Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 1151–1155, February 1988
Genetics

Structural conservation among three homologous introns of bacteriophage T4 and the group I introns of eukaryotes

(catalytic RNA/RNA processing/prokaryotic splice junctions/translation and splicing)

DAVID A. SHUB*, JONATHA M. GOTT*, MING-QUN XU*, B. FRANZ LANG†, FRANÇOIS MICHEL‡, JÖRG TOMASCHEWSKI§, JOAN PEDERSEN-LANE¶, AND MARLENE BELFORT¶

*Department of Biological Sciences, State University of New York at Albany, Albany, NY 12222; †Département de Biochimie, Université de Montréal, Montréal, PQ H3C 3J7, Canada; ‡Centre de Génétique Moléculaire du Centre National de la Recherche Scientifique, 91190 Gif-sur-Yvette, France; §Arbeitsgruppe Moleculare Genetik, Lehrstuhl Biologie der Mikroorganismen, Ruhr-Universität Bochum, D-4630 Bochum 1, Federal Republic of Germany; and ¶Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, NY 12201

Communicated by Thomas R. Cech, October 12, 1987

ABSTRACT **Three group I introns of bacteriophage T4 have been compared with respect to their sequence and structural properties. The introns include the *td* intervening sequence, as well as the two newly described introns in the *nrdB* and *sunY* genes of T4. The T4 introns are very closely related, containing phylogenetically conserved sequence elements that allow them to be folded into a core structure that is characteristic of eukaryotic group IA introns. Similarities extend outward to the exon sequences surrounding the three introns. All three introns contain open reading frames (ORFs). Although the intron ORFs are not homologous and occur at different positions, all three ORFs are looped-out of the structure models, with only the 3′ ends of each of the ORFs extending into the secondary structure. This arrangement invites interesting speculations on the regulation of splicing by translation. The high degree of similarity between the T4 introns and the eukaryotic group I introns must reflect a common ancestry, resulting either from vertical acquisition of a primordial RNA element or from horizontal transfer.**

The discovery of catalytic properties of RNA molecules has strengthened the notion that RNA was the primordial nucleic acid (1). These RNAs are exemplified by introns in fungal mitochondria and the large rRNA of *Tetrahymena* (reviewed in ref. 2). Placed into two groups based on conserved features of sequence and secondary structure (3), the group I and group II introns are excised by related yet distinct mechanisms, which can sometimes proceed autocatalytically (2). Self-splicing introns also exist in at least three genes of bacteriophage T4 (4–6). These T4 mRNA precursors are spliced by a mechanism resembling that used by group I introns of eukaryotes, involving 5′ guanosine addition and circularization of the intron (7–9). Since all of the information for both catalysis and specificity of splicing is contained in the precursor RNA, it is of interest to compare the structural properties of these functionally similar RNAs from such phylogenetically distant organisms. Examination of these structures may also give clues to possible regulatory roles played by these introns in gene expression.

RESULTS AND DISCUSSION

Splice Junction of the *nrdB* Gene. Gott *et al.* (5) described the existence of two T4 introns, one of which mapped to *nrdB*, the gene for the small subunit of ribonucleoside diphosphate reductase. Sjöberg *et al.* (6) have also inferred an intron in *nrdB* by nucleotide sequence analysis. By comparing homologies with the small subunit of ribonucleoside diphosphate reductase from *Escherichia coli*, they concluded that two coding sequences were separated by about 600 base pairs (bp) of nonhomologous DNA. This corresponded closely to the size determined for the intron that could be autocatalytically excised from RNA from the *nrdB* gene (5).

We have determined the precise boundaries of the RNA-processing event by primer-extension sequencing of RNA extracted from cells infected by phage T4 (Fig. 1*A*). This RNA contains both precursor and mature mRNA species. With a primer complementary to a region near the beginning of exon 2, a unique sequence was obtained until position 1799, the 3′ splice site. After this the contiguous sequence of the pre-mRNA was superimposed on the sequence of the mRNA extending back from position 1200 (6). Thus, the intron comprises 598 bp, in excellent agreement with our estimate of 0.6 kbp obtained from gel migration of excised intron RNA (5). This splice junction, which differs from that initially proposed by Sjöberg *et al.* (6), has been confirmed for the purified *in vitro* ligation product (data not shown). The intron boundaries are in accord with typical group I introns: splicing occurs after a uridine in exon I and after a guanosine at the 3′ end of the intron (3, 13).

As noted by Sjöberg *et al.* (6), the intron occurs in a region where the T4 protein displays high homology with the equivalent proteins from *E. coli* and eukaryotes (Fig. 1*B*) (10–12). However, rather than having a deletion of eight amino acids as previously predicted (6), our splice-junction data indicate that the T4 protein has an insertion of one amino acid. A new codon (UGU) is created at the splice junction, placing cysteine at a position homologous with the *E. coli* protein (10) (Fig. 1*B*).

Boundaries of the Third T4 Intron. Previously, a third T4 intron was mapped to a 4.9-kb *Xba* I restriction fragment via hybridization with GTP-end-labeled T4 RNA (5). We have further localized the intron to a 1.3-kb *Acc* I–*Xba* I subfragment (data not shown). Sequence determination confirmed that the intron falls within a region between genes *49* and *55* that contains 13 unidentified open reading frames (ORFs) (14). We provisionally call this gene *sunY* (*s*plit gene, *un*known function, *why?*).

In vitro transcription products from this region were typical of autocatalytic processing events (unpublished data) and revealed an intron of 1.0–1.1 kbp. Fig. 2 shows the sequence of the *sunY* splice junction, which defines a 1033-nucleotide (nt) intron that also conforms to the group I boundary rule, beginning after a uridine and ending with guanosine. Splicing results in the in-frame fusion of two coding regions identified by Tomaschewski and Rüger (14), with an ORF falling within the intron. A protein of the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: IGS, internal guide sequence; ORF, open reading frame; nt, nucleotide(s).

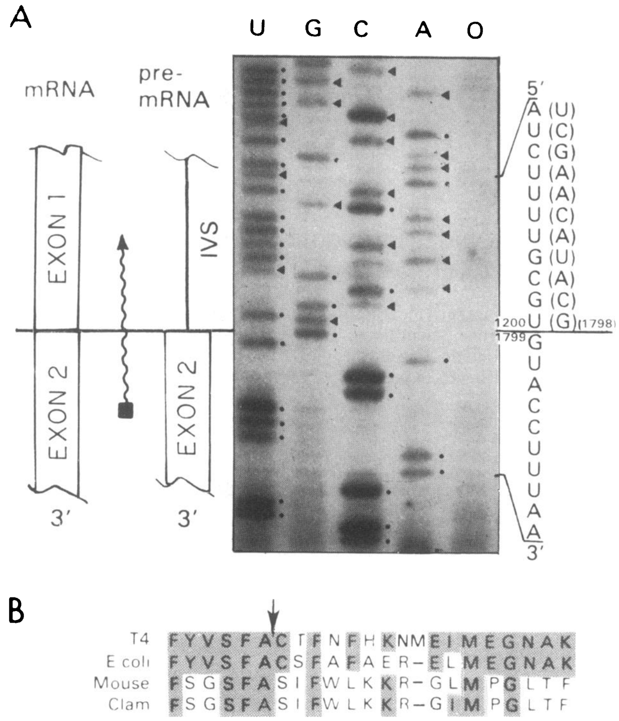

A
U G C A O
mRNA
pre-mRNA
EXON 1
IVS
EXON 2
EXON 2
3′
3′
5′
A (U)
U (C)
C (G)
U (A)
U (A)
U (C)
U (A)
G (U)
C (A)
G (C)
1200 U (G) (1798)
1799 G
U
A
C
C
U
U
U
A
A
3′
B
T4 F Y V S F A C T F N F H K N M E I M E G N A K
E coli F Y V S F A C S F A F A E R – E L M E G N A K
Mouse F S G S F A S I F W L K K R – G L M P G L T F
Clam F S G S F A S I F W L K K R – G I M P G L T F

FIG. 1. *nrdB* splice junction. (*A*) *nrdB* mRNA sequence. An exon 2-specific oligonucleotide (black square) was used to prime cDNA synthesis (wavy arrow) from RNA extracted from *E. coli* B that had been infected with T4 for 12 min at 30°C (7). Lanes are labeled with the complement of the dideoxynucleoside triphosphate. The primer was complementary to residues 1864–1885 (6). The horizontal line is at the splice junction, above which the sequence of the mRNA (dots) and pre-mRNA (i.e., intron, triangles) are interdigitated. The exon junction sequence is shown on the right, with the intron sequence after the 3′ splice site in parentheses. (*B*) Amino acid sequence (standard one-letter symbols) at the splice junction. The deduced amino acid sequence across the T4 *nrdB* splice junction (arrow) is compared with the ribonucleoside diphosphate reductase small subunit sequence of *E. coli* (10), mouse (11), and clam (12).

predicted size for the spliced exon product is made during infection but is lacking when the intron is mutationally disrupted, indicating that *sunY* is a functional T4 gene (A. Zeeh and D.A.S., unpublished data).

Predicted Secondary Structures of the T4 Introns and Their Relationship to Eukaryotic Introns. Detailed predictions of a group I core structure, consisting of both local and long-range complementary base-pairing regions, have been made (3, 15, 16). This model has been consistent with the sequence of every group I intron determined subsequently (13, 17). The three T4 introns bear a remarkable resemblance, both in secondary structure (Fig. 3) and in conserved sequence elements (Fig. 4), to the group I introns of eukaryotes. They belong to a subset, group IA, that is characterized by systematic variations in the highly conserved sequences P, Q, R, and S (Fig. 4) and by one (or, rarely, two) extra stem–loop structures between the P7 and P3 pairings (the P7.1 and P7.2 pairings in Fig. 3) (3, 13).

Identification of P and Q sequences (parts of which pair to form the P4 stem) and the R and S sequences (parts of which pair to form the P7 helix) is critical to generating the correct core structure. Although our assignments of P and Q sequences deviate from the consensus, compensatory changes allow the two sequences to pair in the expected way for each of the T4 introns (Fig. 4). Our structure for the *td* intron predicts P and Q sequences that are different from those originally assigned by Chu *et al.* (8) but agree with their recent reassignments (9).

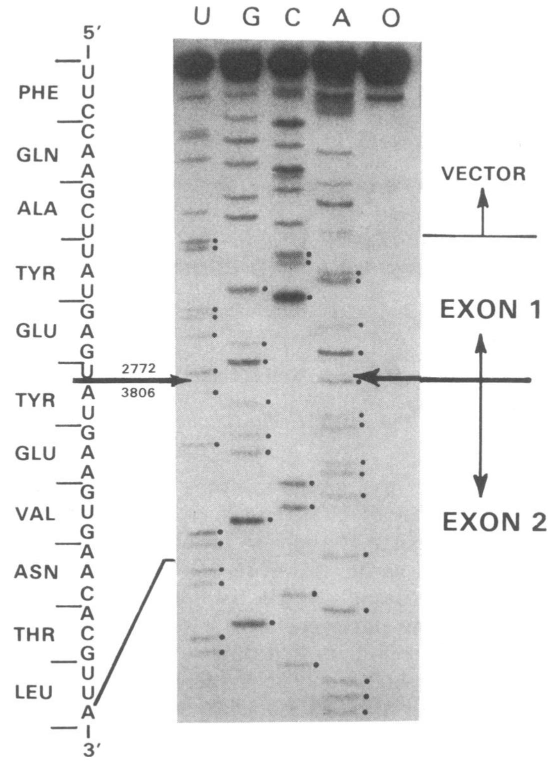
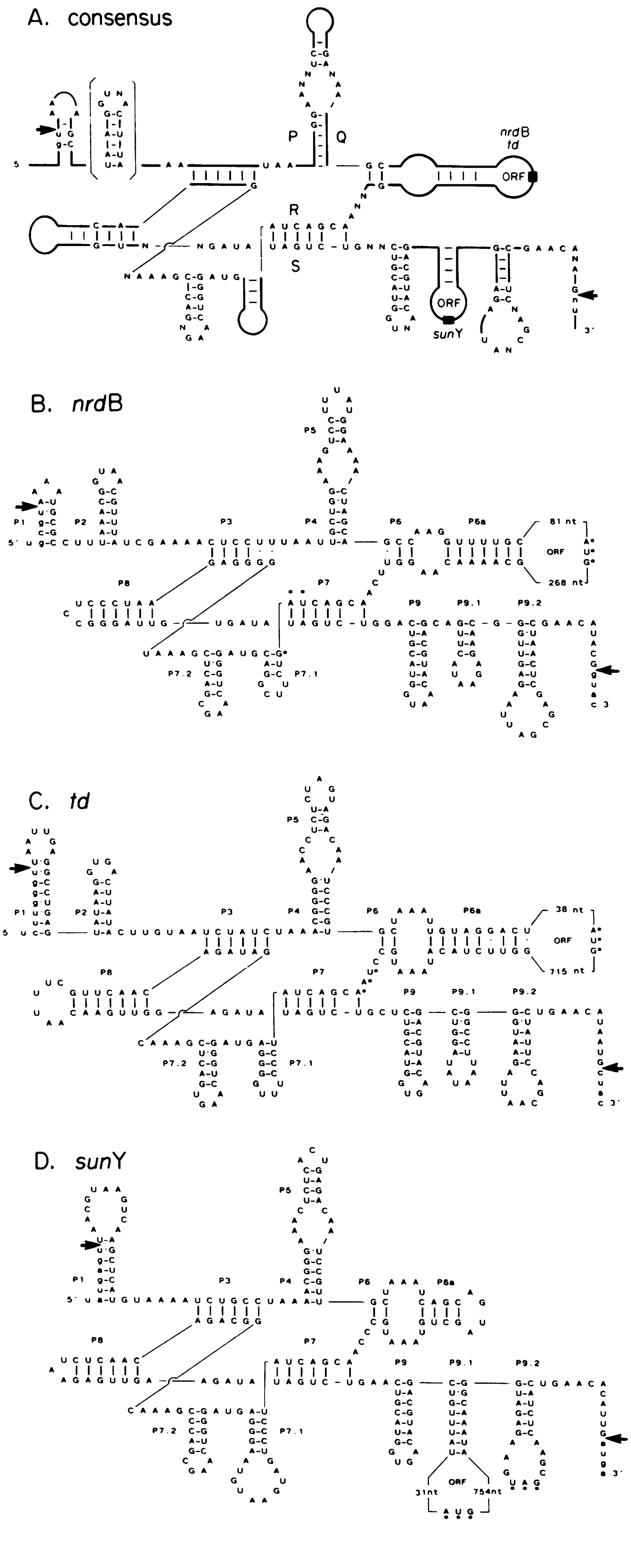


FIG. 2. *sunY* splice junction. RNA was synthesized with phage T3 RNA polymerase, using pMAX1 [a 1.3-kb *Acc* I–*Xba* I fragment that includes the entire *sunY* intron in pBSM13+ (5, 14)] as template. Numbering of residues (on the left) is according to ref. 14. An exon 2-specific oligonucleotide, complementary to residues 3865–3886, was used to prime cDNA synthesis from purified ligated exons. Lanes are labeled as in Fig. 1, with horizontal arrows at the splice junction. The deduced amino acid sequence of *sunY* appears on the left.

The T4 introns conform to the group I intron core structure not only with respect to the central P3, P4, and P7 pairings, but also with respect to the intron–exon alignment (P1 pairing). The internal guide sequence (IGS), near the 5′ end of the intron, has been proposed to precisely align the 5′ splice site by pairing with the upstream exon, with the terminal uridine of exon 1 always paired with a guanosine in the IGS. These requirements are exactly met in all three intron models (Fig. 3).

Fig. 3 also shows the striking similarities among the T4 introns. First, each contains two stem–loop structures (P7.1 and P7.2) between the R sequence of P7 and the P3 pairing, consistent with their status as group IA introns. Second, the three introns have in common two additional stem–loops after P9. Third, there is extensive conservation of sequences at equivalent positions of the core structure (Fig. 3*A*), extending well beyond the sequences that are conserved in all group I introns. Among these conserved regions are the P7.2 stem–loop and its unpaired flanking regions, the P9 and P9.2 stem–loop structures, and the unpaired residues extending to the 3′ end of the intron. While this homology clearly reflects the common ancestry of the T4 introns, it is interesting that these sequences have diverged less than other structural elements (e.g., P3 and P4) that are central to the core structure of group I introns. This implies a common function for the conserved T4 sequences. The only major structural differences among the T4 introns are the positions of the intron ORFs and the lack of the P2 stem in the *sunY* intron. P2 is also an optional element in eukaryotic group I introns (3, 13).

A significant degree of sequence conservation also exists beyond the intron boundaries (Fig. 5), extending 10–12 nt to



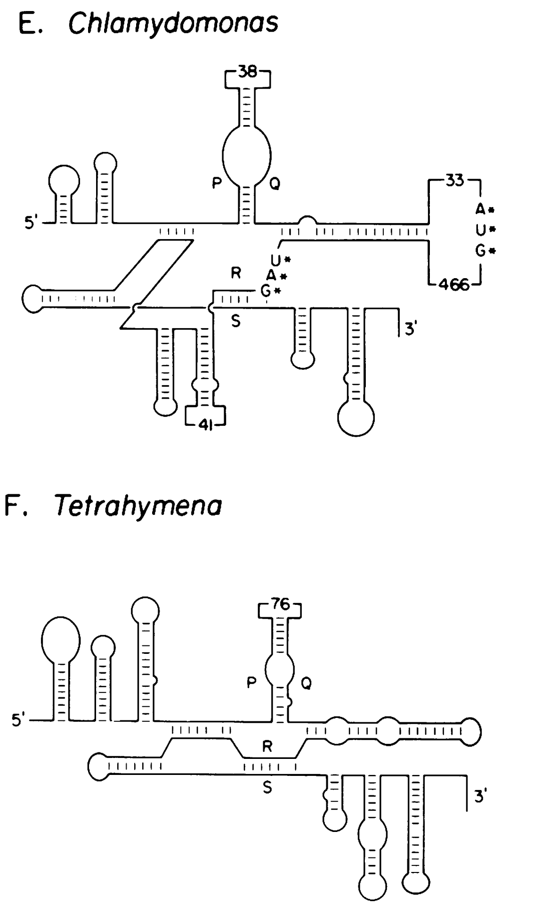


FIG. 3. Secondary structure models. (*A*) Schematic representation of the sequence and structural elements conserved among the T4 introns. Nucleotides common to RNAs from each of the three intron-containing T4 genes are shown, with conserved spacings between them indicated by Ns and conserved base pairings indicated by lines between the phosphodiester backbone. Thin lines represent connections between adjacent nucleotides, whereas thicker lines indicate varying lengths of nonconserved nucleotides. P, Q, R, and S are the highly conserved sequences among group I introns (refs. 3 and 13; see Fig. 4). Splice sites are indicated by arrows between the exon (lowercase) and intron (uppercase) nucleotides. Black squares represent the intron ORFs, which extend from stems that are conserved in structure but not in sequence. Although conserved between the *td* and *nrdB* genes, the P2 stem is absent from the *sunY* gene and is shown in brackets. (*B–D*) Predicted secondary structures for intron sequences from the *nrdB*, *td*, and *sunY* genes, respectively. The pairings P1 through P9.2 refer to structural elements characteristic of group I introns (3, 13). The nomenclature and intron displays are according to recently revised conventions (18). The initiator AUG and stop codons of each intron ORF are indicated by asterisks. Numbers indicate the number of nucleotides between given structural elements. (*E* and *F*) Schematic illustrations of the phosphodiester backbone and base pairings predicted for the rRNA introns from *Chlamydomonas reinhardii* (19) and *Tetrahymena thermophila* (3, 16). P, Q, R, and S and the position of the *C. reinhardii* intron ORF are shown.

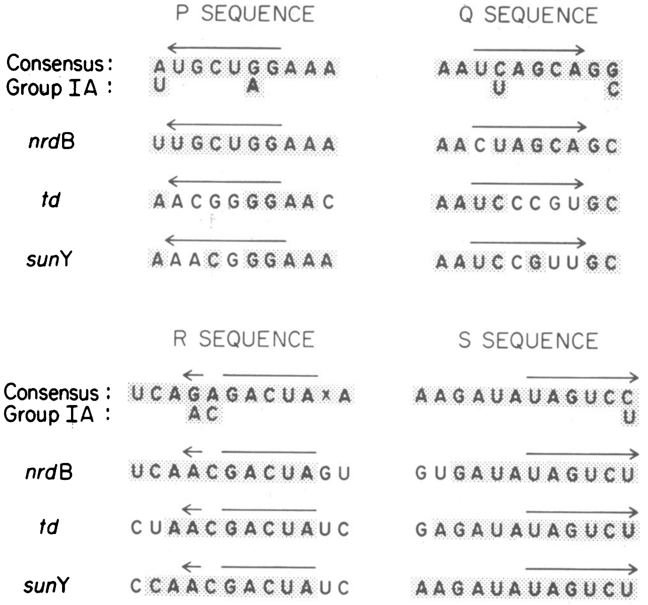


FIG. 4. Group I intron conserved sequences. Highly conserved sequences of group I introns are shown, with the variations characteristic of group IA given below (3). Bases involved in pairings (Fig. 3) are indicated by arrows (13, 17). The portions of the T4 sequences that match the consensus are shaded.

each side of the splice junction. This may reflect a target site for intron insertion and/or interactions with the core structure of the intron.

Although many group I introns contain ORFs looped out from the core structure, the position of the ORFs within the T4 introns (P6 and P9.1) is not typical. In addition, unlike many introns of this type, the T4 intron ORFs overlap structural elements thought to be involved in splicing. Both the *td* and *nrdB* intron ORFs extend past the P6 stem, ending within the highly conserved R sequence, whereas the *sunY* intron has an ORF that overlaps P9.2, the last structural element of the intron. These arrangements are not without precedent, however. An intron in the large rRNA gene of *Chlamydomonas reinhardii* chloroplasts (19) bears a remarkable structural resemblance to the *td* and *nrdB* introns (Fig. 3), sharing with them all of the unusual features listed above. Similarly, the placement of the ORF within the *sunY* intron resembles that in three fungal mitochondrial introns (17, 20).

Although the sequences of the T4 introns (outside of the conserved group I elements) are not similar to their eukaryotic counterparts, the striking structural similarities between the prokaryotic and eukaryotic introns suggest a common evolutionary origin. These structural features are likely to reflect common functions carried out by RNA molecules from these widely divergent sources.

Support for the Structure Models. Although these structure models remain to be tested rigorously, mutational analysis of the *td* intron is fully consistent with the proposed structures. First, based on the distribution of 30 nondirected splicing-defective mutations, two functional domains have been defined (21, 22). The 5′ and 3′ domains encompass the first 90 and last 160 nt of the *td* intron (22), exactly those regions folded into the independently derived core structure. Second, a 631-nt intron deletion removing the major central portion of the ORF is splicing-proficient (7, 22), in agreement with the exclusion of most of the ORF from the core structure. Third, one splicing-defective mutation lies within the ORF, immediately 5′ to the stop codon (22). The location of this mutation supports the involvement of the 3′ end of the ORF in the active splicing conformation, exactly as predicted by the model.

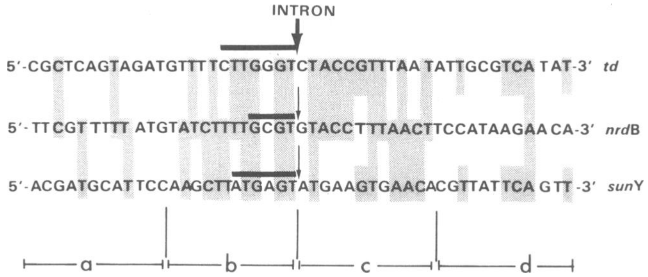

INTRON
5′-CGCTCAGTAGATGTTTTCTTGGGTCTACCGTTTAA TATTGCGTCA TAT-3′ td
5′-TTCGTTTTT ATGTATCTTTTGCGTGTACCTTTAACTTCCATAAGAACA-3′ nrdB
5′-ACGATGCATTCCAAGCTTATGAGTATGAAGTGAACACGTTATTCA GTT-3′ sunY
a
b
c
d

FIG. 5. Exon sequences flanking the three T4 introns. Sequence identities surrounding the splice junctions (arrow) are shaded. Identical nucleotides in equivalent positions are found at 8/24 residues surrounding the splice junction (b + c), as opposed to 1/24 three-way identities in the flanking regions (a + d). Solid bars indicate exon sequences that form the P1 helix by pairing with the respective IGS.

ORFs Contained Within the T4 Introns. Like several other group I introns (3, 13), the T4 introns contain ORFs bounded entirely by core structure elements. In contrast to the extensive conservation in the core structure sequences, the ORFs are unrelated in both nucleic acid and protein sequence. Several considerations suggest that these ORFs are functional *in vivo*. First, their codon utilization is typical of T4 genes. A sensitive indicator of functional coding sequences is provided by the relative frequencies of each base at each position within the codons (23). Table 1 compares a compilation of 11,370 codons of T4 genes (G. Stormo, personal communication) with the intron-containing T4 genes. The ORFs created by the ligated exon sequences, as well as the three intron ORFs, reflect the distinctive T4 pattern, with highly significant correlation coefficients of 0.97 and 0.92, respectively (23).

The second feature suggesting that these ORFs are expressed is the presence of a sequence characteristic of T4 late promoters (24) preceding each ORF. In addition, the *nrdB* intron ORF is preceded by a perfect match to the T4 middle promoter consensus sequence (25). We have confirmed that these promoters are used in T4 infection (unpublished observations).

Finally, the predicted amino acid sequences of the ORFs have ordered properties characteristic of functional proteins. For example, the 291-nt *nrdB* ORF has two repeated domains containing the motif: Cys-Xaa-Xaa-Cys-(12 or 13 amino acids)-His-Xaa-Xaa-Xaa-Cys. This is an example of a repeated structural motif that binds divalent metal ions (e.g., Zn^{2+}) tightly and occurs frequently in nucleic acid-binding proteins (reviewed in ref. 26). The *td* and *sunY* ORF prod-

Table 1. Frequencies of bases at each position of T4 codons

Sequences compiled	Position	Frequency			
		Ade	Cyt	Gua	Ura
T4 genes* (11,370 codons)	1	0.32	0.32	0.55	0.20
	2	0.37	0.42	0.24	0.31
	3	0.30	0.26	0.21	0.49
nrdB, *sunY*, and *td* ligated exons (1279 codons)	1	0.33	0.38	0.50	0.21
	2	0.38	0.38	0.26	0.31
	3	0.29	0.24	0.24	0.48
nrdB, *sunY*, and *td* intron ORFs (600 codons)	1	0.34	0.33	0.46	0.26
	2	0.37	0.44	0.30	0.26
	3	0.29	0.22	0.24	0.48

*Excluding *nrdB*, *sunY*, and *td*.

ucts are predicted to be very basic, consistent with the possibility that they may also be nucleic acid-binding proteins. In addition, a sequence within the *td* ORF is highly homologous to ORFs (of unknown function) in mitochondrial introns of filamentous fungi (27). The large size of the *td* and *sunY* ORFs, 735 and 774 nt, respectively, also argues strongly for their functional importance in T4.

Although they seem to be functional genes that are likely to be expressed from their own promoters, expression of the T4 intron ORFs from the polycistronic pre-mRNAs made early after infection seems unlikely, since each sequence presents barriers to efficient translation. For *nrdB*, the first AUG codon of the intron ORF is preceded by a Shine–Dalgarno sequence of only 3 nt, with a spacing of only 3 nt between the putative initiation codon and the Shine–Dalgarno sequence. Such short spacings are not found in efficiently utilized ribosome initiation sites (28). The *td* (8) and *sunY* (unpublished observations) intron ORFs have Shine–Dalgarno sequences sequestered in an extensive secondary structure. Transcription from the late promoter sequences should free these ORFs for translation.

Translation of the *td* and *nrdB* intron ORFs from the pre-mRNA would disrupt the long-range R–S pairing (P7), which is essential for splicing, whereas translation of the *sunY* ORF would disrupt the P9.2 helix. The existence of two splicing-defective *td* mutations within the P9.2 stem–loop suggests that this structural element is essential for the splicing of T4 introns (22). Thus, translation of the intron ORFs from the pre-mRNA would presumably block splicing.

Regulatory Implications of the Intron Structure. The variable occurrence of these introns in the T-even phages (ref. 29 and unpublished data) implies that they are not essential for phage viability. Nevertheless, the streamlined nature of phage genomes and the rarity of introns in prokaryotic mRNA suggest that when introns do occur they should perform an important (presumably regulatory) role. It has been suggested that the requirement for guanosine (or one of its 5′ phosphates) as a substrate in the splicing reaction could bring the genes containing these introns under general growth-rate regulation (5). We now propose that the structural organization of the T4 introns presents another opportunity for regulation.

Several late T4 genes are initially transcribed on an early polycistronic message. Early translation is prevented by sequestration of translation signals in secondary structures involving sequences absent from the late, monocistronic mRNA (30, 31). It has been proposed that this gene arrangement allows the abnormal early expression of those genes under conditions of cellular stress (31). If the T4 intron ORFs were regulated in a similar manner, their translation would prevent the splicing and expression of the genes that contain introns. Since two of these genes (*td* and *nrdB*) provide precursors for DNA synthesis in excess of those provided by the host cell, their expression should be dispensable under just such suboptimal growth conditions.

It is perhaps not coincidental that the occurrence of genes within introns is limited to eubacteria, mitochondria, and chloroplasts, where no physical barrier exists between transcription and translation. The presence of introns in genomes where transcription and translation are coupled allows the regulatory coupling of splicing and translation, in a manner analogous to the attenuation of transcription in bacteria. We have proposed a negative regulatory role for this coupling in three genes in phage T4 and one in *Chlamydomonas* but would not be surprised if ribosomal movement plays a positive role in the splicing of some mitochondrial genes where translation of the intron is required for splicing (32).

We have described a remarkable similarity in conserved elements of both sequence and structure between three introns in bacteriophage T4 and the group I introns of eukaryotes. While these homologies undoubtedly reflect a common ancestry, we cannot tell whether the conservation stems from the preservation of an ancient molecular design that existed before the divergence of eukaryotes and prokaryotes or from a more recent horizontal gene transfer.

We thank Gary Stormo for T4 codon-utilization data, Gisela Mosig for preliminary sequence from the *sunY* region and for electron microscopy data that suggest the production of circular RNA, and P. Scott Chandry and John Burke for help in displaying the intron structures. J.M.G. received support from a training grant from the U.S. Army Research Office. This work was supported by National Institutes of Health Grants GM37746 (to D.A.S.) and GM33314 and GM39422 (to M.B.), National Science Foundation Awards DMB8609066 (to D.A.S.) and DMB8502961 and DMB850552 (to M.B.), and grants from the Deutsche Forschungsgemeinschaft to B.F.L. and to Wolfgang Rüger in support of J.T.

1. Gilbert, W. (1986) *Nature (London)* **319**, 618.
2. Cech, T. R. (1986) *Cell* **44**, 207–210.
3. Michel, F. & Dujon, B. (1983) *EMBO J.* **2**, 33–38.
4. Chu, F. K., Maley, G. F., Maley, F. & Belfort, M. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 3049–3053.
5. Gott, J., Shub, D. A. & Belfort, M. (1986) *Cell* **47**, 81–87.
6. Sjöberg, B.-M., Hahne, S., Mathews, C. Z., Mathews, C. K., Rand, K. N. & Gait, M. J. (1986) *EMBO J.* **5**, 2031–2036.
7. Ehrenman, K., Pedersen-Lane, J., West, D., Herman, R., Maley, F. & Belfort, M. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 5875–5879.
8. Chu, F. K., Maley, G. F., West, D. K., Belfort, M. & Maley, F. (1986) *Cell* **45**, 157–166.
9. Chu, F. K., Maley, G. F. & Maley, F. (1987) *Biochemistry* **26**, 3050–3057.
10. Carlson, J., Fuchs, J. A. & Messing, J. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 4294–4297.
11. Thelander, L. & Berg, P. (1986) *Mol. Cell. Biol.* **6**, 3433–3442.
12. Standart, N. M., Bray, S. J., George, E. L., Hunt, T. & Ruderman, J. V. (1985) *J. Cell Biol.* **100**, 1968–1976.
13. Waring, R. B. & Davies, R. W. (1984) *Gene* **28**, 277–291.
14. Tomaschewski, J. & Rüger, W. (1987) *Nucleic Acids Res.* **15**, 3632–3633.
15. Davies, R. W., Waring, R. B., Ray, J. A., Brown, T. A. & Scazzocchio, C. (1982) *Nature (London)* **300**, 719–724.
16. Cech, T. R., Tanner, N. K., Tinoco, I., Jr., Weir, B. R., Zuker, M. & Perlman, P. S. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3903–3907.
17. Michel, F. & Cummings, D. J. (1985) *Curr. Genet.* **10**, 69–79.
18. Burke, J. M., Belfort, M., Cech, T. R., Davies, R. W., Schweyen, R. J., Shub, D. A., Szostak, J. W. & Tabak, H. F. (1987) *Nucleic Acids Res.* **15**, 7217–7221.
19. Rochaix, J. D., Rahire, M. & Michel, F. (1985) *Nucleic Acids Res.* **13**, 975–984.
20. Waring, R. B., Brown, T. A., Ray, J. A., Scazzocchio, C. & Davies, R. W. (1984) *EMBO J.* **3**, 2121–2128.
21. Hall, D. H., Povinelli, C. M., Ehrenman, K., Pedersen-Lane, J., Chu, F. & Belfort, M. (1987) *Cell* **48**, 63–71.
22. Belfort, M., Chandry, P. S. & Pedersen-Lane, J. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, in press.
23. Stormo, G. D. (1987) in *Nucleic Acid and Protein Sequence Analysis, A Practical Approach*, eds. Bishop, M. J. & Rawlings, C. J. (IRL, Oxford), pp. 231–258.
24. Kassavetis, G. A., Zentner, P. G. & Geiduschek, E. P. (1986) *J. Biol. Chem.* **261**, 14256–14265.
25. Guild, N., Gayle, M., Sweeney, R., Walker, T., Modeer, T. & Gold, L. (1987) *J. Mol. Biol.*, in press.
26. Berg, J. M. (1986) *Science* **232**, 485–487.
27. Michel, F. & Dujon, B. (1986) *Cell* **46**, 323.
28. Stormo, G. D., Schneider, T. D. & Gold, L. M. (1982) *Nucleic Acids Res.* **10**, 2971–2996.
29. Pedersen-Lane, J. & Belfort, M. (1987) *Science* **237**, 182–184.
30. Macdonald, P. M., Kutter, E. & Mosig, G. (1984) *Genetics* **106**, 17–27.
31. McPheeters, D. S., Christensen, A., Young, E. T. & Gold, L. (1986) *Nucleic Acids Res.* **14**, 5813–5826.
32. Lazowska, J., Jacq, C. & Slonimski, P. P. (1980) *Cell* **22**, 333–348.