The DNA polymerase genes of several HMU-bacteriophages have similar group I introns with highly divergent open reading frames

Heidi Goodrich-Blair⁺ and David A. Shub^{*}

Department of Biological Sciences, Center for Molecular Genetics, University at Albany, SUNY, 1400 Washington Avenue, Albany, NY 12222, USA

Received June 6, 1994; Revised and Accepted July 25, 1994

GenBank accession nos U04812 and U04813

ABSTRACT

A previous report described the discovery of a group I, self-splicing intron in the DNA polymerase gene of the *Bacillus subtilis* bacteriophage SPO1 (1). In this study, the DNA polymerase genes of three close relatives of SPO1: SP82, 2C and ϕ e, were also found to be interrupted by an intron. All of these introns have group I secondary structures that are extremely similar to one another in primary sequence. Each is interrupted by an open reading frame (ORF) that, unlike the intron core or exon sequences, are highly diverged. Unlike the relatives of *Escherichia coli* bacteriophage T4, most of which do not have introns (2), this intron seems to be common among the relatives of SPO1.

INTRODUCTION

Group I introns interrupt the genes of a diverse range of organisms: eukaryotic organelles, protist nuclei, gram-negative and gram-positive bacteriophage and eubacteria (3). Regardless of their location all group I introns share secondary structure elements and conserved sequences that constitute the core structure required for splicing (4,5). This secondary structure is important for the proper folding of the intron into its catalytic form and is even sufficient for many group I introns to self-splice in the absence of protein cofactors (6).

Although this common structure is critical for splicing, group I introns can be of variable length due to the presence of additional non-conserved sequences that are not required for folding. In fact, group I introns are often interrupted by entire open reading frames (ORFs), some of which encode maturases required for splicing or site-specific DNA endonucleases that initiate intron homing (3). The ORFs sometimes evolve independently of the introns they inhabit, the most striking case being the intron in the ND1 gene of *Neurospora* mitochondria, where two independently isolated strains have unrelated ORFs inserted into different intron locations (7).

Splicing of group I introns occurs via two sequential transesterification reactions. The first is initiated by the 3'

hydroxyl group of a guanosine cofactor that attacks the 5' splice site (s.s.). This initial reaction results in the release of exon I and covalent attachment of the non-coded guanosine to the 5' end of the intron. In the second transesterification reaction the 3' s.s. is attacked by the 3' hydroxyl of exon I, yielding ligated exons and a free intron (6). Because group I introns retain the non-coded guanosine at their 5' ends they will be end-labeled if the splicing reaction is initiated by a radiolabeled guanosine cofactor. As many group I introns can self-splice *in vitro*, introns can be identified in deproteinized extracts by end-labeling with [³²P]-GTP (8).

GTP-labeling has led to the discovery of several self-splicing introns including two of the three introns in the *Escherichia coli* bacteriophage T4 (9,10), and one intron in the *Bacillus subtilis* bacteriophage SPO1 (1). Interestingly, these bacteriophage introns all interrupt genes involved in DNA synthesis. The T4 introns are found within *nrdB* (9,11) and *sunY/nrdD* (10,12), both of which encode ribonucleotide reductases, and in *td*, encoding thymidylate synthase (13) (all of which are required for the synthesis of DNA precursors), while the SPO1 intron interrupts its gene for DNA polymerase. The consistent presence of bacteriophage group I introns within DNA synthesis genes has prompted speculation that these introns might be part of a regulatory mechanism controlling this pathway (1).

Many independently isolated members of the T-even family of bacteriophage (of which T4 is a member) have been screened for the presence of the *td*, *nrdB* and *sunY* introns. Of the twenty-one isolates for which complete data are available, only T4 harbors all three introns and most do not have any (2). It was, therefore, fortunate that T4 (rather than one of its intron-less relatives) was chosen as the archetype and was consequently the subject of intron searches.

SPO1 is also a member of a closely related family of bacteriophages—based on the presence of 5-hydroxymethyluracil (HMU) rather than thymine in their DNA, genetic and restriction maps, DNA/DNA cross hybridization, antigenic cross-specificity, genetic complementation and capsid morphology (14). Considering the possibility that, like the T-even phage, HMU-

*To whom correspondence should be addressed

⁺Present address: Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA 02115, USA

phage might harbor several optional introns, only one of which had been identified in SPO1, we screened the independently isolated HMU-phage SP82, 2C, ϕ e and SP8 (15–18) for self-splicing group I introns by GTP-labeling their RNA.

MATERIALS AND METHODS

Strains

B. subtilis 168 was used for maintenance of SP82G (subsequently referred to as SP82), SPO1 and 2C. ϕ e was propagated on *B. subtilis* 3610, both of which were provided by A.L. Sonenshein. SP8 and its host *B. subtilis* strain Marburg (ATCC 15563) were obtained from the Félix d'Hérelle Reference Center maintained by H.W. Ackermann.

RNA isolation

B. subtilis cells were grown in modified M9S (1) to an OD₅₄₀ of 0.3 to 0.4 (approximately 1×10^8 cells per ml) and infected at a multiplicity of approximately ten. Samples of cells were placed into pre-chilled tubes, harvested and washed twice with 10 mM Tris–HCl (pH 7.5) and 100 μ g/ml chloramphenicol. After pelleting, cells were lysed and RNA was extracted by a lysozyme freeze–thaw method (9).

GTP labeling

Deproteinized RNA was incubated with 0.66 μ M [α -³²P]-GTP (3000 Ci/mmol) for 60 min at 37°C in 50 mM Tris–HCl (pH 7.5), 50 mM NH₄Cl, 30 mM MgCl₂, 1 mM spermidine, 5 mM DTT and 1 unit/ μ l RNasin (8). Reactions were terminated by the addition of 15 volumes of 25 mM EDTA followed by ethanol precipitation. Samples were resuspended in H₂O and denatured at 65°C for 15 min in the presence of two volumes of acrylamide gel dye (95% formamide, 10 mM EDTA, 0.1% xylene cyanol and 0.1% bromophenol blue) before loading on denaturing polyacrylamide gels.

Phage DNA isolation

DNA was isolated from CsCl purified phage particles by three successive phenol and one chloroform:isoamyl alcohol (24:1) extractions followed by dialysis against three changes of 10 mM Tris–HCl (pH 7.5) and 0.1 mM EDTA. DNA concentrations were determined spectrophotometrically.

Polymerase chain reaction amplification

Approximately 0.1 μ g of phage DNA was amplified in 10 mM Tris–HCl (pH 8.3), 50 mM KCl, 1.5 mM MgCl₂, 0.01% gelatin, 200 μ M dTTP, dCTP, dGTP, and dATP and 2.5 units AmpliTaq recombinant Taq DNA polymerase (Cetus corp.). For symmetric amplifications two primers flanking the region to be amplified were added to 0.3 μ M final concentration. For asymmetric amplifications the primer identical to the strand to be made in excess was added to 0.3 μ M while the limiting primer was added to 0.003 μ M. Reactions were subjected to 20–35 rounds of denaturation at 92°C, annealing at the optimal temperature for each set of primers, and extension at 72°C in a thermal cycler. A final extension step of 72°C for 10 min was completed at the end of the multiple cycles. Typically one-tenth of the reaction was analyzed by agarose gel electrophoresis.

DNA sequencing

Amplified DNA was separated from unincorporated deoxynucleotides by elution on a Sephadex G-50 spin column.

Asymmetrically amplified templates were used as single stranded DNA in standard dideoxy chain termination sequencing with the Sequenase sequencing kit (United States Biochemicals). Symmetrically amplified templates were sequenced with a Taq Dideoxy Terminator Cycle Sequencing Kit (Applied Biosystems) and analyzed on an ABI automated sequencer. Plasmid DNA was denatured with 0.2 N NaOH in the presence of 2 mM EDTA for 5 min at room temperature then precipitated in 1.3 M ammonium acetate and 4 volumes of 100% at –70°C for 30 min. Approximately 2.5 μ g of denatured plasmid DNA was annealed to reverse primer (end-labeled with [γ -³²P]-ATP and T4 polynucleotide kinase) and extended with Sequenase T7 DNA polymerase in dideoxytermination mixes. Use of end-labeled primer and omission of the usual labeling step allowed sequencing of discrepancies 17 and 19 nucleotides from the primer.

RNA sequencing

Deproteinized 20 min RNA from cells infected with either SPO1 or SP82 was used as a template for sequencing with AMV reverse transcriptase (Life Sciences Inc.) (1). The primers 5 and 21 anneal to premessage or spliced RNA immediately downstream of the 3' s.s. of SPO1 and SP82 respectively. 20 min after infection the majority of the message is spliced and extension of the primers yields sequence of exon I.

Oligonucleotides

- 1: 5'-GGTCTCATGGAATTCCGTAACGGAA-3' [1706–1731]
- 2: 5'-CAGACAAGTATACGTAACCTAACTG-3' [complement: 1999–1974]
- 5: 5'-TACGTGGGAATTGCTGAGCG-3' [complement: 1872–1853]
- 21: 5'-CTTCTTGGGAATTGTTGAGCG-3' [SP82 complement: 1074–1053]
- 22: 5'-GGNCGN(C/T)TGAG(T/C)AG(T/C)AA(T/C)GANCC-3' [1827–1849]
- 23: 5'-TANAC(A/G/T)ATNCCGAA(T/C)TG(A/G/T)AT(T/C)TT(T/C)TT-3' [complement: 2149–2124]
- 24: 5'-TCNAGNAC(T/G/A)ATNGAGTCGTGNAC-3' [complement: 2527–2505]

Degenerate positions within the oligonucleotides are in parentheses. Numbering in brackets refers to the SPO1 sequence of Scarlato and Gargano (19) except for primer 21 which refers to the SP82 sequence (GenBank accession number U04812). The sequences are the same as the published strand except when indicated as the complement.

RESULTS

An intron within several members of the HMU-family of bacteriophage

RNA was isolated from *Bacillus subtilis* cells infected with HMU-phage SP82, 2C, ϕ e and SP8 as well as from SPO1 (as a control). RNA samples taken 7 min after the onset of phage infection—roughly corresponding to the middle stage of gene expression (20)—were deproteinized, incubated with [α -³²P]-GTP and separated on a denaturing polyacrylamide gel (Fig. 1). A distinct labeled species, slightly larger than the 882-nucleotide intron of SPO1, was apparent in RNA from all of the phage except SP8. RNA from early (3 min) and late (20 min) periods of gene expression were also tested (with the exception of ϕ e from which only 7 min RNA was isolated) but no other labeled species indicative of introns were detected (data not shown). As SP8 might harbor a group I intron that is not efficiently spliced *in vitro*, we attempted to optimize conditions by varying NH₄Cl concentration (1–200 mM) and temperature (37°C and 42°C) but did not obtain evidence for self-splicing (data not shown).

The size and time of expression of the GTP-labeled bands of SP82, 2C and ϕ e RNA suggested they might be introns similar to that of SPO1. If this were the case we expected the GTP-labeled bands to cross hybridize with the SPO1 intron. Indeed, in the case of SP82 and 2C DNA the same restriction fragments hybridized to an SPO1 RNA probe (comprised of 882 bp intron and 284 bp flanking exon sequence) and also to their own GTP-labeled RNAs (21). ϕ e DNA was resistant to cleavage with restriction endonucleases, but spots of total ϕ e DNA anneal to end-labeled ϕ e RNA in addition to the SPO1 RNA probe. However, SP8 total DNA does not hybridize detectably to the SPO1 *in vitro* transcribed RNA, ϕ e GTP-labeled RNA or to its own RNA that had been incubated with GTP (data not shown). These results suggest that the GTP-labeled RNAs of 2C, SP82 and ϕ e are indeed self-excised introns that are similar to that of SPO1. However, SP8 does not appear to harbor an intron related to SPO1. In fact, since the SPO1 RNA probe includes 284 bp of exon sequence, SP8 is likely to be highly diverged in this region of the DNA polymerase coding region.

The HMU-phage introns interrupt the DNA polymerase gene

To determine if the end-labeled bands are indeed introns interrupting the DNA polymerase gene, we amplified a portion of this gene from each phage using the polymerase chain reaction (PCR). Primers 1 and 2 (Fig. 2A) were designed to anneal to the SPO1 DNA polymerase sequence flanking the intron. The

size expected for amplification products from intron-less alleles is based on the sequences of related, intron-less DNA polymerases (21) and any amplification products from intron-plus alleles should be larger. Amplifications with primers 1 and 2 yielded products from SP82 and 2C consistent with the presence of an intron (Fig. 2B). However, no products were obtained using ϕ e or SP8 DNAs so neither the presence nor absence of an intron could be confirmed. In order to resolve this issue we designed degenerate primers 22, 23 and 24 (Fig. 2A) based on conserved amino acid (aa) regions between SPO1, *E. coli* PolI, and other related DNA polymerases (22). Using the upstream primer, 22, we were able to amplify a large fragment from ϕ e in combination with any of the downstream primers (2, 23 and 24). It appears, therefore, that ϕ e has an intron-containing DNA polymerase gene that is more diverged from the others in the region spanned by primer 1. Amplification of SP8 DNA using primer 1 also did not yield any bands. When primer 22 was used in combination with either primer 2 or 23, minor products were obtained that did not correspond to the size expected for either an intron-plus or an intron-minus DNA polymerase gene. In fact, the products decrease in size as the distance between the amplification primers increases (see Fig. 2B), suggesting that they are not annealing to the expected DNA polymerase sequences. Using primers 22 and 24 we were able to amplify a band of the size expected for an intron-minus allele (approximately 700 bp). However, because this band was not prominent or reproducibly present it was not analyzed further.

Sequence of the HMU-phage introns

SP82 and 2C DNA fragments amplified with primers 1 and 2 and ϕ e DNA amplified with primers 22 and 2 were sequenced on both strands. SP82 and 2C were identical throughout the region sequenced and are therefore treated as one in this discussion. Because primer 22 anneals immediately upstream of the 5' s.s. (see Fig. 2A) ϕ e exon I DNA was not amplified for sequencing. Within the 109 bp of SP82 exon I sequence obtained (from the 3' end of primer 1 to the 5' s.s.; see Fig. 2A) there was one nucleotide difference with SPO1. This nucleotide is at the third position of a codon and does not alter the amino acid sequence.

In experiments not shown here we sequenced the messenger RNA of SP82 and SPO1 near the *Eco*RI site of exon I (Fig. 2A). Two discrepancies with the published DNA sequence of SPO1 were found (nucleotides 1727 and 1729 in Fig. 2 of ref. 19). These nucleotides are included within the sequence of primer 1, which was used to amplify templates for sequencing here as well as in our previous work (nucleotides 6 and 8 in Fig. 3 of ref. 1). Resequencing of DNA from plasmid α puc8, cloned directly from the SPO1 genome (23), confirmed the presence of G and C at these positions, respectively, rather than the published C and G. Correction of these nucleotides creates two changes in the published amino acid sequence: asparagine 555 to lysine and glycine 556 to alanine (19).

123 bp of SP82 and ϕ e exon II sequences (from the 3' s.s. to [but not including] primer 2; see Fig. 2A) are compared to SPO1 in Fig. 3. Twenty-two base pairs are variable between the phage. Most of these differences are in the third positions of codons and do not affect the predicted amino acid sequence. Only three residues are affected by sequence differences, the SPO1 amino acids phenylalanine 615 (numbering refers to ref. 19), aspartic acid 627 and glycine 628. In SP82 these amino acids are histidine, aspartic acid and aspartic acid respectively and in

Figure 1. *In vitro* labeling of HMU-phage RNA. RNA was extracted from *B. subtilis* cells seven minutes after infection with SP8, ϕ e, 2C, SP82 or SPO1 phage. After deproteinization and labeling with [α - 32 P]-GTP the RNAs were separated on a 5% polyacrylamide/8 M urea gel next to end-labeled DNA size markers, M1, a 123 bp ladder, and M2, ϕ X174 RF DNA cut with *Hae*III. Sizes of the ϕ X174/*Hae*III fragments are indicated to the right.

ϕ e they are phenylalanine, glutamic acid and aspartic acid respectively.

Intron secondary structure

The sequence of the SP82 intron, with SPO1 and ϕ e differences, is shown folded according to our previous proposal for the SPO1 intron (1), except that a representation of the entire region between P7 and P8 is included (Fig. 4). Variation between the three sequences is rare, occurring within only four regions. These differences support the existence of the proposed secondary structures in two of these regions (P7.2 and P9) and are neutral for a third, P6a. However, they are inconsistent with our previous proposal for the fourth region, between P3 and P4.

The SP82 intron has only nine nucleotide differences with the SPO1 intron—all within P9—and the validity of this stem is supported by the covariation of base-paired nucleotides. The ϕ e sequence varies from SP82 and SPO1 in all four of the regions indicated above. One nucleotide difference weakens the upper portion of P9, but the seven base pairs remaining are typical of most introns of this type (5). In the P7.2 stem the differences are compensatory and provide additional support for the existence of this structure. Two changes in P6a increase the ϕ e terminal loop by one nucleotide and destroy one base pair. Although the ϕ e P6a is therefore less stable than those of SPO1 and SP82, this helix is marginally stable in many other group I introns (4).

On the other hand, differences between ϕ e and the other phage

sequences are not consistent with our previous proposal in the region between P3 and P4 (Fig. 4). In most group I introns this region comprises only 3 to 5 nucleotides (5). Lacking guidance from comparative sequence analysis, our original model of the SPO1 intron in this region was based on maximal predicted stability ($\Delta G = -18.1$ kCal) (24). However, three of the four differences in the ϕ e sequence in this region do not support this prediction. On the other hand, the optimal ϕ e folding ($\Delta G = -19.4$ kCal) is consistent with the variations in SPO1 and SP82 (Fig. 4). In fact, using the RNA-folding computer program *mfold* (24), structures resembling the single ϕ e interrupted stem loop appear as the next two suboptimal foldings of the SP82/SPO1 sequence. Although large insertions between P3 and P4 are very rare among the group I introns, the few examples can be represented as two (rather than three) locally folded RNA helices, giving further support to the conjecture that the alternative structure shown in Fig. 4 may be closer to the correct representation of this region (5,25,26).

HMU-phage intron-ORFs

Each of the phage introns has an ORF that begins in the loop of P8 (L8) and ends within the 3' part of the stem of essential secondary structure P8 (Fig. 4). The SPO1 intron-ORF of 525 nucleotides potentially encodes a 174 amino acid, 19.8 kDa protein. This is smaller than both the SP82 intron-ORF of 558 nucleotides and the ϕ e intron-ORF of 546 nucleotides, which

Figure 2. PCR amplification of HMU intron-containing DNA. **A.** SPO1 intron-containing DNA is schematically represented. Nucleotide positions beginning with the guanosine of the *EcoRI* site within primer 1 are indicated within parentheses. The 5' s.s. and 3' s.s., indicated by vertical arrows, mark the boundaries of the intron. The intron-ORF begins at position 443 and ends at position 967. The primers used for amplification are represented by horizontal arrows, with the arrowhead indicating the 3' end of the oligonucleotide. **B.** Phage DNAs were PCR amplified with six different primer combinations, and 1/10th of each reaction was separated on a 1.2% agarose gel. The three downstream primers 2, 23 and 24 were used in combination with primer 1, shown in the left panel, or with primer 22, shown in the right panel. No DNA (ND) or SPO1 (O1), SP82 (82), 2C, ϕ e or SP8 phage DNAs were amplified using each primer set. λ DNA digested with *Bsr*II was used as a DNA size marker (M).

potentially encode proteins of 21.3 and 20.4 kDa respectively. It is this size difference that causes the SPO1 intron to appear smaller in the GTP-labeling assay (Fig. 1).

ϕ E	AAC	GCT	CAA	CAA	TTC	CCA	AGA	AAG	GTT	AAT	ACT	CCT	ACC	CTT
82	AAC	GCT	CAA	CAA	TTC	CCA	AGA	AAG	GTG	AAT	ACA	CCT	ACT	CTC
O1	AAC	GCT	CAG	CAA	TTC	CCA	CGT	AAG	GTG	AAC	ACG	CCA	ACA	TTA
	N	A	Q	Q	F	P	R	K	V	N	T	P	T	L
ϕ E	TTC	CAA	TAT	AAC	TTT	GAG	ATT	AAA	AAG	ATG	TTT	AAT	TCT	AGG
82	TTT	CAG	TAT	AAT	CAC	GAG	ATT	AAG	AAA	ATG	TTT	AAC	TCT	AGG
O1	TTT	CAG	TAT	AAC	TTT	GAG	ATT	AAG	AAA	ATG	TTT	AAC	TCT	AGG
	F	Q	Y	N	F	E	I	K	K	M	F	N	S	R
ϕ e	TTT	GGT	GAA	GAT	GGT	GTA	ATC	GTA	CAG	TTT	GAC	TAC	TCT	
82	TTT	GGG	GAT	GAT	GGT	GTA	ATT	GTA	CAG	TTT	GAC	TAC	TCT	
O1	TTT	GGG	GAT	GGT	GGT	GTA	ATT	GTA	CAG	TTT	GAC	TAC	TCT	
	F	G	D	G	G	V	I	V	Q	F	D	Y	S	

Figure 3. Exon II sequence comparison. Nucleotide sequences of ϕ e, SP82 (82) and SPO1 (O1) extending from the 3' s.s. to the beginning of primer 2. (Sequence of primer 2 was not included for comparison because it was used for the amplification of all sequencing templates). Bold nucleotides indicate variable positions between the phage. The predicted amino acid sequence is shown beneath the SPO1 nucleotide sequence, and is the same for all three phage except where differences are indicated by bold letters. GenBank accession numbers: SPO1, M37686; SP82, U04812; ϕ e, U04813.

In contrast to the high sequence conservation within the intron core, the intron-ORFs have diverged considerably. The predicted amino acid sequences of the ORFs are aligned in Fig. 5. SPO1 has 43.6 and 43.1% identity with SP82 and ϕ e respectively, while the latter two are 69.9% identical. Interestingly, the *nucleotide* sequences of the three phage are identical from the beginning of L8 through the first base of the 14th codon of the ORFs, with only one exception in the ORF that does not alter the amino acid sequence. Identity decreases dramatically at this point and does not resume until the coding sequence overlaps the 3' portion of the essential P8 helix. The amino acids at the amino terminus may be especially crucial for protein activity, resulting in strict selective pressure to maintain the nucleotide sequence (note the high conservation with phage SPP1 gene 36.1 in this region). However, the lack of variation—even at third positions—suggests these nucleotides may also be important for RNA structure. For example, the sequence in this region might be constrained by the conformation required for splicing, or for optimal translational efficiency. With regard to the latter proposal, it is interesting to note that the optimal folding of this region predicted by *mfold* includes an unstructured ribosome binding region flanked by stem-loops (not shown).

The intron-ORFs were subjects of searches against a non-redundant compilation of protein sequence databases using default parameters on the BLAST network service at NCBI (27). One protein sequence, ORF 36.1 from the unrelated *B. subtilis* phage SPP1 (28) gave statistically significant alignments with all the

Figure 4. Intron secondary structure. Phylogenetically conserved structures P1 through P9 are shown for the SP82 intron, following our previously published model for SPO1 (1). Exon sequences are shown in lower case letters and intron sequences in upper case, with the 5' and 3' splice sites indicated by open arrows. Numbering of nucleotides indicates their distance from the first guanosine of the *EcoRI* site within exon I (see Fig. 2A). The ORF begins within L8 (AUG at position 443) and ends within P8 (underlined UGA). The SPO1 P9 stem loop (shown in the lower right corner) has nine nucleotide differences with SP82 that are indicated by asterisks. Variable nucleotides between SP82 and ϕ e are circled, with the ϕ e sequence indicated by small arrows. The two outlined stem loop structures between P3 and P4 can be drawn as a single interrupted stem loop shown boxed in the upper right corner.

HMU intron ORFs (Fig. 5). The function of the SPP1 ORF is not known but it seems likely to us that it is functionally equivalent to the HMU intron ORFs. Two of the HMU intron ORFs are site-specific DNA endonucleases (21, Goodrich-Blair and Shub, in prep.). In this context it may be noteworthy that, like the SP82 intron ORF, SPP1 gene 36.1 could not be cloned in *E. coli* without frame shift or other inactivating mutations (J.C. Alonso, personal comm.). Other than SPP1 gene 36.1, no protein sequence was found that was extensively similar to the HMU intron ORFs, indicating that there are few closely related proteins in the databases. There were, however, a number of weak matches to other bacterial and bacteriophage proteins. One of these (boxed in Fig. 5) corresponds to the recently described HI-NL-H motif, which is found in group I intron endonucleases, a bacterial restriction endonuclease, and the Zn-finger-like domains of several group II intron ORFs (29).

DISCUSSION

We have shown that the DNA polymerase genes of the *B. subtilis* HMU-phage SP82, ϕ e and 2C are interrupted by an intron similar to the one previously reported in SPO1 (1). We were unable, however, to find conclusive evidence for or against the presence of an intron in another isolate, SP8. All attempts to amplify the region of the SP8 DNA polymerase gene that contains the intron insertion site have been unsuccessful. Electron microscopy of our preparation of SP8 has confirmed that its morphology resembles that of other HMU-phage, which argues against the possibility that we accidentally cultured an unrelated *B. subtilis* phage. As DNA polymerase is an essential gene for the other HMU-phage, we conclude that SP8 has a DNA polymerase gene that is highly divergent, at least within the region studied here. We cannot, however, exclude the possibility that this gene harbors an intron that is incapable of self-splicing *in vitro*.

Although a data set of only five phage is not statistically representative, the presence of introns in the DNA polymerase genes of at least four of them suggests that the intron is commonly

found in this phage family. Analysis of additional isolates of HMU-phage will be required to determine if this is actually the case, and to allow comparison to the variable and rare distribution of introns among the T-even phage (2).

The four introns we have identified are extremely similar—even within regions that are not typically conserved between group I introns. In fact, SP82 and 2C are identical, and the others, excluding the ORFs, are identical except for a small number of differences in P6a, P7.2, and P9 and in the region between P3 and P4. All of the differences in P7.2 and all but two in P9 are compensatory, confirming there are functional constraints maintaining the existence of these structures. Although the variation among these introns has caused us to revise our proposal for the secondary structure between P3 and P4 (Fig. 4), our representation must still be considered tentative. Mutagenic analysis should provide insight into possible roles for this region in splicing, and may also aid in identifying interactions with other parts of the intron RNA. Finally, although the lack of variation in the 33 nucleotides between P7 and P7.2, gives no insight into the correct structure for P7.1, the high degree of similarity of other phage introns in this region supports the structure shown (5,30).

In contrast to the introns, the intron-ORFs have diverged considerably from one another, suggesting rapid evolution without strict selective pressure. The SPO1 intron-ORF is the furthest diverged: its predicted amino acid sequence has only 43% identity with ϕ e or SP82. Both the SPO1 and SP82 ORFs encode endonucleases (21, Goodrich-Blair and Shub, in prep.) and although the ϕ e intron-ORF protein product has not been tested, its similarity to SP82 strongly suggests it is also an endonuclease. As the SPO1 and SP82 endonucleases have different DNA targets, the high similarity of amino acid sequence in their amino termini may indicate structures involved in the active site, while the high variability in the carboxyl termini may indicate involvement in DNA site selection.

The high similarity of SPP1 ORF 36.1 to the HMU phage intron-ORFs, especially in the highly conserved amino terminal half, suggests that it is also an endonuclease. The HI-NL-H motif proteins (29) are the latest example of distantly related endonucleases, sharing a common amino acid motif, that are found in introns and also as free-standing genes (the other two best-known examples being the *dodecapeptide* and GIY-YIG motifs; see ref. 3 for a review). The distribution of these proteins is consistent with the hypothesis that introns and their ORFs are separately mobile elements (3,31–33).

In a DNA cross hybridization study, phage 2C could not be distinguished from SPO1 and SP82, while SP8 and ϕ e had 92% and 83% homology to 2C respectively (34). Another study (including only SPO1, SP82 and ϕ e) concluded that SPO1 and SP82 are approximately 80% homologous with one another, while ϕ e is at most 50% similar to either of them (35). Although the sequences presented here of the core intron and part of exon II are consistent with these previous findings—SPO1, SP82 and 2C have fewer differences with one another than any do with ϕ e—the overall similarity of the phage in this region is higher than indicated by the hybridization studies. In addition, the sequences of the intron-ORFs are not consistent with the hybridization data, as ϕ e and SP82 are much more similar to one another than they are to SPO1. These observations suggest that the HMU introns, and especially their ORFs, may be the result of a fairly recent transfer between closely related phage. Consistent with this are the facts that the intron-ORFs of SPO1

	**	*****	*****	*	+	++	**	+	**	+	+	+	+	+
phiE	ME..WKDIKG	YEGHYRISDN	GDIFSLKSNK	VLKTMNSN	KN	GYTYIHLTKG								
sp82	ME..WKDIKG	YEGHYQISDN	GDIFSLKSNR	VLKTMKNKKK	KN	GYTYIHLTKD								
sp01	ME..WKDIKG	YEGHYQVSNT	GEVYSIKSGK	TLKH.QIPKD		GYHRIGLFKG								
spp1	MEEVWRDIPG	YVGIYEVSTL	GRVRKKEGGR	ILKARKSR..		GYMRLWLKYD								
phiE	GKKKSFTIHR	LVALHFCGEG	GEDIVVDHID	QDRDNHCSN	LRWVSRKENS									
sp82	GKKKAFTHIR	LVALHFCGGY	EESIVVDHID	RNRHNNHFSN	LRWVSRKENS									
sp01	GKGKTFQVHR	LVALHFCGEG	EESIVVDHID	GNKDNNLSTN	LRWVTQKINV									
spp1	KLKKDWYVHR	LVALAFLCRP	PGKELINHID	GDKTNNYFKN	LEWCDHKENL									
phiE	+			+	+	+	+	+	+	+	+	+	+	+
sp82	NNISADTRAK	VSEVSKRNAR	KVSQMTGRKK	RPVISISPDP	VEKYHG...	S								
sp01	SNISADTKKN	IITAVRKNAR	SQSQRSNRKI	RPVISISPDP	VIRHHR...	S								
spp1	ENQMSRGTLN	VS.....KAQ	QIAKIKNQK.	PIIVISPDG	IEKEYP...	S								
	MH.....		AYMTGANK	SPIKVTLVDP	RTKKHLFPYS									
phiE	IMEASRYTGV	GNSSIWKALN	KGSVLASGYK	F...KHALNG										
sp82	ITQASKETGV	TNANISNNLN	KGSRLMSGYR	FLYKDRQLNG										
sp01	TKCACEELGL	TRGKVTDVLK	GHRIHKGYT	FRYK...LNG										
spp1	MGAASRFLGH	HSGYISRKVG	NGNFADGYI	IKLEA....										

Figure 5. Alignment of HMU intron-ORFs. The predicted amino acid sequences of the HMU intron ORFs and SPP1 ORF 36.1 (GenBank accession number X67865) were aligned using the PileUp program of Genetics Computer Group Inc. (GCG). Amino acid identities among all four sequences are marked by asterisks (*), while agreement of three sequences is indicated by a plus (+). Gaps introduced by the program to optimize the alignment are indicated by dots. The region corresponding to the HI-NL-H protein sequence motif (29) is boxed.

and SP82 encode endonucleases (21, Goodrich-Blair and Shub, in preparation) and that other intron-encoded endonucleases mediate intron mobility, with attendant co-conversion of flanking exon sequences (3).

ACKNOWLEDGEMENTS

We thank Richard Bonocora for technical assistance and Carole Keith for help with the figures. H.G.-B. was supported by an Arvid J. and Mary A. Burke Graduate Fellowship. This work was supported by grant GM37746 from the National Institutes of Health.

REFERENCES

- Goodrich-Blair, H., Scarlato, V., Gott, J.M., Xu, M.-Q., and Shub, D.A. (1990) *Cell*, **63**, 417–424.
- Eddy, S. (1992) Ph.D. thesis, Univ. of Colorado, Boulder.
- Lambowitz, A.M. and Belfort, M. (1993) *Ann. Rev. Biochem.*, **62**, 587–622.
- Cech, T.R. (1988) *Gene*, **73**, 259–271.
- Michel, F. and Westhof, E. (1990) *J. Mol. Biol.*, **216**, 585–610.
- Cech, T.R. (1990) *Annu. Rev. Biochem.*, **59**, 543–568.
- Mota, E.M. and Collins, R.A. (1988) *Nature*, **332**, 654–656.
- Garriga, G. and Lambowitz, A.M. (1984) *Cell*, **38**, 631–641.
- Gott, J.M., Shub, D.A., and Belfort, M. (1986) *Cell*, **47**, 81–87.
- Shub, D.A., Gott, J.M., Xu, M.-Q., Franz-Lang, B., Michel, F., Tomaschewski, J., Pedersen-Lane, J., and Belfort, M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 1151–1155.
- Sjöberg, B.M., Hahne, S., Mathews, C.Z., Mathews, C., Rand, K.N., and Gait, M.J. (1986) *EMBO J.*, **5**, 2031–2036.
- Young, P., Ohman, M., Xu, M.-Q., Shub, D.A., and Sjöberg, B.-M. (1994) *J. Biol. Chem.*, **269**, 20229–20232.
- Chu, F.K., Maley, G.F., Maley, F., and Belfort, M. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 3149–3153.
- Stewart, C. (1993) In A. Sonenshein (ed.), *Bacillus subtilis* and Other Gram-Positive Bacteria. American Society for Microbiology, Washington, DC, pp. 813–829.
- Green, D.M. (1964) *J. Mol. Biol.*, **10**, 438–451.
- Pène, J.J. and Marmur, J. (1964) *Fed. Proc.*, **23**, 318.
- Roscoe, D.H. and Tucker, R.G. (1966) *Virology*, **29**, 157–166.
- Romig, W.R. and Brodetsky, A.M. (1961) *J. Bacteriol.*, **82**, 135–141.
- Scarlato, V. and Gargano, S. (1992) *Gene*, **118**, 109–113.
- Gage, L.P. and Geiduschek, E.P. (1971) *J. Mol. Biol.*, **57**, 279–300.
- Goodrich-Blair, H. (1993) Ph.D. thesis, University at Albany, SUNY.
- Braithwaite, D.K. and Ito, J. (1993) *Nucleic Acids Res.*, **21**, 787–802.
- Goodrich, H.A., Gott, J.M., Xu, M.-Q., Scarlato, V., and Shub, D.A. (1989) In Cech, T.R. (ed.), *Molecular Biology of RNA*. Alan R. Liss, Inc., New York, pp. 59–66.
- Zucker, M., Jaeger, J.A., and Turner, D.H. (1991) *Nucleic Acids Res.*, **19**, 2707–2714.
- Trinkl, H. and Wolf, K. (1986) *Gene*, **45**, 289–297.
- Cummings, D.J., Michel, F., and McNally, K.L. (1989) *Curr. Genet.*, **16**, 381–406.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
- Weise, F., Chai, S., Lüder, G., and Alonso, J.C. (1994) *Virology*, **202**, 1046–1049.
- Shub, D.A., Goodrich-Blair, H. and Eddy, S.R. (1994) *TIBS*, **19**, in press.
- Bechhofer, D.H., Hue, K.K., and Shub, D.A. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, in press.
- Perlman, P.S. and Butow, R.A. (1989) *Science*, **246**, 1106–1109.
- Sharma, M., Ellis, R.L., and Hinton, D.M. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 6658–6662.
- Shub, D.A. and Goodrich-Blair, H. (1992) *Cell*, **71**, 183–186.
- Truffaut, N., Revet, B., and Soulie, M. (1970) *Eur. J. Biochem.*, **15**, 391–400.
- Lawrie, J.M., Downard, J.S., and Whitely, H.R. (1978) *J. Virol.*, **27**, 725–737.