Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 3557–3561, June 1985
Biochemistry

Tobacco chloroplast tRNA^{Lys}(UUU) gene contains a 2.5-kilobase-pair intron: An open reading frame and a conserved boundary sequence in the intron

(molecular cloning/DNA sequence/blot hybridization/precursor RNA/codon usage)

MAMORU SUGITA*, KAZUO SHINOZAKI†, AND MASAHIRO SUGIURA†‡

*Department of Botany, Hokkaido University, Sapporo 060, Japan; and †Department of Biology, Faculty of Science, Nagoya University, Chikusa, Nagoya 464, Japan

Communicated by Motoo Kimura, January 28, 1985

ABSTRACT **The nucleotide sequence of a tRNA^{Lys}(UUU) gene on tobacco (*Nicotiana tabacum*) chloroplast DNA has been determined. This gene is located 215 base pairs upstream from the gene for the 32,000-dalton thylakoid membrane protein on the same DNA strand and has a 2526-base-pair intron in the anticodon loop. The intron boundary sequence does not follow the G-U/A-G rule but is similar to those of tobacco chloroplast split genes for tRNA^{Gly}(UCC) and ribosomal proteins L2 and S12. The intron contains one major open reading frame of 509 codons. The codon usage in the open reading frame resembles those observed in the genes for tobacco chloroplast proteins so far analyzed. The primary transcript of this tRNA gene is 2.7 kilobases long.**

Chloroplast genomes code for all components of rRNAs and probably for a complete set of tRNAs necessary for the protein synthesis in chloroplasts. Chloroplast tRNAs show high sequence homology with prokaryotic tRNAs (1). However, chloroplast tRNA genes show both prokaryotic and eukaryotic features (2). They contain sequences similar to prokaryotic tRNAs in their structural parts except for 3′ CCA ends and to prokaryotic "Pribnow boxes" and "−35 regions" in their upstream regions. Some tRNA genes from chloroplasts in plants contain introns (3–10). Their introns are very long [451–949 base pairs (bp)] compared with those of nuclear tRNA genes (13–60 bp) (11).

Recently it was found that tobacco chloroplast tRNA^{Gly}(UCC) gene contains a 691-bp intron in the D stem (7). The intron sites of all other tRNA genes so far analyzed were located in the anticodon loops (1). Here we present the nucleotide sequence of a tRNA^{Lys} gene on tobacco chloroplast DNA. This gene contains a 2526-bp intron in the anticodon loop. Some of the chloroplast tRNA genes have, therefore, very unique structures.

MATERIALS AND METHODS

Recombinant plasmid pTB11, which contains 10.2-kbp *Bam*HI partial fragments of *Nicotiana tabacum* (var. Bright Yellow 4) chloroplast DNA, was constructed as described by using pBR322 (12). Sequencing and blot hybridization were performed as described (13).

RESULTS AND DISCUSSION

The DNA Sequence. Tobacco chloroplast DNA is a circular molecule with a size of about 160 kbp that contains a 26-kbp inverted repeated sequence; the repeats are separated by a small (20-kbp) and a large (90-kbp) single-copy region. To

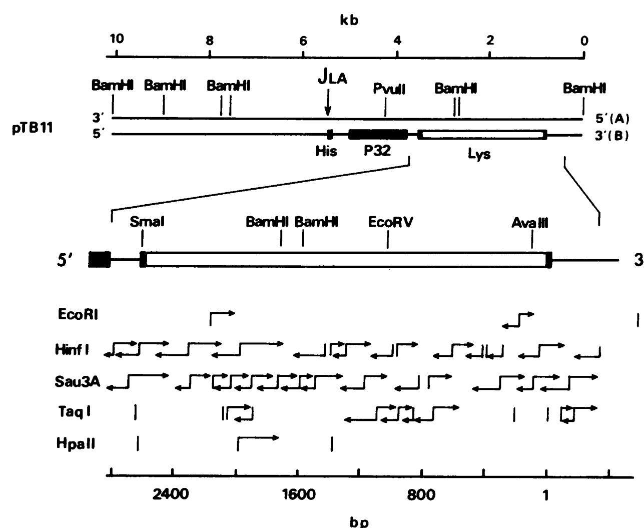

FIG. 1. (*Upper*) Physical map of the cloned 10.2-kbp partial *Bam*HI fragment from tobacco chloroplast DNA and the strategy for sequencing part of it. Strand B codes for tRNA^{His} (His), P32, and tRNA^{Lys} (Lys). J_{LA} is the junction between the right segment of the 26-kbp inverted repeated sequence (the left side) and the large single-copy region (the right side). Coding regions are shown by thick lines and introns by boxes. (*Lower*) Expanded physical map of the tRNA^{Lys} gene region. Vertical lines indicate restriction sites. Horizontal arrows show directions and extents of the DNA segments analyzed.

study the tRNA genes, we have cloned tobacco chloroplast DNA fragments produced by partial digestion with *Bam*HI followed by size fractionation (2, 13). One of the recombinant plasmids, pTB11, contains 1.1-, 1.2-, 0.2-, 4.8-, 0.14-, and 2.8-kbp *Bam*HI fragments of tobacco chloroplast DNA in this order in pBR322. The inserted sequence has been shown to contain the junction (J_{LA}) between the right segment of the inverted repeated sequence (IR_A) and the large single-copy region, the tRNA^{His}(GUG) gene (14), and the gene for the 32,000-dalton thylakoid membrane protein (P32) (13) in its 4.8-kbp *Bam*HI fragment (Fig. 1). On digestion with *Pvu* II, the 4.8-kbp *Bam*HI fragment yields 3.2- and 1.6-kbp subfragments. Total tobacco chloroplast tRNA hybridized to both subfragments (data not shown). The 3.2-kbp subfragment was shown to contain the tRNA^{His}(GUG) gene and the COOH-terminal half of the P32 gene, and the 1.6-kbp subfragment was shown to contain the NH_2-terminal half of the P32 gene. We then sequenced the region upstream from the P32

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: bp, base pairs; P32, 32,000-dalton thylakoid membrane protein.
‡To whom reprint requests should be addressed.

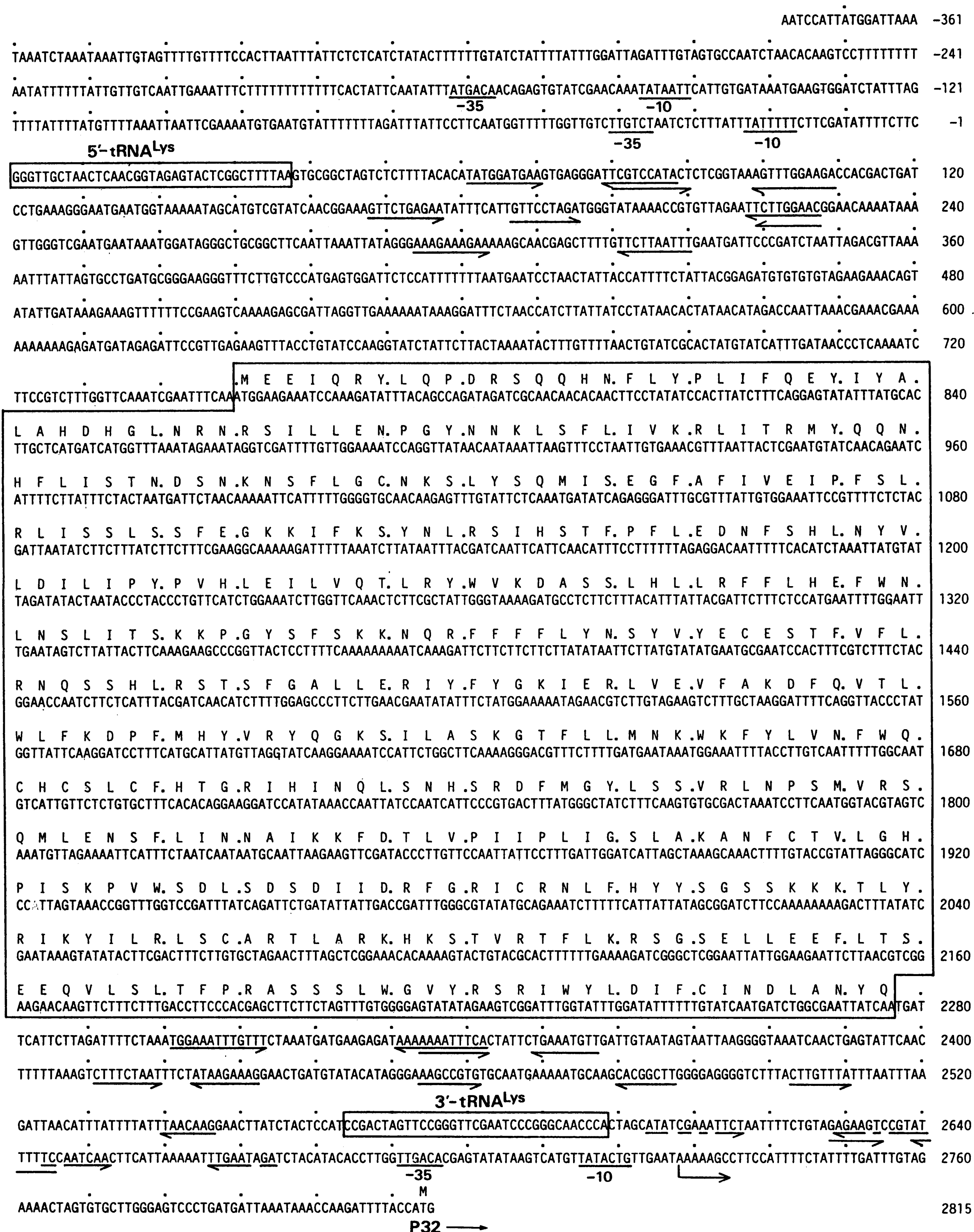

FIG. 2. DNA sequence of the 3193-bp region containing the tRNA^{Lys} gene. The tRNA-like strand (strand A in Fig. 1) is presented. The tRNA-like sequences (5′-tRNA^{Lys} and 3′-tRNA^{Lys}) and the open reading frame of 509 codons are boxed. The deduced amino acid sequence is shown above the DNA sequence. Sequences similar to Pribnow boxes ("−10") and −35 regions ("−35") are underlined. Horizontal arrows indicate inverted repeats. →, Transcription initiation site of the P32 gene. The sequence between positions 2513 and 2815 has been reported (13).

gene by the strategy shown in Fig. 1. Fig. 2 shows the sequences of a 3193-bp portion (the right *Hin*fI site to the putative ATG initiation codon of the P32 gene, see Fig. 1).

The tRNA^{Lys} Gene. A sequence corresponding to a 3′ half of a tRNA gene was found between positions 2564 and 2598 and was 215 bp upstream from the P32 gene in the same orientation as the P32 gene in the 1.6-kbp *Pvu* II–*Bam*HI subfragment (Fig. 2). As this sequence shows a substantial homology with the 3′ half of *Escherichia coli* tRNA^{Lys}(UUU), it seems to be a part of a gene for tRNA^{Lys}.

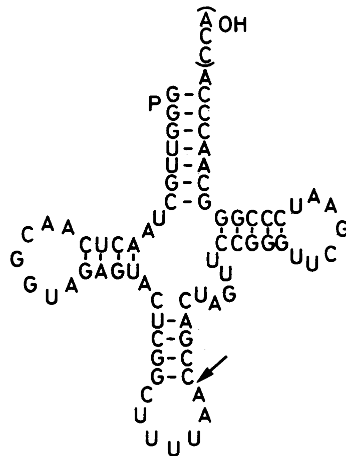

FIG. 3. Sequence of unmodified bases and cloverleaf structure of the tRNA^{Lys} predicted from the DNA sequence. An arrow indicates a possible intron site.

However, the sequence immediately before position 2564 was not able to form a cloverleaf structure. We then searched for a sequence complementary to the aminoacyl stem (positions 2591–2597, G-C-A-A-C-C-C). Surprisingly, we could find the complementary sequence (G-G-G-T-T-G-C) between positions 1 and 7, about 2.5 kbp apart from the 3′ half in the 2.8-kbp *Bam*HI fragment (Fig. 2), and the sequence between positions 1 and 37 was able to form a tRNA structure with the above sequence (positions 2564–2598) (Fig. 3). These two sequences show 72% and 56% homologies with those of *E. coli* and *Bacillus subtilis* tRNA^{Lys}(UUU)s, respectively (Fig. 4). Based on the sequence homology, the two sequences represent the gene for tRNA^{Lys}(UUU). The location of this tRNA^{Lys} gene is consistent with that based on tRNA/DNA hybridization studies reported by Bergmann *et al.* (15). One unusual feature in the sequence of the tRNA^{Lys} gene is the presence of a C-C mismatch base pair in the anticodon stem (Fig. 3).

The 2526-bp spacer between the 5′ and 3′ half of the tRNA gene sequences should be an intron. This is the longest intron so far found in tRNA genes and in chloroplast genes. The intron boundary sequences of the tRNA^{Lys} gene were found to be similar to those of tobacco chloroplast tRNA^{Gly}(UCC) gene (7), ribosomal protein L2 gene (ref. 16; our unpublished data), and ribosomal protein S12 gene (17) and also to those of *Euglena* genes for the large subunit of ribulose biphosphate carboxylase (18) and P32 (19) (Fig. 5). Their intron boundary sequences do not obey the G-U/A-G rule characteristic of eukaryotic mRNA precursors. From the sequence homologies with *E. coli* and *B. subtilis* tRNA^{Lys}(UUU)s (Fig. 4) and with the intron boundaries (Fig. 5), the 2526-bp intron in the tRNA^{Lys}(UUU) gene is most likely located at position 38–39 in the anticodon loop (Fig. 3). This intron site is the same as that found in maize and spinach spacer tRNA^{Ile}(GAU) genes (20).

Chloroplast tRNA genes containing introns have been divided into two groups (9, 21). Now we propose that they can be classified into three groups. (*i*) The tRNA^{Leu}(UAA) genes compose the first group (5, 9). Their introns can form unique secondary structures. Bonnard *et al.* (9) have reported that the split tRNA^{Leu}(UAA) gene from *Vicia faba* can be folded into a secondary structure that is very similar to the postulated structure of the intron from the autosplicable rRNA precursor of *Tetrahymena*. The tRNA gene transcripts in this group may be spliced by an autocatalytic reaction, as is the case for the *Tetrahymena* rRNA precursor (22). (*ii*) Spacer tRNA^{Ala}(UGC) and tRNA^{Ile}(GAU) genes (3, 4) and tRNA^{Val}(UAC) genes (6, 8, 10) constitute the second group. Their intron sequences are relatively similar to each other (6). (*iii*) The tRNA^{Gly}(UCC) gene (7) and the tRNA^{Lys}(UUU) gene (this paper) form the third group. Their intron boundary sequences are similar to each other and to those found in split chloroplast protein genes (Fig. 5). If the intron site of the tRNA^{Val}(UAC) gene is shifted one base (to position 38–39) from the position suggested (6), its possible intron boundary sequence aligns well with the above sequences (unpublished data). These tRNA precursors may be spliced by the same mechanism that worked for chloroplast mRNA precursors (17).

The Open Reading Frame. We found a long open reading frame between positions 750 and 2276 and two other short open reading frames (positions 209–349 and 370–534) in the intron of the tRNA^{Lys}(UUU) gene on the same strand (Fig. 2). No significant open reading frames were found in the opposite strand. The long open reading frame is 1527 bp long (509 codons; designated ORF_{509}) and starts with ATG and

```
     5'                                     intron                                       3'
(a)  GGGTCGTTAGCTCAGTTGGTAGAGCAGTTGACTTTTAA                  TCAATTGGTCGCAGGTTCGAATCCTGCACGACCCA
     **** * ** ****   ******* * * * *******                   * * * **  * *********** *  * *****
(b)  GGGTTGCTAACTCAAC GGTAGAGTACTCGGCTTTTAAGTGCGG······CTCCATCCGACTAGTTCCGGGTTCGAATCCCGGGCAACCCA
     * *    ** ****   ******* *   * *******                   *     **    ******* ***       * **
(c)  GAGCCATTAGCTCAGTTGGTAGAGCATCTGACTTTTAA                  TCAGAGGGTCGAAGGTTCGAGTCCTTCATGGCTCA
```

FIG. 4. Comparison of tobacco chloroplast tRNA^{Lys}(UUU) (line b) with *E. coli* (line a) and *B. subtilis* (line c) tRNA^{Lys}(UUU)s. Asterisks indicate homologous nucleotides.

```
                                  5'-exon                  intron                      3'-exon

Tobacco tRNAGly(UCC)              TGGTAAAA   GTGTGATTCGTTCTATT·······ATCGTCGTCGACTATAAC   CCCTAGCC
                                             *** *  * **    **            *  ** ***  *
Tobacco tRNALys(UUU)              GCTTTTAA   GTGCGGCTAGTCTCTTT·······GAACTTATCTACTCCAT    CCGACTAG
                                             ****** * *     **           *   ******* **
Tobacco ribosome protein L2       ACCTTTGA   GTGCGGTTTGAACTATT·······AGAAGAATCTACTTCAA    CCGATATG
                                             ***** * *  * *               *  *** **   *
Tobacco ribosome protein S12      GCGTTCTA   GTGCGTTGTAGATT··············ATGATCCACCCTAC   AATATGGG

Euglena conserved sequence                   GTGYG                           TTTARTTTTAT
```

FIG. 5. Comparison of the exon/intron boundary sequence of tobacco chloroplast tRNA^{Lys} gene with those of tobacco chloroplast genes for tRNA^{Gly} (7), ribosomal protein L2 (unpublished data), and ribosomal protein S12 (16). The conserved intron boundary sequence in *Euglena* chloroplast genes (16) is also shown.

Table 1. Codon usage in ORF_{509}

First position	Second position	U			C			A			G		Third position
		a	b		a	b		a	b		a	b	
U	Phe	28	46	Ser	22	43	Tyr	25	40	Cys	5	9	U
	Phe	15	38	Ser	7	24	Tyr	3	21	Cys	4	4	C
	Leu	24	62	Ser	14	18	Ter	0	6	Ter	1	1	A
	Leu	13	39	Ser	5	6	Ter	0	0	Trp	8	21	G
C	Leu	16	46	Pro	5	47	His	15	22	Arg	5	42	U
	Leu	1	7	Pro	0	13	His	3	16	Arg	5	11	C
	Leu	9	32	Pro	5	18	Gln	14	61	Arg	11	24	A
	Leu	4	17	Pro	2	13	Gln	4	22	Arg	3	5	G
A	Ile	20	87	Thr	9	51	Asn	24	64	Ser	9	23	U
	Ile	6	35	Thr	4	34	Asn	7	27	Ser	1	10	C
	Ile	10	28	Thr	3	29	Lys	20	63	Arg	8	34	A
	Met	8	63	Thr	2	15	Lys	10	13	Arg	3	4	G
G	Val	7	55	Ala	7	105	Asp	13	63	Gly	3	90	U
	Val	3	7	Ala	2	35	Asp	3	21	Gly	3	19	C
	Val	8	78	Ala	3	49	Glu	18	116	Gly	8	60	A
	Val	3	16	Ala	2	19	Glu	3	35	Gly	4	23	G

Codon usage in the open reading frame of 509 codons (ORF_{509}; columns a) compared with that in the sum of *rbcL, atpA, atpB, atpE, atpH, psbA*, and *rps19* genes (columns b). Ter, termination.

ends with TGA. The amino acid sequence of the putative polypeptide is shown in Fig. 2. The predicted polypeptide is 60,177 Da, basic, and hydrophilic (48% polar residues). The codon usage in ORF_{509} (Table 1) is similar to those in the large subunit of ribulosebisphosphate carboxylase, the proton-translocating ATPase α and β subunits, and the P32 genes from tobacco chloroplasts.

Open reading frames have been reported in some introns. For example, the second intron (*box3*) in the yeast mitochondrial apocytochrome *b* gene contains an open reading frame of 423 codons whose product has been proposed to be involved in splicing and maturation of its own precursor mRNA (23). The 2295-bp intron in the *Neurospora crassa* mitochondrial 24S rRNA gene contains an open reading frame of 426 codons, which could correspond to ribosomal protein S5 (24). ORF_{509} is similar in size to the above yeast and *Neurospora* open reading frames but shows no apparent sequence homology with them. No sequence similarity was also observed between ORF_{509} and *Euglena* chloroplast EF-Tu (25) or any of *E. coli* ribosomal proteins. At present no function of the tobacco ORF_{509} is known.

Expression of the $tRNA^{Lys}$ Gene. It is important to examine how the $tRNA^{Lys}$ gene containing such a long intron is expressed in the chloroplasts. Total tobacco chloroplast RNA extracted from young tobacco leaves was electrophoresed in a 1.2% agarose gel. The RNA was immobilized in nitrocellulose filter sheets and hybridized with the ^{32}P-labeled 877-bp *Sma* I–*Bam*HI fragment containing the 3′ half exon and the 3′ part of ORF_{509}, the 1356-bp *Bam*HI–*Eco*RI fragment containing the 5′ part of ORF_{509}, and the 750-bp *Eco*RI fragment containing the 5′ half exon. These three DNA probes hybridized to RNA bands of about 2.7, 2.2, 1.3, and 1.0 kb and several other minor RNA bands (Fig. 6). From its size, the 2.7-kb RNA should be an unspliced precursor molecule for the $tRNA^{Lys}$. The DNA probes containing the 3′ and 5′ exon halves hybridized to 4S-size RNA also, indicating that the precursor is processed to mature tRNA. The 4S RNA band was rather faint because of the low retention of small-size RNAs in nitrocellulose paper.

There were several discrete RNA bands between the 2.7-kb and 4S RNAs (e.g., 2.2, 1.3, and 1.0 kb), and these are likely to be the processing intermediates. These RNAs were moderately stable within the chloroplasts, suggesting a possible function for the RNAs. If ORF_{509} is a real gene for a polypeptide, one or more of the intermediates may serve as the mRNA. Several inverted repeats were found near both ends of the intron (Fig. 2), and these may be involved in the splicing and processing of the primary transcript.

The Pribnow box-like sequences T-A-T-A-A-T-T (positions −157 to −151) and T-A-T-T-T-T-T (positions −23 to −17) and the −35 region-like sequences A-T-G-A-C-A (posi-

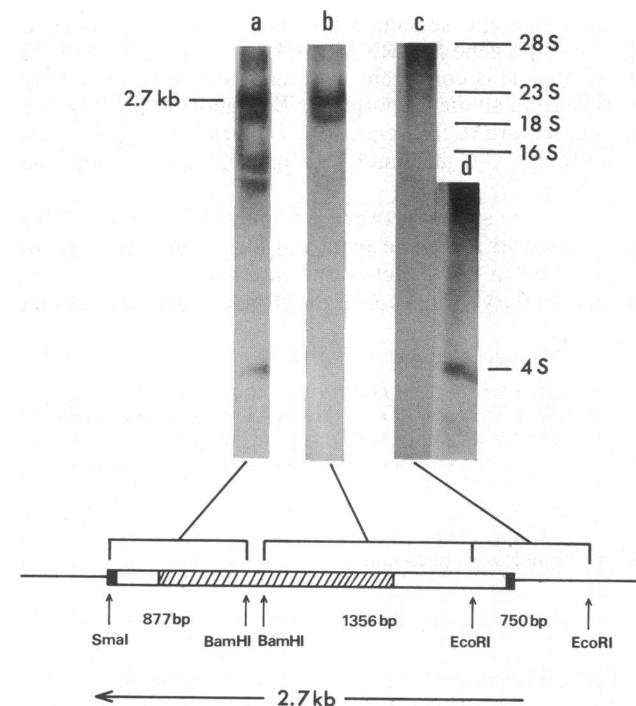

FIG. 6. Detection of precursor and mature RNA molecules for $tRNA^{Lys}$. (*Upper*) Autoradiographs of RNA blots of tobacco chloroplast RNA hybridized with the ^{32}P-labeled 877-bp *Sma* I/*Bam*HI fragment (lane a), 1356-bp *Bam*HI/*Eco*RI fragment (lane b), and 750-bp *Eco*RI fragment (lane c). Lane d is an overexposed autoradiograph of lane c. Size markers are *E. coli* 23S, 16S, and 4S RNAs and calf liver 28S and 18S rRNAs. (*Lower*) Location of the DNA probes used. Exons are shown as filled boxes and an intron is shown as an open box in which the hatched area indicates the open reading frame of 509 codons. An arrow indicating 2.7 kilobases (kb) indicates a primary transcript.

tions −182 to −177) and T-T-G-T-C-T (positions −41 to −36) were found in the region upstream from the 5′ exon of the $tRNA^{Lys}$ gene (Fig. 2). Three inverted repeats (positions 2604–2640, 2628–2645, and 2644–2674) were found in the downstream region from the 3′ exon and in the vicinity of the transcriptional starting site (position 2728) of the P32 gene (13). These structures may be transcriptional initiation and termination signals for the $tRNA^{Lys}$ gene. The distance between these two structures is about 2.7 kilbases, which agrees well with the size of a $tRNA^{Lys}$ transcript detected by the blot-hybridization analysis.

We thank Dr. M. Ohme for the preparation of tobacco chloroplast RNA. This work was supported in part by Grants-in-Aid from the Ministry of Education, Science and Culture and from the Toray Science Foundation.

1. Gauss, D. H. & Sprinzl, M. (1984) *Nucleic Acids Res.* **12,** r1–r131.
2. Sugiura, M., Shinozaki, K. & Ohme, M. (1985) in *Molecular Form and Function of the Plant Genome*, eds. Vloten-Doting, L., Groot, G. S. P. & Hall, T. C. (Plenum, New York), pp. 325–334.
3. Koch, W., Edwards, K. & Kössel, H. (1981) *Cell* **25,** 203–213.
4. Takaiwa, F. & Sugiura, M. (1982) *Nucleic Acids Res.* **10,** 2665–2676.
5. Steinmetz, A., Gubbins, E. J. & Bogorad, L. (1982) *Nucleic Acids Res.* **10,** 3027–3037.
6. Deno, H., Kato, A., Shinozaki, K. & Sugiura, M. (1982) *Nucleic Acids Res.* **10,** 7511–7520.
7. Deno, H. & Sugiura, M. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 405–408.
8. Krebbers, E., Steinmetz, A. & Bogorad, L. (1984) *Plant Mol. Biol.* **3,** 13–20.
9. Bonnard, G., Michel, F., Weil, J. H. & Steinmetz, A. (1984) *Mol. Gen. Genet.* **194,** 330–336.
10. Zurawski, G. & Clegg, M. T. (1984) *Nucleic Acids Res.* **12,** 2549–2559.
11. Ogden, R. C., Knapp, G., Peebles, C. L., Kang, H. S., Beckmann, J. S., Johnson, P. F., Fuhrman, S. A. & Abelson, J. N. (1980) in *Transfer RNA: Biological Aspects*, eds. Söll, D., Abelson, J. N. & Schimmel, P. R. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 173–190.
12. Sugiura, M. & Kusuda, J. (1979) *Mol. Gen. Genet.* **172,** 137–141.
13. Sugita, M. & Sugiura, M. (1984) *Mol. Gen. Genet.* **195,** 308–313.
14. Sugita, M., Kato, A., Shimada, H. & Sugiura, M. (1984) *Mol. Gen. Genet.* **194,** 200–205.
15. Bergmann, P., Seyer, P., Burkard, G. & Weil, J. H. (1984) *Plant Mol. Biol.* **3,** 29–36.
16. Zurawski, G., Bottomley, W. & Whitfeld, P. R. (1984) *Nucleic Acids Res.* **12,** 6547–6558.
17. Hallick, R. B., Gingrich, J. C., Johanningmeier, U. & Passavant, C. W. (1985) in *Molecular Form and Function of the Plant Genome*, eds. Vloten-Doting, L., Groot, G. S. P. & Hall, T. C. (Plenum, New York), in press.
18. Koller, B., Gingrich, J. C., Stiegler, G. L., Farley, M. A., Delius, H. & Hallick, R. B. (1984) *Cell* **36,** 545–553.
19. Karabin, G. D., Farley, M. & Hallick, R. B. (1984) *Nucleic Acids Res.* **12,** 5801–5812.
20. Guillemaut, P. & Weil, J. H. (1982) *Nucleic Acids Res.* **10,** 1653–1659.
21. Michel, F. & Dujon, B. (1983) *EMBO J* **2,** 33–38.
22. Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E. & Cech, T. R. (1982) *Cell* **31,** 147–157.
23. Lazowska, J., Jacq, C. & Slonimski, P. P. (1980), *Cell* **22,** 333–348.
24. Burke, J. M. & RajBhandary, U. L. (1982) *Cell* **31,** 509–520.
25. Montandon, P. E. & Stutz, E. (1983) *Nucleic Acids Res.* **11,** 5877–5892.