entry with no reference in the published Cyanobase (55), has been scored (*P* = 1.2e^{-08}) as being next best to the self-match by the 216 aa Dod-containing sequence delimited by Nintein and inteinC domains of *Mle* RecA intein using the TblastN. A group of 20 Dod domains distinct from the other Dod proteins was identified in the course of the Blast- and HMM-based iterative database searches initiated independently with *Mle* RecA and *Ssp* sll1360x sequences. This group also includes MJ1098 and 17 intronic (putative) DNases (see Fig. 3D for accession numbers). The major group was identified through iterative searches initiated with intein Dod domains and is composed of 44 Dods including 38 of intein origin, three *M.jannaschii* ORFs (MJ0314, MJ0398 and U67501x), and four other Dods. An HMM trained on an alignment of this group did not give significant scores to the minor group Dods (unpublished) but matched free-standing and intronic Dod proteins outside the training set, indicating that this group can be enlarged. It matched (25.31 bits) *Pan* ND5i3 which has a close homolog *Pan* COIi2 (61). The latter was hit (13.99 bits, a marginal, next to the true score) downstream of the Dod domain by an HMM trained on the 46 inteinC. These as well as other two intronic proteins (unpublished) contain the Dod domain fused with a domain closely resembling inteinC. The position of the last residue in each block in the intein, Hh-C domain and Dod protein is listed to the right of the block. The last residue of block G belongs to C-extein. Accession numbers are shown at the right. GB, GenBank; TE, Trembl; SP, SwissProt. Intein designation and nomenclature were as described (2,24) with minor modifications. First three letters denote organism and the following letters depict protein. Intron-encoded proteins contain suffices i in their names. Pol-1, DNA-polymerase, 1st intein; Pol-IIIα and Pol-IIIγ/τ, DNA-polymerase IIIα and IIIγ/τ subunits; GyrA and GyrB, DNA-Gyrase A and B subunits; Pps, unknown function; DnaB, DnaB replicative helicase; Rpol_A″, RNA polymerase subunit A″; TFIIB, transcription factor IIB; COI, cytochrome oxidase; ND5, NADH ubiquinone reductase subunit 5; M75 and HH, hedgehog proteins. *Cel, Caenorhabditis elegans; Dme, Drosophila melanogaster; Mja, M.jannaschii; Mxe, Mycobacterium xenopi; Mle, Mycobacterium leprae; Pan, Podospora anserina; Ppu, Porphyra purpurea; Psp, Pyrococcus* sp.*; Sce, S.cerevisiae; Ssp, Synechocystis* sp.*; Tli, Thermococcus litoralis*. 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 blocks within the sll1572 ORF is indicated taking into account a 21 aa residue extention from the postulated initiator Met (55) to the 5′-direction.

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 EX_1H-HX_3H 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 sll2005 intein or its ancestor, given the origin of the GyrB HN domain relatives (Fig. 3A).

DNA-polymerase IIIγ/τ Dod domain is encoded in an alternative ORF to the intein ORF and has complex relationships with the other Dods

A putative DNase, in this case of the Dod-type, was also identified in an ORF alternative to that of the Pol-IIIγ/τ sll1360 (named sll1360x). This ORF is located within a genome region delimited

Figure 2. Multiple alignment of *M.jannaschii* KlbA protein and its closest relatives. The alignment was generated with the ClustalW program (the most variable N-terminal and C-terminal parts of the alignment as well as the main portion of the intein are not shown). Genpeptide (GP) or *M.jannaschii* database (MJ) identificators are listed at the left and GB identificators at the right. The positions of the most left residue in the raws of sequences as well as the distances to the end of the protein are indicated at the left and right sides, respectively. *Cli, Chlorobium limicola; Mvo, Methanococcus voltae; Sso, Sulfolobus solfataricus*. The intein-less version of *Mja* KlbA (consisting of 553 aa) have matched the five other proteins (474–552 aa, P <1.7e^{-26}) on top of the klbA gene product (298 aa) using TblastN. MJ1287_8, a frameshifted fusion of the previously reported MJ1287 and MJ1288 ORFs (36). Other designations see Figure 1 legend.

in the intein ORF sll1360 by blocks B and F (the Pol-IIIγ/τ scheme in Fig. 3B and see below). It was found that the Dod domains of Pol-IIIγ/τ sll1360x and *Mle* RecA intein are grouped together and separately from dozens of homologs associated with the other inteins (named minor and major groups, respectively) (Fig. 1). In line with this observation, the Dod domains of the minor and major groups are different in many positions of blocks Dod1 and Dod2, the most conserved sequences of these proteins. [For instance in the Dod2, only positions 156 (Leu dominancy) and 164 (Gly dominancy) have a similar residue profile and in at least seven positions (153, 155, 157, 158, 159, 161 and 162), the two groups differ strikingly (Fig. 3C)]. Each of the two groups also includes non-described previously free-standing Dod-proteins from *M.jannaschii* as well as intronic homologs (legend to Fig. 1). This indicates that the Dod domains of the different genomic origins are evolutionary interleaved and may form a common pool in one genome, e.g. *M.jannaschii* encoding canonical and non-canonical inteins and free-standing Dods, or in a species community.

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

The sll1360x Dod domain is flanked by additional domains on either side (the Pol-IIIγ/τ sll1360x scheme in Fig. 3B). 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. 3D) 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 sll1360x Dod domain (the sll1360x scheme in Fig. 3B) 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 sll1360x 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 *Halobacterium salinarium* phage ϕH [(57), the C-terminal half of the alignment placed between the GyrB and sll1360x schemes in Fig. 3B]. In the two sequences, the majority of the 34% identical residues belong to the active-site forming amino acids and include among others a CECGCG run. In the sll1360x 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 sll1360x scheme in Fig. 3B).

In the *Ssp* Pol-IIIγ/τ, the conserved N-terminus of the first CG domain of the sll1360x ORF overlaps with the C-terminal border of the sll1360 Nintein and the conserved C-terminus of the YK domain overlaps with the N-terminus of the sll1360 inteinC (the Pol-IIIγ/τ scheme along with the sll1360 and sll1360x ORFs translations in Fig. 3B). This overlapping organization indicates a complex expression mechanism of the ORFs representing the sll1360/sll1360x 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 sll1360/sll1360x mRNA by a ribosome slippage mechanism (58) which would allow two frameshifts, back/forward (–1/+1), with reading of the sll1360x domains in-frame and between the Nintein and inteinC domains (the Pol-IIIγ/τ scheme in Fig. 3B). Either of these mechanisms would affect host protein expression. [Very recently, splicing of the *Ssp* Pol-IIIγ/τ intein has been demonstrated in *Escherichia coli* (58)].

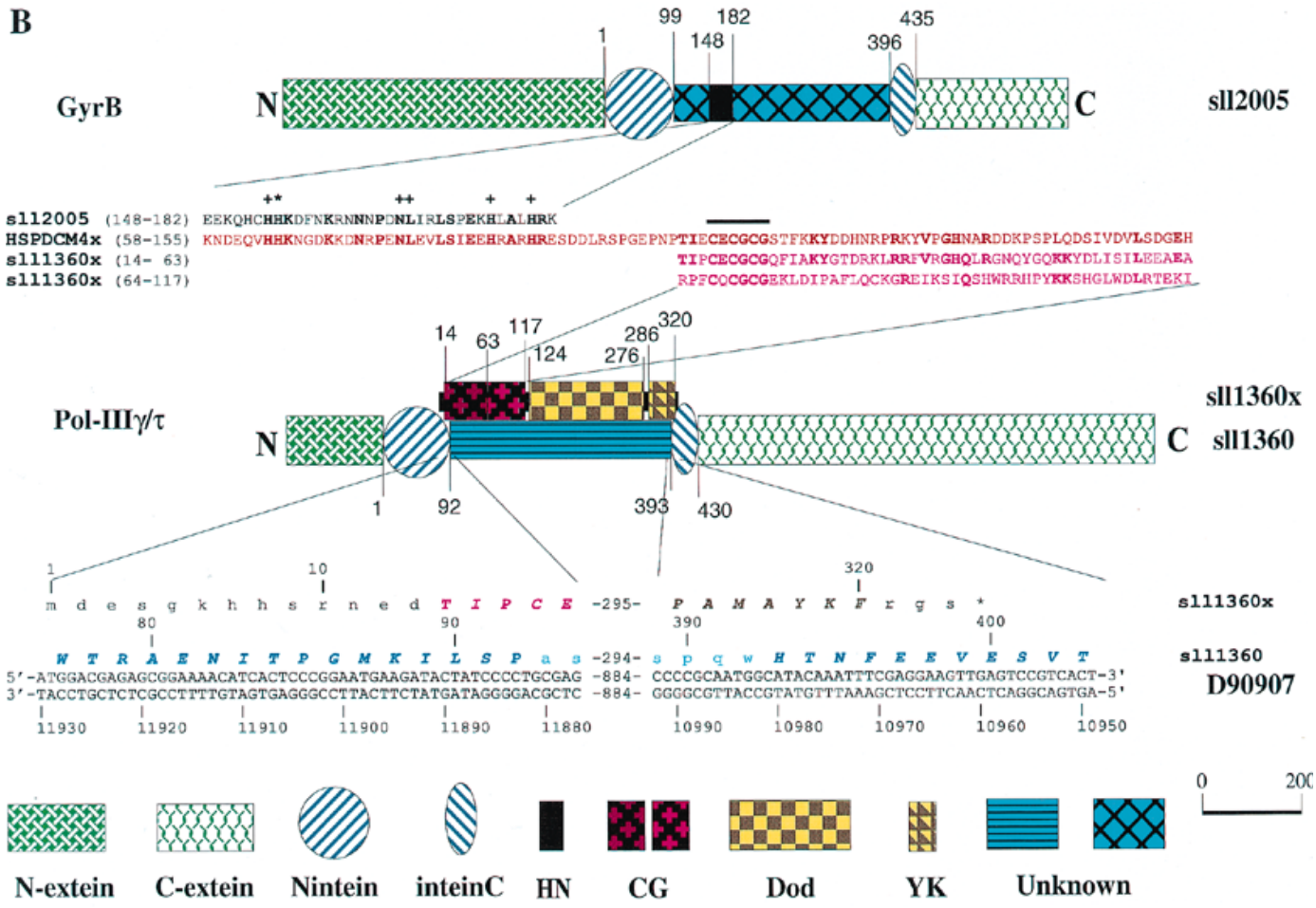
Unrelated Pol-IIIγ/τ and GyrB DNase-containing ORFs are 'linked' by a phage ORF through CG and HN domains

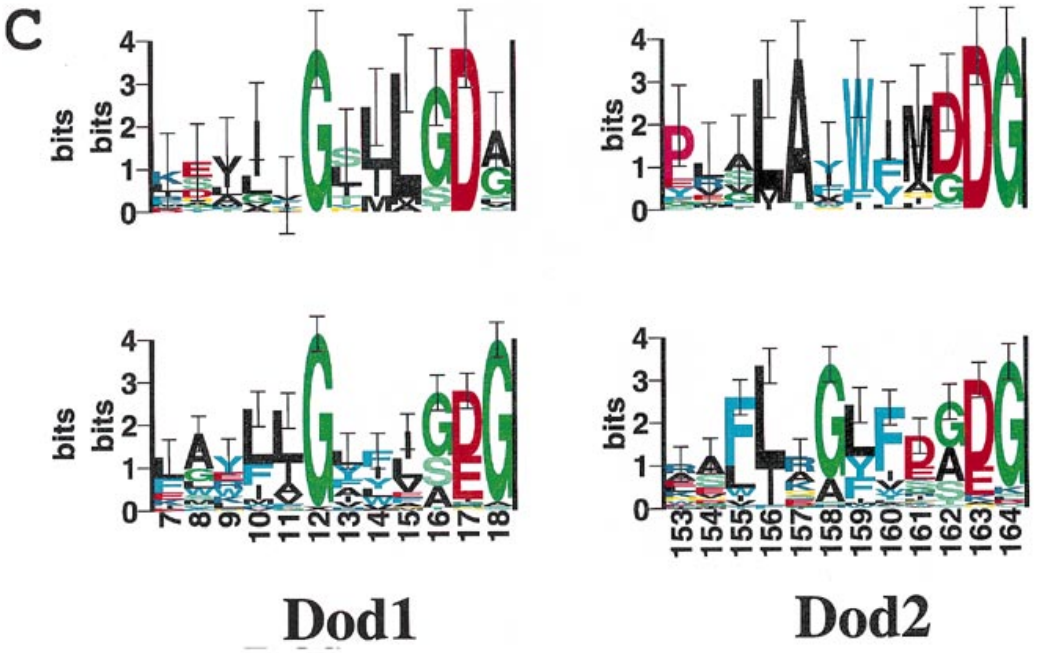
Not only the CG domain unites the sll1360x and HSPDCM4x ORFs, they also both include a DNase domain, although belonging to the two different families, Dod and HN, respectively. Within the HSPDCM4x ORF, the HN and CG domains are separated by only a short peptide with loop-forming potential (SDDLRSPGEPNP) (HSPDCM4x sequence beneath the GyrB scheme in Fig. 3B). Remarkably, the HN domain of the HSPDCM4x ORF revealed a strong sequence similarity to that of the GyrB sll2005 intein. Thus, the phage ORF, through the HN and CG domains, 'links' two unrelated *Synechocystis* DNase-containing ORFs (the GyrB sll2005 and Pol-IIIγ/τ sll1360x schemes and the alignment between them in Fig. 3B). The CG domain may cooperate with different DNase domains given the

A

```
sp82i          ESLVVDHIDRNRHNNHFSNLRWVSRKENSSNISADTK  70-106 GB:U04812
phi-Ei         EDLVVDHIDQDRDNNHCSNLRWVSRKENSNNISADTR  69-105 GB:U04813
sp01i          EGLVVDHKDGNKDNNLSTNLRWVTQKINVENQMSRGT  69-105 SP:P34081
spp1           GKELINHIDGDKTNNYFKNLEWCDHKENLMHAYMTGA  70-106 GB:X67865
LL-r1t         NKATVDHIDGNRKNNSIDNLRWATYSENNSRFETIGV  68-104 GB:U38906
LL-Hi          EKNTVNHINGIKTDNRAVNLEWASRSEQMYHAYQHHL  74-110 GB:L37351
Y38_BPT7       KGYYIDHIDGNPLNDALDNLRLALPKENSWNMKTPKS  28-64  SP:P03797
RBNBi          RTDEIHHKDGNRENNDLDNLMCLSIQEHYDIHLAQKD  21-57  GB:X59078
T5Ytr2         PELEVDHKDRDRLNLSKDNLQVLSKIEHQRKTNKDNG  70-106 GB:Y00364
T5ytr1         SELEVDHKDRNKLNFSLDNLVVMTKQDHRIKTTVERG  73-109 GB:Y00364
A87Rn          SNWTVHHIDNDPSNNHCDNLVWASPETQRKEQRPMET 159-195 GB:U42580
A87Rc          HHLIVDHIDDDKQNARLDNLQLLTNQENSKKRHLKKY 358-394 GB:U42580
A354Rn         LKHTADHYNGIRNDNYIDNIRWATPEEQAKNRNMPCT  45-81  GB:U42580
A354Rc         PSEIIRHKYDDKLDFRPENLLIGTQSQNISDAHDNGK 212-248 GB:U42580
A422Rn         ETFTIDHIDRNPLNNSVSNLRWVNKHVQLENRRDICR  76-112 SP:P34081
A422Rc         EGVIINHKDHNKQNASLDNLEILTRSENAIAAHDAGK 254-290 SP:P34081
consensus          &dHid $@ nn &dNL w&#  en

Ssp GyrB       EKQHCHHKDFNKRNNNPDNLIRLSPEKHLALHRKHIS 149-185 GB:D90908
Cons.family     e hH          nl       h   h
```





D

Block		<--1-->	<-----2---->	<---3--->		
Ama orf176i	8	PSIFIPRVN..	LRKLQELVVPHI..	HTSLLYKIHL	0	TE:Q33760
Ama orf242i	11	YVIYIPKGD..	VDKIRPFLVPHM..	HASMMYKLGL	0	GB:U41288
Ama orf231i	8	YRIRFSRLPDN	ITRLRDLTLPHM..	HPLMIYKLGL	0	TE:Q37087
Ama orf246i	15	YLLYIDNAS..	MDILRPQVKPHM..	HPDMLYKIAK	1	GB:U41288
Ama orf261i	10	YILYIHTSS..	FDKFASIVRPHM..	HPDMLYKLGE	0	GB:U41288
Ama orf233i	10	YCIYIKAES..	MQLIADICKPHT..	HPSMYYKFNG	2	GB:U41288
Kla COIi	8	YVIYIKVKS..	IKLLYPNIKNYI..	IPSMKYKFEY	12	TE:Q34835
Pan COBi3	9	PRIYIFKES..	MDTLITLVKPYF..	IKEMLYKLGL	0	GB:X55026
Sce COBi	11	FVIYVWKES..	MPILTKIVSPYI..	IPSMKYKLGN	2	SP:P03873
Csm COBi	12	YRIYVVAKS..	MPRLRSLVSCYL..	DPSMHYKINN	5	TE:Q06824
Kth Lsri	8	PIIFIPSDN..	YKHYYDLIIPYI..	IPEMKYKLPT	0	TE:Q34839
Sce Lsri	8	PIIYIDSMS..	YLIFYNLIKPYL..	IPQMMYKLPN	9	SP:P03882
Pca Lsri	8	PMIYIPNKD..	YELFYNLVNPYI..	MPEMKYKLTF	2	TE:Q34807
Chu Lsri	13	YRIAIPEAS..	SGAFREVIKPFL..	VDCMRYKVSD	6	TE:Q32001
Ceu Lsri	14	WALSINAPD..	YDKFRALITKIDLI	PTIFPYKLHR	3	TE:Q02449
Cpa Ssri2	10	YYIGLPAKT..	FGRFINIIEPYVRE	VPALCYKLNE	7	TE:Q39559
Cpa Ssri1	15	CRISIKDQEN.	LLNLLRYIAPVVP.	VREMLYKVCF	35	TE:Q39560
Mja MJ1098	9	YLRLFKSKQ..	VHAFCSIVEPFI..	PPSYRNKVRC	13	TE:Q58498
Ssp sll1360x	12	YYTKMGAVA..	TRKWLSDLEIYS..	IPAMAYKFRG	1	GB:D90907
Mle RecA	10	AVLVFSTAA..	TAKFQSLIAPYV..	APSMEYKLLP	20	SP:P35901
consensus		y&i!i	& & && p&&	p m yKl		

Figure 3. (Above and overleaf). *Synechocystis* DNA-gyrase B subunit and DNA-polymerase IIIγ/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_1H-HX_3H family (unpublished). (**B**) Domain organization and sequence features of Ssp GyrB and Pol-IIIγ/τ inteins and an archael phage ORF. Two schemes depicting the intein/extein organization of *Ssp* GyrB and Pol-IIIγ/τ are shown. Between the schemes, alignments of HN domains (left; the EX_1H-HX_3H 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-IIIγ/τ sll1360x 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 *Halobacterium salinarium* phage ϕH. 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 product 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-IIIγ/τ sll1360 region embedded between the Nintein and inteinC (not shown). Beneath the *Ssp* Pol-IIIγ/τ 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 (54) 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; light green, S,T,C; 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 thermotolerans; Pca, Pichia canadensis;* COB, apocytochrome *b*; Lsr and Ssr, large and small subunit ribosomal RNAs.

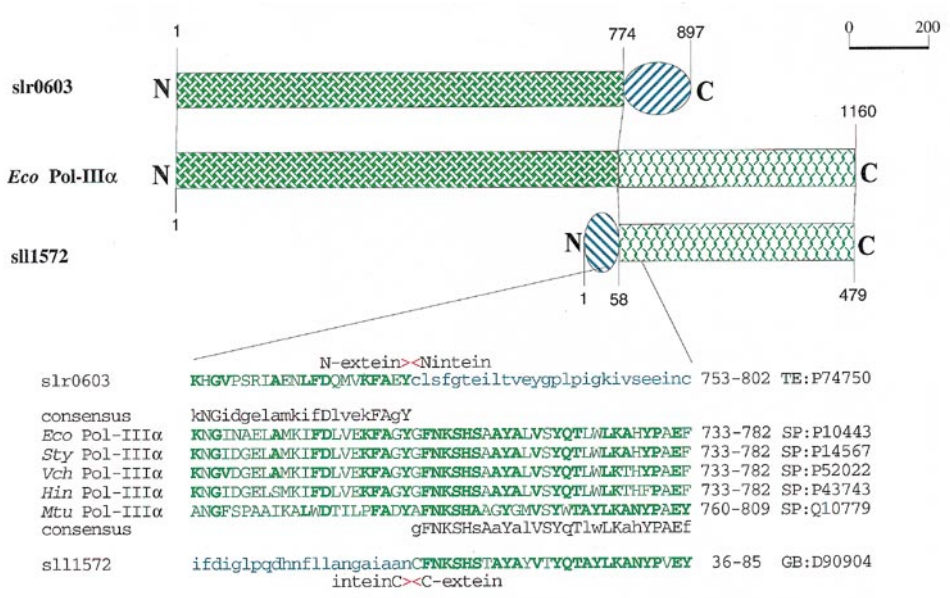


Figure 4. The extein/intein organization of the disrupted DNA-polymerase IIIα subunit of *Synechocystis*. A scheme of the intein/extein organization of *Ssp* Pol-IIIα in comparison with *E.coli* intein-free Pol-IIIα, and an alignment of *Ssp* Pol-IIIα with bacterial Pol-IIIα around the site of the intein insertion are shown. *Eco*, *Escherichia coli*; *Sty, Salmonella typhimurium; Vch, Vibrio cholera; Hin, Haemophilus influenza*. For other abbreviations see Figures 1 and 3. Bold, identical residues in the *Ssp* Pol-IIIα exteins and Pol-IIIα of other origins.

structural organizations of Pol-IIIγ/τ sll1360x 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-IIIγ/τ 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 IIIα Nintein and inteinC are associated with the disrupted host protein encoded by two spatially separated ORFs

The Nintein domain identified within the Pol-IIIα slr0603 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-IIIα is encoded within the 479 aa sll1572 ORF which is located far distant from slr0603 in the *Synechocystis* genome (55). The 1–774 aa portion of the slr0603 protein combined with the 58–479 aa portion of the sll1572 are perfectly aligned with Pol-IIIα 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-IIIα and consists of the Nintein and inteinC domains encoded by two spatially separated ORFs.

Assuming that an active form of *Ssp* Pol-IIIα 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-IIIα might function as a non-covalently bound complex of the exteins with the intein being splicing-defective, although possibly proteolyticaly active and involved in formation of the complex.

Although it is presently unclear how the *Ssp* Pol-IIIα 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-IIIα 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 COIi2). However, in the *Ssp* Pol-IIIα, no vestiges of a DNase domain were found in the vicinity of either slr0603 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 *Mja* KlbA 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-IIIγ/τ 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 Spaan for encouragement and facilities, Gerard Canters for office space, Tom Schneider and Dave Gillespie for helping to install the logos 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

1 Dujon,B., Belfort,M., Butow,R.A., Jacq,C., Lemieux,C., Perlman,P.S. and Vogt,V.M. (1989) *Gene*, **82**, 115–118.
2 Perler,F.B., Davis,E.O., Dean,G.E., Gimble,F.S., Jack,W.E., Neff,N., Noren,C.J., Thorner,J. and Belfort,M. (1994) *Nucleic Acids Res.*, **22**, 1125–1127.
3 Belfort,M., Reaban,M.E., Coetzee,T. and Dalgaard,J.Z. (1995) *J. Bacteriol.*, **177**, 3897–3903.
4 Cooper,A.A. and Stevens,T.H. (1995) *Trends Biochem. Sci.*, **20**, 351–356.
5 Colston,M.J. and Davis,E.O. (1994) *Mol. Microbiol.*, **12**, 359–363.
6 Lambowitz,A.M. and Belfort,M. (1993) *Annu. Rev. Biochem.*, **62**, 587–622.
7 Shub,D.A. and Goodrich Blair,H. (1992) *Cell*, **71**, 183–186.
8 Cech,T.R. (1990) *Annu. Rev. Biochem.*, **59**, 543–568.
9 Kane,P.M., Yamashiro,C.T., Wolczyk,D.F., Neff,N., Goebl,M. and Stevens,T.H. (1990) *Science*, **250**, 651–657.
10 Dujon,B. (1989) *Gene*, **82**, 91–114.
11 Gimble,F.S. and Thorner,J. (1992) *Nature*, **357**, 301–306.
12 Jacquier,A. and Dujon,B. (1985) *Cell*, **41**, 383–394.
13 Dalgaard,J.Z., Garrett,R.A. and Belfort,M. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 5414–5417.
14 Sharma,M., Ellis,R.L. and Hinton,D.M. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 6658–6662.
15 Bechhofer,D.H., Hue,K.K. and Shub,D.A. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11669–11673.
16 Loizos,N., Tillier,E.R. and Belfort,M. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11983–11987.
17 Shub,D.A., Gott,J.M., Xu,M.Q., Lang,B.F., Michel,F., Tomaschewski,J., Pedersen Lane,J. and Belfort,M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 1151–1155.
18 Hensgens,L.A., Bonen,L., de Haan,M., van der Horst,G. and Grivell,L.A. (1983) *Cell*, **32**, 379–389.
19 Waring,R.B., Davies,R.W., Scazzocchio,C. and Brown,T.A. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 6332–6336.
20 Michel,F., Jacquier,A. and Dujon,B. (1982) *Biochimie*, **64**, 867–881.
21 Michel,F. and Dujon,B. (1986) *Cell*, **46**, 323.
22 Gorbalenya,A.E. (1994) *Protein Sci.*, **3**, 1117–1120.
23 Shub,D.A., Goodrich Blair,H. and Eddy,S.R. (1994) *Trends Biochem. Sci.*, **19**, 402–404.
24 Perler,F.B., Olsen,G.J. and Adam,E. (1997) *Nucleic Acids Res.*, **25**, 1087–1093.
25 Pietrokovski,S. (1994) *Protein Sci.*, **3**, 2340–2350.
26 Hirata,R., Ohsumk,Y., Nakano,A., Kawasaki,H., Suzuki,K. and Anraku,Y. (1990) *J. Biol. Chem.*, **265**, 6726–6733.
27 Koonin,E.V. (1995) *Trends Biochem. Sci.*, **20**, 141–142.
28 Hodges,R.A., Perler,F.B., Noren,C.J. and Jack,W.E. (1992) *Nucleic Acids Res.*, **20**, 6153–6157.
29 Chong,S., Shao,Y., Paulus,H., Benner,J., Perler,F.B. and Xu,M.Q. (1996) *J Biol. Chem.*, **271**, 22159–22168.
30 Cooper,A.A., Chen,Y.J., Lindorfer,M.A. and Stevens,T.H. (1993) *EMBO J.*, **12**, 2575–2583.
31 Hirata,R. and Anraku,Y. (1992) *Biochem. Biophys. Res. Commun.*, **188**, 40–47.
32 Xu,M.Q. and Perler,F.B. (1996) *EMBO J.*, **15**, 5146–5153.
33 Duan,X., Gimble,F.S. and Quiocho,F.A. (1997) *Cell*, **89**, 555–564.
34 Gimble,F.S. and Stephens,B.W. (1995) *J. Biol. Chem.*, **270**, 5849–5856.
35 Pietrokovski,S. (1996) *Trends Genet.*, **12**, 287–288.
36 Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J.F., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.L., Geoghagen,N.S.M., Weidman,J.F., Fuhrmann,J.L., Venter,J.C., *et al.* (1996) *Science*, **273**, 1058–1073.
37 Fsihi,H., Vincent,V. and Cole,S.T. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 3410–3415.
38 Riera,J., Robb,F.T., Weiss,R. and Fontecave,M. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 475–478.
39 Eddy,S.R. (1996) *Curr. Opin. Struct. Biol.*, **6**, 361–365.
40 Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) *J. Mol. Biol.*, **235**, 1501–1531.
41 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
42 Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) *Proteins*, **9**, 180–190.
43 Eddy,S.R. (1995) Ismb, 114–120.
44 Henikoff,S. (1996) *Curr. Opin. Struct. Biol.*, **6**, 353–360.
45 Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.
46 Eddy,S.R., Mitchison,G. and Durbin,R. (1995) *J. Comput. Biol.*, **2**, 9–23.
47 Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol.*, **147**, 195–197.
48 Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
49 Birney,E., Thompson,J.D. and Gibson,T.J. (1996) *Nucleic Acids Res.*, **24**, 2730–2739.
50 Altschul,S.F., Carroll,R.J. and Lipman,D.J. (1989) *J. Mol. Biol.*, **207**, 647–653.
51 Altschul,S.F. and Gish,W. (1996) *Methods Enzymol.*, **266**, 460–480.
52 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
53 Brendel,V., Bucher,P., Nourbakhsh,I.R., Blaisdell,B.E. and Karlin,S. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 2002–2006.
54 Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
55 Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S., Kimura,T., Hosouchi,T., Matsuno,A., Muraki,A., Nakazaki,N., Naruo,K., Okumura,S., Shimpo,S., Takeuchi,C., Wada,T., Watanabe,A., Yamada,M., Yasuda,M. and Tabata,S. (1996) *DNA Res.*, **3**, 185–209.
56 Eddy,S.R. and Gold,L. (1991) *Genes Dev.*, **5**, 1032–1041.
57 Stolt,P., Grampp,B. and Zillig,W. (1994) *Biol. Chem. Hoppe Seyler*, **375**, 747–757.
58 Farabaugh,P.J. (1996) *Microbiol. Rev*, **60**, 103–134.
59 Liu,X. and Zhuma,H. (1997) *FEBS Lett.*, **408**, 311–314.
60 Porter,J.A., Ekker,S.C., Park,W.J., von Kessler,D.P., Young,K.E., Chen,C.H., Ma,Y., Woods,A.S., Cotter,R.J., Koonin,E.V. and Beachy,P.A. (1996) *Cell*, **86**, 21–34.
61 Cummings,D.J., McNally,K.L., Domenico,J.M. and Matsuura,E.T. (1990) *Curr. Genet.*, **17**, 375–402.
62 Dalgaard,J.Z., Moser,M.J., Hughey,R. and Mian,I.S. (1997) *J. Comput. Biol.*, **4**, 193–214.
63 Tanaka Hall,T.M., Porter,J.A., Young,K.E., Koonin,E.V., Beachy,P.A. and Leahy,D.J. (1997) *Cell*, **91**, 85–97.
64 Pietrokovski,S. (1998) *Protein Sci.*, **7**, 64–71.