The EMBO Journal vol.5 no.9 pp.2043-2049, 1986

The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression

K.Shinozaki, M.Ohme, M.Tanaka, T.Wakasugi, N.Hayashida, T.Matsubayashi, N.Zaita, J.Chunwongse, J.Obokata, K.Yamaguchi-Shinozaki, C.Ohto, K.Torazawa, B.Y.Meng, M.Sugita[1], H.Deno[2], T.Kamogashira[3], K.Yamada[4], J.Kusuda[5], F.Takaiwa[6], A.Kato[6], N.Tohdoh[7], H.Shimada[8] and M.Sugiura

Centre for Gene Research and Department of Biology, Nagoya University, Chikusa, Nagoya 464, and [5]National Institute of Genetics, Mishima 411, Japan

Present addresses: [1]Department of Botany, Hokkaido University, Sapporo 060; [2]Mitsui Petrochemical Industries, Waki 740; [3]Otsuka Pharmaceutical Co., Tokushima 771-01; [4]Yamanouchi Pharmaceutical Co., Tokyo 174; [6]National Institute of Agrobiological Resources, Tsukuba 305; [7]National Cancer Center, Tokyo 104; [8]Mitsui-Toatsu Chemical Inc., Mobara 297, Japan

Communicated by J.H.Weil

The complete nucleotide sequence (155 844 bp) of tobacco (*Nicotiana tabacum* var. Bright Yellow 4) chloroplast DNA has been determined. It contains two copies of an identical 25 339 bp inverted repeat, which are separated by a 86 684 bp and a 18 482 bp single-copy region. The genes for 4 different rRNAs, 30 different tRNAs, 39 different proteins and 11 other predicted protein coding genes have been located. Among them, 15 genes contain introns. Blot hybridization revealed that all rRNA and tRNA genes and 27 protein genes so far analysed are transcribed in the chloroplast and that primary transcripts of the split genes hitherto examined are spliced. Five sequences coding for proteins homologous to components of the respiratory-chain NADH dehydrogenase from human mitochondria have been found. The 30 tRNAs predicted from their genes are sufficient to read all codons if the 'two out of three' and 'U:N wobble' mechanisms operate in the chloroplast. Two sequences which autonomously replicate in yeast have also been mapped. The sequence and expression analyses indicate both prokaryotic and eukaryotic features of the chloroplast genes.

Key words: DNA sequence/gene map/intron/tobacco chloroplast/transcription

Introduction

Chloroplasts are intracellular organelles present in plants, which contain the entire enzymic machinery for the process of photosynthesis. The discovery of non-Mendelian mutants of the chloroplast phenotype at the beginning of this century suggested the existence of a separate genetic system in chloroplasts. Since the demonstration of a unique DNA species in chloroplasts, over 20 years ago, intensive studies of the structure and expression of chloroplast genomes have been made (Dyer, 1984; Crouse *et al.*, 1984; Groot, 1985).

Chloroplast DNAs of higher plants are circular molecules with a size of 120-160 kbp. One of the outstanding features of chloroplast DNAs of most higher plants is the presence of two copies of a large inverted repeat (IR). These sequences (IR_A and IR_B) are separated by a large and a small single-copy region (LSC and SSC, respectively). Chloroplast DNAs are known to contain all the chloroplast rRNA genes (four genes in higher plants) and tRNA genes (~35 genes) and probably all the genes for proteins synthesized in the chloroplast (~100 genes) (Dyer, 1984; Gray *et al.*, 1984).

To understand the chloroplast genetic system more fully, we have determined the entire DNA sequence of the tobacco chloroplast genome. Tobacco plant has been chosen for our study because it has been a favoured material for studies of inheritance and evolution (Smith, 1974). There are many interspecific hybrids, chloroplast mutants and cell lines with altered chloroplast ribosomes. Moreover, tobacco cells provide a model system for studying somatic cell genetics, because of the recent technical advances in cell and protoplast cultures and protoplast fusion (Galun, 1981; Medgyesy *et al.*, 1985). We report here the overall arrangement of identified genes and possible protein-coding regions and summarize our present knowledge of transcription in the chloroplasts. More detailed reports of portions of the sequence have been published (see refs in Table I).

Results and Discussion

DNA sequence analysis

The clone bank of the entire tobacco chloroplast DNA as a set of overlapping restriction endonuclease fragments (Sugiura *et al.*, 1986) was used for sequencing. Overlapping DNA fragments are essential to cover the entire genome: otherwise very short restriction fragments are overlooked.

The physical map and gene map are shown in Figure 1. The maps are presented in linearized forms by cutting at the junction (J_{LA}) between IR_A and LSC (Sugiura *et al.*, 1986). J_{LA} has been designated zero and nucleotides are numbered proceeding towards the LSC. The DNA strand which codes for the large subunit of ribulose-1,5-bisphosphate carboxylase has been designated as A and the complementary strand as B (Deno *et al.*, 1983). The nomenclature for genes follows the proposals of Hallick and Bottomley (1983). The chloroplast DNA is divided into four regions (LSC, SSC, IR_A and IR_B) (Sugita *et al.*, 1984). LSC and SSC are 86 684 bp and 18 482 bp long, respectively. IR_A and IR_B have been sequenced separately and found to be completely identical (25 339 bp). The entire genome size is thus 155 844 bp long. The complete DNA sequence has been deposited with the EMBL database. Table I lists the genes and major open reading frames (ORF) with their positions, transcripts and other features.

rRNA and tRNA genes

The rRNA genes are arranged in the order of 16, 23, 4.5 and 5S rDNA in both IRs (Takaiwa and Sugiura, 1980, 1982a,b; Tohdoh and Sugiura, 1982). There are consequently two copies of each, or eight rRNA genes per genome. The coding regions for the mature 16 and 23S rRNAs have been determined by S1 mapping and those for the mature 4.5 and 5S rRNAs by sequencing the mature RNAs.

Thirty different tRNA genes have been identified in the DNA sequence (Kato *et al.*, 1981, 1985; Tohdoh *et al.*, 1981; Deno

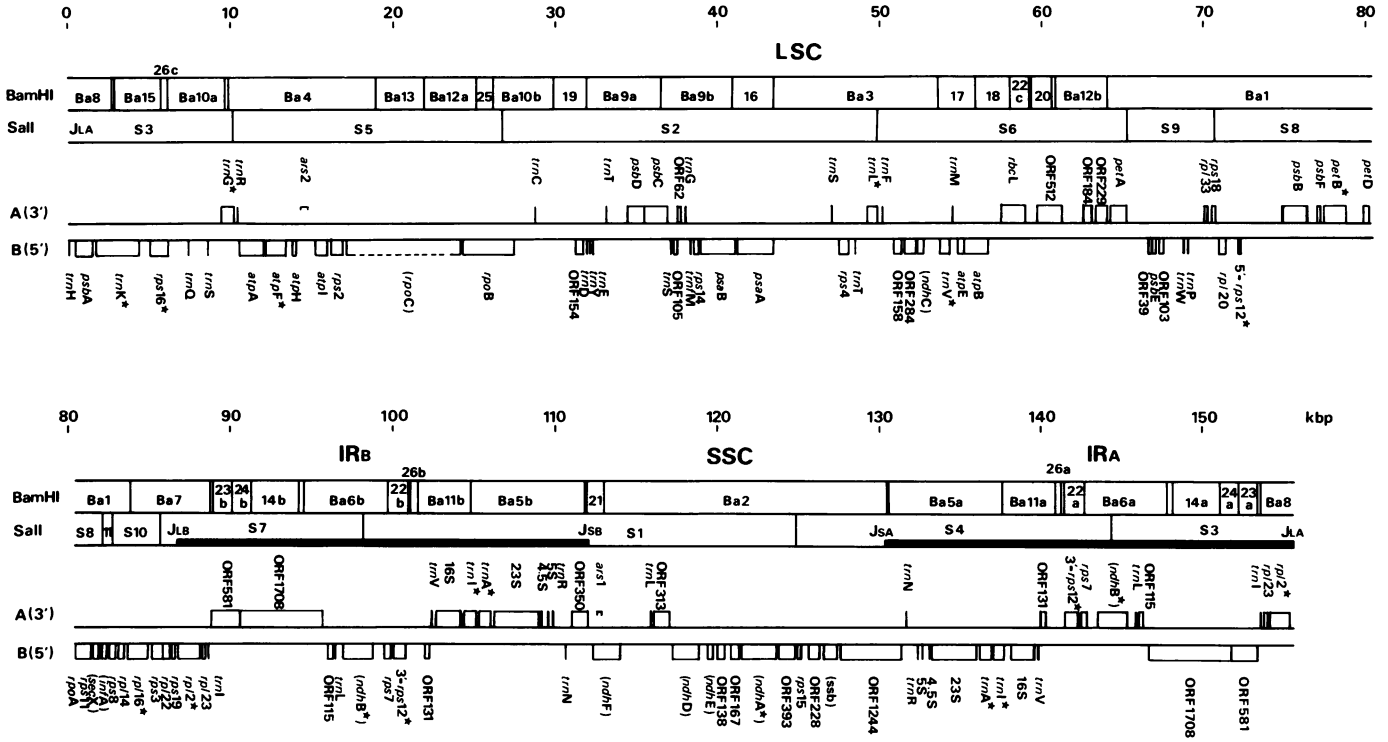

Fig. 1. Physical map and gene map of tobacco chloroplast DNA. The maps are presented by linearized forms by cutting at J_{LA}. The *Bam*HI (Ba) and *Sal*I (S) fragment maps (the upper part) are from Sigiura *et al.* (1986) supplemented with *Bam*HI fragments <0.4 kbp. IR_A and IR_B are shown by bold lines. J_{LB}, J_{SB}, J_{SA} and J_{LA} are junctions between IR and LSC or SSC. The lower part is the gene map. Genes are shown on their coding strands, strand A (A) and strand B (B). Putative genes are in parentheses. Asterisks indicate split genes. ORFs of over 100 codons, ORF62, ORF39 and ORF37 are also included.

et al., 1982; Deno and Sugiura, 1983, 1984; Ohme *et al.*, 1984, 1985; Sugita *et al.*, 1984, 1985; Yamada *et al.*, 1986; Wakasugi Ohome, Shinozati and Sugiura, submitted). Seven of them are located in the IR and therefore the total number of tRNA genes in the genome is 37. Hybridization analysis to total tobacco chloroplast tRNA has revealed that all tRNA genes are expressed. The map position of most of the tRNA genes is consistent with that based on a tRNA/DNA fragment hybridization study (Bergmann *et al.*, 1984). The presence of introns in chloroplast tRNA genes was first demonstrated in maize *trn*I and *trn*A by Koch *et al.* (1981). Six tRNA genes (*trn*I-GAU, *trn*A-UGC, *trn*V-UAC, *trn*L-UAA, *trn*G-UCC and *trn*K-UUU) contain introns which are 503–2526 bp long. Interestingly, *trn*G-UCC contains a 691 bp intron in the D stem (Deno and Sugiura, 1984). This seems to be a unique feature of chloroplast genomes (Quigley and Weil, 1985). The predicted amino acid sequence of ORF509A found in the *trn*K intron (Sugita *et al.*, 1985) has a local homology with that of the ORF present in the yeast mitochondrial *oxi*3 intron 2.

The minimum number of tRNA species required for translation of all codons is thought to be 32 for the universal genetic code. All possible codons are used in the sequences coding for proteins in tobacco chloroplasts (Sugita *et al.*, 1985). We have found genes for 30 tRNAs but could not detect genes for four other tRNAs which recognize codons CUU/C (Leu), CCU/C (Pro), GCU/C (Ala) and CGC/A/G (Arg). If the 'two out of three' mechanism can operate in the chloroplast, as has been shown in an *in vitro* protein synthesizing system from *Escherichia coli* (Samuelsson *et al.*, 1980), the single $tRNA^{Pro}$ (UGG), $tRNA^{Ala}$ (UGC) and $tRNA^{Arg}$ (ACG) can read all four Pro, Ala and Arg codons, respectively (GC pairs in the first and second codon–anticodon interaction). There is a gene for $tRNA^{Leu}$ (UAG) and if this tRNA has an unmodified U in the first position of the anticodon it can read all four Leu codons (CUN) by U:N wobble (Barrell *et al.*, 1980). The bean, spinach and soybean $tRNAs^{Leu}$ (UAG) have unmodified Us in their anticodons ($UA^{m7}G$) (Pillay *et al.*, 1984). These 28 tRNAs are therefore likely to be sufficient to read all codons in the tobacco chloroplast system using the above mechanisms. The possibility that the remaining tRNAs are imported from the cytoplasm or that their genes have unusual structures so as to prevent their detection cannot be excluded.

Genes for stromal polypeptides

Tobacco chloroplast ribosomes are of the 70S type and contain 58–62 ribosomal proteins (Capel and Bourque, 1982), one-third of which are thought to be encoded by chloroplast DNA. The electrophoretic patterns and molecular weight frequency distributions of ribosomal proteins of tobacco chloroplasts are quite similar to those of *E. coli* (Capel and Bourque, 1982). This led us to search for ribosomal protein genes through their homology with *E. coli* ribosomal protein genes (Sugita and Sugiura, 1983; Shinozaki *et al.*, 1986a; Torazawa *et al.*, 1986; Tanaka *et al.*, 1986). We have found 19 different sequences coding for polypeptides homologous to *E. coli* ribosomal proteins S2, S3, S4, S7, S8, S11, S12, S14, S15, S16, S18, S19, L2, L14, L16, L20, L22, L23 and L33. Northern blot hybridization revealed that the 12 different sequences so far examined are all expressed in the chloroplast. We therefore tentatively assigned these 19 sequences to be the genes for ribosomal proteins. Among them *rps*7, *rpl*2 and *rpl*23 are located in IR and therefore the total number of ribosomal protein genes is 22. The *rps*12, *rps*16, *rpl*2 and *rpl*16 contain introns which are 536–1020 bp long.

The most striking feature is that *rps*12 consists of three exons and that its 5′ exon (5′-*rps*12) is located 28 kbp downstream from

the other exons (3′-*rps*12) in IR_B on the same strand, or 86 kbp downstream from the 3′-*rps*12 in IR_A on the opposite strand (Torazawa *et al.*, 1986). A possible spliced mRNA for S12 has been detected by Fromm *et al.* (1986). These findings suggest that the tobacco *rps*12 gene consists of three transcription units and requires *trans* splicing. We propose to designate this gene structure as a 'divided' gene.

The *rpl*23, *rpl*2, *rps*19, *rpl*22, *rps*3, *rpl*16, *rpl*14 and *rps*8 genes are clustered in this order (*rpl*23 cluster) and this arrangement corresponds to that of the homologous genes in *E. coli* S10 - *spc* operons (Tanaka *et al.*, 1986). The 3′-*rps*12 and *rps*7 are arranged as in the *E. coli str* operon and are co-transcribed (Fromm *et al.*, 1986).

It has been suggested that the chloroplast RNA polymerase in higher plants is nuclearly encoded (Lerbs *et al.*, 1985). We have found that ORF337 corresponds to the spinach gene for the α subunit of RNA polymerase (*rpo*A) (Sijben-Müller *et al.*, 1986). ORF1070 has been assigned to be the gene for the β subunit (*rpo*B) (Ohme *et al.*, 1986). A series of four ORFs and three reading frames (RF, a region from a stop codon to the next stop codon) is located downstream from *rpo*B. Segments of the amino acid sequences deduced from four out of the seven ORFs and RFs show striking homology with portions of the β' subunit sequence of *E. coli* RNA polymerase. The sum of these homologous segments (~1360 codons) corresponds to the size of the *E. coli* β' subunit (1407 amino acid residues). These ORF and RFs may represent a split gene for the β' subunit (*rpo*C), although an extra splicing mechanism seems to be required. These findings raise the possibility that the chloroplast is an additional site of synthesis of its RNA polymerase subunits.

RF96 is similar to the spinach gene for the initiation factor IF-1 (*inf*A) although no initiation codon is found. RF96 could be a portion of the tobacco *inf*A or its pseudogene. ORF37 next to *rps*11 has homology with the *E. coli sec*X. The predicted amino acid sequence of ORF273 shows a local homology with that of *E. coli* single-stranded DNA binding protein and ORF273 may be the gene for the corresponding protein (*ssb*). ORF120 and ORF284 resemble the spinach and pea *bhp*A and *bhp*B (Zurawski *et al.*, personal communication). The total number of genes and putative genes for stromal proteins (including *rbc*L) is 28.

Genes for thylakoid polypeptides

Thylakoid membranes of higher plants have five functionally distinct complexes (Dyer *et al.*, 1984; Gray *et al.*, 1984; Herrmann *et al.*, 1985). These are the photosystem I (PSI), the photosystem II (PSII), the light-harvesting chlorophyll protein complex (its proteins are all nuclear coded), the cytochrome b/f complex and the H^+-ATPase complex.

We have found the genes for the P700 apoproteins A1 (*psa*A) and A2 (*psa*B) of PSI, the genes for the 32 kd protein (*psb*A) (Sugita and Sugiura, 1984), P680 apoprotein (*psb*B), 44 kd protein (*psb*C), D2 protein (*psb*D) and cytochrome b559 (*psb*E) of PSII, and the genes for cytochrome f (*pet*A), cytochrome b6 (*pet*B) and subunit 4 (*pet*D) of the cytochrome b/f complex. These genes have been identified by using the corresponding spinach gene probes (gifts from Dr R.G.Herrmann) and through their homology with the published sequences (Fish *et al.*, 1985; Alt *et al.*, 1984; Morris and Herrmann, 1984; Willey *et al.*, 1984; Heinemeyer *et al.*, 1984; Herrmann *et al.*, 1984). The tobacco *pet*B is likely to have a 759 bp intron. We have found that an ORF of 73 codons is located between *psb*B and *pet*B and its deduced N-terminal amino acid sequence matches that reported for the spinach 10 kd phosphoprotein of PSII (Farchaus and Dilley, 1986). We tentatively assigned this ORF to be the gene for the 10 kd phosphoprotein (*psb*F).

Of the nine subunits of the H^+-ATPase complex, six are coded for by the chloroplast DNA (subunits α, β, ϵ, I, III and a). Five genes (*atp*A, *atp*B, *atp*E, *atp*F and *atp*H) have been characterized previously (Deno *et al.*, 1983, 1984; Shinozaki *et al.*, 1983, 1986b). ORF247 shows high homology with the gene for the subunit a of pea chloroplasts (Cozens *et al.*, 1986) and is thought to be the subunit a gene (*atp*I). The gene *atp*F contains a 695 bp intron (Shinozaki *et al.*, 1986b). The total number of genes for thylakoid proteins is 17.

We have found that the predicted amino acid sequences of eight ORFs resemble those of components (ND1-5) of the respiratory-chain NADH dehydrogenase from human mitochondria (Chomyn *et al.*, 1985). ORF207 + ORF170 correspond to ND1, ORF180 + ORF260 to ND2, ORF120 to ND3, ORF509B to ND4, ORF101 to ND4L and ORF710 to ND5 and these are tentatively designated as *ndh*A, *ndh*B, *ndh*C, *ndh*D, *ndh*E and *ndh*F, respectively. The putative *ndh*A and *ndh*B genes contain single introns. Northern blot hybridization revealed that all six *ndh*s are expressed in the chloroplasts. It would therefore appear that these *ndh*s are functional at limited stages in plastid development, as NADH dehydrogenase is a mitochondrial enzyme and the presence of its activity has not been reported in higher plant chloroplasts. A further possibility is that this is an example of transposition in the direction opposite to what has been observed so far (namely the insertion of chloroplast genes into mitochondrial genomes, Stern and Lonsdale, 1982), so that these *ndh*s could be pseudogenes.

ORF39 next to *psb*E corresponds to the spinach ORF39 which has been suggested to be a gene for a component of PSII (Herrmann *et al.*, 1984). ORF62 before *trn*G-GCC has also been found in wheat, maize and spinach genomes and therefore ORF62 may be a gene for a membrane protein (Quigley and Weil, 1985).

Autonomously replicating sequences

The chloroplast DNA segments capable of replication in yeast (*ars*) have been cloned (in collaboration with Dr H.Uchimiya) and one of them, *ars*1 (350 bp segment), has been mapped (Ohtani *et al.*, 1984). *ars*1 is now known to be within ORF710 (*ndh*F). Here we present *ars*2 located between *atp*H and *atp*I. These two segments show stronger *ars* activity than others. The structure and location of tobacco *ars*1 and *ars*2 are similar to those of *Petunia ars*B and *ars*A, respectively (de Haas *et al.*, 1986).

Gene expression

The chloroplast genes are transcribed by the chloroplast RNA polymerase. Fourteen transcripts with definite sizes have so far been identified in the chloroplasts by Northern blot hybridization (see Table I). Some of the genes have been shown to be co-transcribed (e.g. *atp*F-*atp*A, *trn*E-*trn*Y-*trn*D, *atp*B-*atp*E, *rpl*23 cluster, 3′-*rps*12-*rps*7 and *rrn*) while others are transcribed monocistronically (e.g. *psb*A, *trn*K, *rps*16, *trn*G-UCC, *trn*V-UAC and *rbc*L).

Transcriptional initiation sites of the *psb*A, *trn*G-UCC, *trn*EYD, *atp*BE and *rbc*L genes have been identified by S1 mapping. Upstream of these sites there are sequences highly homologous to bacterial '−10' and '−35' regions. *Escherichia coli* RNA polymerase has been shown to recognize the tobacco *rbc*L and *atp*BE promoters and initiate transcription at their authentic initiation sites (Shinozaki and Sugiura, 1982a). Therefore most of the chloroplast promoters, if not all, resemble the prokaryotic promoter organization. Recently essential

Table I. Lists of tobacco chloroplast genes and their transcripts

	Gene	Gene product	Strand	Coding region start	Coding region end	Transcripts size (start - stop)	Protein amino acids (M. W.)	Introns (length) (donor - acceptor)	Reference
	[JLA]	[Junction IRA-LSC]		[155,844	1]				Sugita et al., 1984
	trnH	tRNA-His(GUG)	B	80	6	+		No	Sugita et al., 1984
	psbA	PSII 32kd protein	B	1,595	534	1,240±2b (1,680 - 441±2)	353 (38,950)	No	Sugita and Sugiura, 1984
	trnK	tRNA-Lys(UUU) 3'exon	B	1,844	1,810	2.7kb		1 (2,526bp)	Sugita et al., 1985
		5'exon	B	4,407	4,371	2.7kb		(4,370 - 1,845)	
	ORF509A		B	3,658	2,129	2.7kb			Sugita et al., 1985
	rps16	ribosomal protein S16 3'exon	B	5,311	5,094	1.3kb	85 (9,921)	1 (860bp)	Shinozaki et al., 1986
		5'exon	B	6,211	6,172	1.3kb	<14+71>	(6,171 - 5,312)	
	trnQ	tRNA-Gln(UUG)	B	7,487	7,416	+		No	Deno and Sugiura, 1983
	ORF98		A	7,724	8,020	ND			
	trnS	tRNA-Ser(GCU)	B	8,719	8,632	+		No	Deno and Sugiura, 1983
	trnG	tRNA-Gly(UCC) 5'exon	A	9,499	9,521	0.9kb		1 (691bp)	Deno and Sugiura, 1984
		3'exon	A	10,213	10,260	0.9kb (9,494±1 - ?)		(9,522 - 10,212)	
	trnR	tRNA-Arg(UCU)	A	10,430	10,501	+		No	Deno and Sugiura, 1984
	atpA	ATPase alpha subunit	B	12,148	10,625	cotranscription	507 (55,446)	No	Deno et al., 1983
	atpF	ATPase I subunit 3'exon	B	12,612	12,203	3.0kb	184 (19,085)	1 (695bp)	Shinozaki et al., 1986
		5'exon	B	13,452	13,308		<49+135>	(13,307 -12,613)	
	atpH	ATPase III subunit	B	14,099	13,854	0.8kb	81 (7,990)	No	Deno et al., 1984
	ars2			14,570	15,088				
	atpI	ATPase a subunit	B	16,001	15,258	*	247 (27,002)	No	
	rps2	ribosomal protein S2	B	16,938	16,228	ND	236 (26,943)	No	
	RF862	(E. coli rpoC)	B	19,753	17,165	ND		?	
	ORF134		B	20,277	19,873	ND			
	ORF80		B	20,423	20,181	ND			
	ORF90	(rpoC)	B	20,646	20,374	ND			
	RF236	(E. coli rpoC)	B	21,475	20,765	ND		?	
	RF548	(E. coli rpoC)	B	23,127	21,481	ND		?	
	ORF151	(E. coli rpoC)	B	24,283	23,828	ND		?	
	rpoB	RNA polymerase beta subunit	B	27,501	24,289	ND	1,070 (120,546)	No	Ohme et al., 1986
	trnC	tRNA-Cys(GCA)	A	28,783	28,854	+		No	Wakasugi et al., submitted
	ORF154		B	31,744	31,280	ND			
	trnD	tRNA-Asp(GUC)	B	31,999	31,926	cotranscription		No	Ohme et al., 1985
	trnY	tRNA-Tyr(GUA)	B	32,191	32,108	512b		No	Ohme et al., 1985
	trnE	tRNA-Glu(UUC)	B	32,323	32,251	(32,347 - 31,836)		No	Ohme et al., 1985
	trnT	tRNA-Thr(GGU)	A	33,172	33,243	+		No	Wakasugi et al., submitted
	psbD	PSII D2 protein	A	34,462	35,523	ND	353 (39,535)	No	
	psbC	PSII 44kd protein	A	35,471	36,892	ND	473 (51,909)	No	
	trnS	tRNA-Ser(UGA)	B	37,223	37,132	+		No	Wakasugi et al., submitted
	ORF105		B	37,558	37,241	ND			
	ORF62	(membrane protein ?)	A	37,586	37,774	ND			
	trnG	tRNA-Gly(GCC)	A	38,050	38,120	+		No	Ohme et al., 1984
	trnfM	tRNA-fMet(CAU)	B	38,421	38,348	+		No	Ohme et al., 1984
	rps14	ribosomal protein S14	B	38,873	38,571	ND	100 (11,744)	No	
	psaB	PSI P700 apoprotein A2	B	41,200	38,996	ND	734 (82,310)	No	
	psaA	PSI P700 apoprotein A1	B	43,478	41,226	ND	750 (82,990)	No	
	ORF77		A	44,264	44,497	ND			
	ORF82		B	45,394	45,146	ND			
	ORF74A		B	46,464	46,240	ND			
	trnS	tRNA-Ser(GGA)	A	47,111	47,197	+		No	Yamada et al., 1986
	rps4	ribosomal protein S4	B	48,133	47,528	ND	201 (23,420)	No	
	trnT	tRNA-Thr(UGU)	B	48,577	48,505	+		No	Yamada et al., 1986
	ORF70A		A	48,933	49,145	ND			
	trnL	tRNA-Leu(UAA) 5'exon	A	49,288	49,322	+		1 (503bp)	Yamada et al., 1986
		3'exon	A	49,826	49,875	+		(49,323 - 49,825)	
	trnF	tRNA-Phe(GAA)	A	50,232	50,304	+		No	Yamada et al., 1986
	ORF158		B	51,457	50,981	ND			
	ORF284	(bhpB)	B	52,417	51,563	ND	284 (32,325)	No	
	ORF120 (ndhC)	(mitochondria NADH dehydrogenase ND3), (bhpA)	B	52,659	52,297	*	120 (13,916)	No	
	trnV	tRNA-Val(UAC) 3'exon	B	53,781	53,747	0.75kb		1 (571bp)	Deno et al., 1982
		5'exon	B	54,390	54,353	0.75kb		(54,352 - 53,782)	
	trnM	tRNA-Met(CAU)	A	54,581	54,653	+		No	Deno et al., 1982
	atpE	ATPase epsilon subunit	B	55,276	54,875	cotranscription	133 (14,607)	No	Shinozaki et al., 1983
	atpB	ATPase beta subunit	B	56,769	55,273	2,350 - 2,390b (57,025 - 54,676±2) (57,025 - 54,637±1)	498 (53,554)	No	Shinozaki et al., 1983 Shinozaki and Sugiura, 1982a
	rbcL	RuBisCO large subunit	A	57,587	59,020	1757b (57,405 - 59,161)	477 (52,897)	No	Shinozaki and Sugiura, 1982b Shinozaki and Sugiura, 1982a
	ORF512		A	59,785	61,323	ND			
	ORF184		A	62,630	63,184	ND			
	ORF229		A	63,407	64,096	ND			
	petA	cytochrome f	A	64,327	65,289	5.0kb	320 (35,243)	No	
	ORF99A		A	66,168	66,467	ND			
	ORF39	PSII component	B	66,860	66,741	ND	39 (4,484)		
	psbE	PSII cytochrome b559	B	67,121	66,870	ND	83 (9,395)	No	
	ORF103		B	67,580	67,269	ND			
	trnW	tRNA-Trp(CCA)	B	68,880	68,807	+		No	Ohme et al., 1984
	trnP	tRNA-Pro(UGG)	B	69,118	69,045	+		No	Ohme et al., 1984
	rpl33	ribosomal protein L33	A	70,123	70,323	ND	66 (7,693)	No	
	rps18	ribosomal protein S18	A	70,510	70,815	ND	101 (12,052)	No	
	rpl20	ribosomal protein L20	B	71,401	71,015	1.1kb	128 (15,541)	No	
	5'-rps12	ribosomal protein S12 exon-1	B	72,326	72,213	ND	123 (13,764) <38+78+7>	trans splicing (72,212 100,852) (72,212 141,677)	Torazawa et al., 1986
	ORF73		B	72,686	72,465	ND			
	ORF74B		B	73,547	73,323	ND			
	psbB	PSII P680 apoprotein	A	74,950	76,476	ND	508 (55,855)	No	
	psbF	PSII 10kd phosphoprotein	A	77,098	77,319	ND	73 (7,759)	No	
	petB	cytochrome b6 5'exon	A	77,449	77,454	ND	215 (24,136)	1 (759bp)	
		3'exon	A	78,208	78,849	ND	<2+213>	(77,455 - 78,207)	
	petD	cyt.b/f complex subunit 4	A	79,845	80,264	ND	139 (15,225)	No	
	rpoA	RNA polymerase alpha subunit	B	81,465	80,452	ND	337 (38,612)	No	
	rps11	ribosomal protein S11	B	81,947	81,531	ND	138 (14,883)	No	
	ORF37	(E. coli secX)	B	82,162	82,049	ND	37 (4,460)	No	
	RF96	(E. coli infA)	B	82,465	82,175	ND			
	rps8	ribosomal protein S8	B	83,004	82,600	*	134 (15,790)	No	Tanaka et al., 1986
	rpl14	ribosomal protein L14	B	83,544	83,173	*	123 (13,738)	No	Tanaka et al., 1986
	rpl16	ribosomal protein L16 3'exon	B	84,064	83,669	*	134 (15,214)	1 (1,020bp)	Tanaka et al., 1986
		5'exon	B	85,093	85,085	*	<3+131>	(85,084 - 84,065)	
	rps3	ribosomal protein S3	B	85,896	85,240	*	218 (25,085)	No	Tanaka et al., 1986
	rpl22	ribosomal protein L22	B	86,348	85,881	*	155 (17,769)	No	Tanaka et al., 1986
	rps19	ribosomal protein S19	B	86,680	86,402	*	92 (10,411)	No	Sugita and Sugiura, 1983

L S C

Table I. Continued

	Gene	Gene product		Strand	Coding region start	Coding region end	Transcripts size (start - stop)	Protein amino acids (M. W.)	Introns (donor - acceptor) (length)	Reference
	[JLB]	[Junction LSC-IRB]			[86,684	86,685]				Sugita et al., 1984
	rpl2	ribosomal protein L2	3'exon	B	87,174	86,741	*	274 (30,010)	1 (666bp)	Tanaka et al., 1986
			5'exon	B	88,231	87,841	*	<131+143>	(87,840 - 87,175)	
	rpl23	ribosomal protein L23		B	88,531	88,250	*	93 (10,763)	No	Tanaka et al., 1986
	trnI	tRNA-Ile(CAU)		B	88,770	88,697	+		No	Tanaka et al., 1986
	ORF581			A	88,883	90,628	ND			
	ORF1708			A	90,598	95,724	ND			
	ORF87			A	95,815	96,078	ND			
	ORF92			A	96,116	96,394	ND			
	ORF115			B	96,404	96,057	ND			
	trnL	tRNA-Leu(CAA)		B	96,507	96,427	+		No	Wakasugi et al., submitted
	ORF79			A	96,553	96,792	ND			
	ORF260 (ndhB)	(mitochondria NADH dehydrogenase ND2)	3'exon	B	97,829	97,047	*	361 (39,655) <136+225>	1 (757bp) (98,481 - 97,725)	
	ORF180 (ndhB)	(mitochondria NADH dehydrogenase ND2)	5'exon	B	98,889	98,347	*			
IR_B	rps7	ribosomal protein S7		B	100,004	99,537	cotranscription 1.2kb	155 (17,386)	No	Fromm et al., 1986
IR_B	3'-rps12	ribosomal protein S12	exon-3	B	100,083	100,058		123 (13,764) <38+78+7>	1 (536bp) (100,619 - 100,084)	Fromm et al., 1986
IR_B			exon-2	B	100,851	100,620			(72,212 100,852)	
	ORF70B			A	102,099	102,311	ND			
	ORF131			B	102,343	101,948	*			
	trnV	tRNA-Val(GAC)		A	102,459	102,530	+ (102,436±3 - ?)		No	Tohdoh et al., 1981
	16SrDNA	16S rRNA		A	102,758	104,246			No	Tohdoh and Sugiura, 1982
	trnI	tRNA-Ile(GAU)	5'exon	A	104,547	104,583			1 (707bp)	Takaiwa and Sugiura, 1982a
			3'exon	A	105,291	105,325			(104,584-105,290)	
	trnA	tRNA-Ala(UGC)	5'exon rrnB	A	105,390	105,427	cotranscription 8.2kb		1 (709bp) (105,428-106,136)	Takaiwa and Sigiura, 1982a
			3'exon	A	106,137	106,171				
	23SrDNA	23S rRNA		A	106,325	109,134			No	Takaiwa and Sugiura, 1982b
	4.5SrDNA	4.5S rRNA		A	109,236	109,338			No	Takaiwa and Sugiura, 1980
	5SrDNA	5S rRNA		A	109,595	109,715			No	Takaiwa and Sugiura, 1980
	trnR	tRNA-Arg(ACG)		A	109,973	110,046	+		No	Kato et al., 1985
	trnN	tRNA-Asn(GUU)		B	110,699	110,628	+		No	Kato et al., 1981
	ORF75			B	110,820	110,593	ND			
	ORF350			A	111,025	112,077	ND			
	[JSB]	[Junction IRB-SSC]			[112,023	112,024]				Sugita et al., 1984
	ars1				112,768	113,117				Ohtani et al., 1984
	ORF710 (ndhF)	(mitochondria NADH dehydrogenase ND5)		B	114,198	112,066	*	710 (80,362)	No	
	trnL	tRNA-Leu(UAG)		A	116,067	116,146	+		No	Kato et al., 1985
	ORF313			A	116,250	117,191	ND			
	ORF509B (ndhD)	(mitochondria NADH dehydrogenase ND4)		B	118,958	117,429	*	509 (57,401)	No	
	ORF101 (ndhE)	(mitochondria NADH dehydrogenase ND4L)		B	119,860	119,555	*	101 (11,270)	No	
S S C	ORF99B			B	120,383	120,084	ND			
S S C	ORF138			B	120,612	120,196	ND			
S S C	ORF167			B	121,512	121,009	ND			
S S C	ORF170 (ndhA)	(mitochondria NADH dehydrogenase ND1)	3'exon	B	122,109	121,597	*	333 (37,049) <185+148>	1 (1,242bp) (123,286 - 122,045)	
S S C	ORF207 (ndhA)	(mitochondria NADH dehydrogenase ND1)	5'exon	B	123,840	123,217	*			
S S C	ORF393			B	125,023	123,842	ND			
S S C	rps15	ribosomal protein S15		B	125,398	125,135	ND	87 (10,445)	No	
S S C	ORF228			B	126,482	125,796	ND			
S S C	ORF273	(E. coli ssb)		B	127,561	126,740	ND	273 (33,023)	No	
	[JSA]	[Junction SSC-IRA]			[130,505	130,506]				Sugita et al., 1984
	ORF1244			B	131,501	127,767	ND			
	ORF75			A	131,709	131,936	ND			
	trnN	tRNA-Asn(GUU)		A	131,830	131,901	+		No	
	trnR	tRNA-Arg(ACG)		B	132,556	132,483	+		No	
	5SrDNA	5S rRNA		B	132,934	132,814			No	
	4.5SrDNA	4.5S rRNA		B	133,293	133,191			No	
	23SrDNA	23S rRNA		B	136,204	133,395			No	
	trnA	tRNA-Ala(UGC)	3'exon rrnA	B	136,392	136,358	cotranscription 8.2kb		1 (709bp) (137,101 - 136,393)	
			5'exon	B	137,139	137,102				
	trnI	tRNA-Ile(GAU)	3'exon	B	137,238	137,204			1 (707bp)	
			5'exon	B	137,982	137,946			(137,945 - 137,239)	
	16SrDNA	16S rRNA		B	139,771	138,283			No	
	trnV	tRNA-Val(GAC)		B	140,070	139,999	+ (140,093±3 - ?)		No	
	ORF131			A	140,186	140,581	*			
	ORF70B			B	140,430	140,218	ND			
IR_A	3'-rps12	ribosomal protein S12	exon-2	A	141,678	141,909	cotranscription 1.2kb	123 (13,764) <38+78+7>	1 (536bp) (72,212 141,677) (141,910 - 142,445)	
IR_A			exon-3	A	142,446	142,471				
IR_A	rps7	ribosomal protein S7		A	142,525	142,992		155 (17,386)	No	
IR_A	ORF180 (ndhB)	(mitochondria NADH dehydrogenase ND2)	5'exon	A	143,640	144,182	*	361 (39,655) <136+225>	1 (757bp) (144,048 - 144,804)	
IR_A	ORF260 (ndhB)	(mitochondria NADH dehydrogenase ND2)	3'exon	A	144,700	145,482	*			
	ORF79			B	145,976	145,737	ND			
	trnL	tRNA-Leu(CAA)		A	146,022	146,102	+		No	
	ORF115			A	146,125	146,472	ND			
	ORF92			B	146,413	146,135	ND			
	ORF87			B	146,714	146,451	ND			
	ORF1708			B	151,931	146,805	ND			
	ORF581			B	153,646	151,901	ND			
	trnI	tRNA-Ile(CAU)		A	153,759	153,832	+		No	
	rpl23	ribosomal protein L23		A	153,998	154,279	*	93 (10,763)	No	
	rpl2	ribosomal protein L2	5'exon	A	154,298	154,688	*	274 (30,010)	1 (666bp)	
			3'exon	A	155,355	155,788	*	<130+144>	(154,689 - 155,354)	
	[JLA]	[Junction IRA-LSC]			[155,844	1]				

ORFs of over 70 codons, ORF62, ORF39 and ORF37 are included. ORFs of their gene products or gene names in parentheses ( ) are putative genes. Numbers of amino acids in parentheses < > are those of exons of split genes. Plus (+) and asterisks (*) indicate transcripts detected by Southern and Northern blot hybridization, respectively, but their lengths were not determined. ND: not determined.

regions in the spinach *trn*M2, *rbc*L, *atp*BE and *psb*A promoters have been experimentally identified to be similar to the prokaryotic '−35' and '−10' regions (Gruissem and Zurawski, 1985). Many other genes also contain sequences similar to prokaryotic promoters in front of their coding regions and these sequences are most likely to be their promoters although they are not yet defined functionally (Crouse *et al.*, 1984; Kung and Lin, 1985). Some genes (e.g. *trn*K, *rps*16, *trn*V-UAC and *rrn*) seem to have multiple promoters.

Transcriptional termination sites of the *psb*A, *trn*EYD, *atp*BE and *rbc*L genes have also been identified by S1 mapping. Short inverted repeat sequences have been found just before the stop points. This indicates a further prokaryotic feature of the chloroplast genes. One interesting observation is that *atp*BE has two terminators both of which are located within *trn*M encoded on the opposite strand.

Fifteen identified and putative genes have been shown to contain introns. Among them, both primary and spliced transcripts have so far been detected for *trn*K, *rps*16, *atp*F and *trn*V. We have proposed that introns found in chloroplast genes can be classified into three groups (Shinozaki *et al.*, 1986a). Twelve out of the 15 introns belong to the group III introns which have conserved sequences at their boundaries. There seem to be three splicing mechanisms in the chloroplast. The *trn*L-UAA transcript has been suggested to be auto-spliced (Bonnard *et al.*, 1984). It would be interesting to elucidate molecular mechanisms for splicing operating in chloroplasts.

Conclusions

We have so far found genes for 34 different stable RNAs and 39 different proteins, putative genes for 11 different proteins and 38 different ORFs (over 70 codons, ORF62, ORF39 and ORF37, ORFs found on the complementary strand of functional genes are omitted), which represent a total of 122. Twenty-four out of these 122 sequences are in IR, so that the total number is 146 in the whole genome. This is an expected coding capacity, considering the size of tobacco chloroplast DNA.

The sequence and expression analyses have shown both prokaryotic and eukaryotic features of the chloroplast genes. The genes coding for rRNAs, tRNAs and some of proteins (e.g. ribosomal proteins) have substantial sequence homology with the prokaryotic counterparts. The basic regulatory sequences (promoters, terminators and ribosomal binding sites) are also similar to those in prokaryotic genomes. Some of the gene clusters resemble the corresponding clusters of *E. coli* and cyanobacteria (e.g. *rrn*, *rpl*23 and *atp* clusters).

Some of the chloroplast genes contain introns similar to those which have been found in eukaryotic genomes. However, introns found in the tRNA genes are very long (up to 2526 bp) and one intron is located in an unusual position, namely the D-stem region of *trn*G-UCC. The chloroplast splicing mechanisms seem to be more complex than eukaryotic splicing systems. The *rps*12 gene is divided into three parts which are far away from each other, and hence it is most likely to consist of three different transcription units and to require *trans* splicing (a divided gene).

The endosymbiotic theory, which proposes that chloroplasts derived from an ancestral photosynthetic prokaryote related to cyanobacteria, has been supported in part by comparisons between chloroplast and cyanobacterial *rrn* operons (Tomioka and Sugiura, 1983). This leads us to speculate that ancestral photosynthetic prokaryotes had introns in their genomes and that existing chloroplast genomes have retained these intron sequences.

Further studies are necessary to establish a complete gene map of the tobacco chloroplast genome.

Materials and methods

The clone bank of the entire tobacco (*Nicotiana tabacum* var. Bright Yellow 4) chloroplast DNA as a set of overlapping restriction endonuclease fragments was constructed (Sugiura *et al.*, 1986). IR_A and IR_B have separately been cloned using a cosmid, pHC79. Physical maps of the cloned fragments were constructed and their DNA sequences were determined initially by the chemical method (Maxam and Gilbert, 1977) and later by the dideoxynucleotide procedure (Sanger *et al.*, 1977) using the M13mp10/11 and M13mp18/19 phages and *E. coli* JM109 (Yanisch-Perron *et al.*, 1985). The whole sequence of each region was obtained on both strands and at least twice on one strand. To join up the sequences of adjacent clones, the sequence of a different clone overlapping the junction was determined. DNA sequence data were compiled and analysed in an NEC PC98XA computer using the GENETYX program (Software Development Co., Tokyo, Japan) and in a FACOM M160 compuer using the programs of Wilbur−Lipman (1983) and Staden (1980). Southern and Northern blot hybridizations were carried out as described (Sugiura and Kusuda, 1979; Ohme *et al.*, 1984, 1985).

Acknowledgements

We thank Drs T.Maruyama and T.Gojobori for their help in FACOM computer analysis, Dr H.Uchimiya for *ars* cloning, Dr R.G.Herrmann for spinach probes, and Drs J.C.Gray, M.Edelman, R.A.Dilley, G.Zurawski, J.Mason, R.Bottomley and P.Whitfeld for unpublished data. We also thank Mrs Mie Kusuda, Y.Sawano and C.Sugita for technical assistance, Drs A.Hirai, N.Tomioka, M.Kumano, T.Mikami and Y.Nakabori for helpful discussion, and Dr K.I.Miura for encourgement. This work was supported in part by a Grant-in-Aid for Special Distinguished Research from the Ministry of Education, Science and Culture, a grant from the Ministry of Agriculture, Forestry and Fisheries and a grant from the Toray Science Foundation.

A printout of the complete DNA sequence is available from M.Sugiura, Center for Gene Research, Nagoya University, Chikusa 464, Japan.

References

Alt,J., Morris,J., Westhoff,P. and Herrmann,R.G. (1984) *Curr. Genet.*, **8**, 597−606.

Barrell,B.G., Anderson,S., Bankier,A.T., de Bruijn,M.H.L., Chen,E., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F., Schreier,P.H., Smith,A.J.H., Staden,R. and Young,I.G. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 3164−3166.

Bergmann,P., Seyer,P., Burkard,G. and Weil,J.H. (1984) *Plant Mol. Biol.*, **3**, 29−36.

Bonnard,G., Michel,F., Weil,J.H. and Steinmetz,A. (1984) *Mol. Gen. Genet.*, **194**, 330−336.

Capel,M.S. and Bourque,D.P. (1982) *J. Biol. Chem.*, **257**, 7746−7755.

Chomyn,A., Mariottini,P., Cleeter,M.W.J., Ragan,C.I., Matsuno-Yagi,A., Hatefi,Y., Doolittle,R.F. and Attardi,G. (1985) *Nature*, **314**, 592−597.

Cozens,A.L., Walker,J.E., Phillips,A.L., Huttly,A.K. and Gray,J.C. (1986) *EMBO J.*, **5**, 217−222.

Crouse,E.J., Bohnert,H.J. and Schmitt,J.M. (1984) In Ellis,R.J. (ed), *Chloroplast Biogenesis*. Cambridge University Press, Cambridge, pp. 83−136.

de Haas,J.M., Boot,K.J.M., Haring,M.A., Kool,A.J. and Nijkamp,H.J.J. (1986) *Mol. Gen. Genet.*, **202**, 48−54.

Deno,H. and Sugiura,M. (1983) *Nucleic Acids Res.*, **11**, 8407−8414.

Deno,H. and Sugiura,M. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 405−408.

Deno,H., Kato,A., Shinozaki,K. and Sugiura,M. (1982) *Nucleic Acids Res.*, **10**, 7511−7520.

Deno,H., Shinozaki,K. and Sugiura,M. (1983) *Nucleic Acids Res.*, **11**, 2185−2191.

Deno,H., Shinozaki,K. and Sugiura,M. (1984) *Gene*, **32**, 195−201.

Dyer,T.A. (1984) In Baker,N.R. and Barber,J. (eds), *Chloroplast Biogenesis*. Elsevier, Amsterdam, pp. 23−69.

Farchaus,J. and Dilley,R.A. (1986) *Arch. Biochem. Biophys.*, **244**, 94−101.

Fish,L.E., Kück,U. and Bogorad,L. (1985) *J. Biol. Chem.*, **260**, 1413−1421.

Fromm,H., Edelman,M., Koller,B., Goloubinoff,P. and Galun,E. (1986) *Nucleic Acids Res.*, **14**, 883−898.

Galun,E. (1981) *Annu. Rev. Plant Physiol.*, **32**, 237−266.

Gray,J.C., Phillips,A.L. and Smith,A.G. (1984) In Ellis,R.J. (ed), *Chloroplast Biogenesis*. Cambridge University Press, Cambridge, pp. 137−163.

Groot,G.S.P. (1985) In van Vloten-Doting,L., Groot,G.S.P. and Hall,T.C. (eds), *Molecular Form and Function of the Plant Genome*. Plenum Press, New York, pp. 175−181.

Gruissem,W. and Zurawski,G. (1985) *EMBO J.*, **4**, 3375–3383.
Hallick,R.B. and Bottomley,W. (1983) *Plant Mol. Biol. Rep.*, **1**, 38–43.
Heinemeyer,W., Alt,J. and Herrmann,R.G. (1984) *Curr. Genet.*, **8**, 543–549.
Herrmann,R.G., Alt,J., Schiller,B., Widger,W.R. and Cramer,W.A. (1984) *FEBS Lett.*, **176**, 239–244.
Herrmann,R.G., Westhoff,P., Alt,J., Tittgen,J. and Nelson,N. (1985) In van Vloten-Doting,L., Groot,G.S.P. and Hall,T.C. (eds), *Molecular Form and Function of the Plant Genome*. Plenum Press, New York, pp. 233–256.
Kato,A., Shimada,H., Kusuda,M. and Sugiura,M. (1981) *Nucleic Acids Res.*, **9**, 5601–5607.
Kato,A., Takaiwa,F., Shinozaki,K. and Sugiura,M. (1985) *Curr. Genet.*, **9**, 405–409.
Koch,W., Edwards,K. and Kössel,H. (1981) *Cell*, **25**, 203–213.
Kung,S.D. and Lin,C.M. (1985) *Nucleic Acids Res.*, **13**, 7543–7549.
Lerbs,S., Bräutigam,E. and Parthier,B. (1985) *EMBO J.*, **4**, 1661–1666.
Maxam,A.M. and Gilbert,W. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 560–564.
Medgyesy,P., Fejes,E. and Maliga,P. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 6960–6964.
Morris,J. and Herrmann,R.G. (1984) *Nucleic Acids Res.*, **12**, 2837–2850.
Ohme,M., Kamogashira,T., Shinozaki,K. and Sugiura,M. (1984) *Nucleic Acids Res.*, **12**, 6741–6749.
Ohme,M., Kamogashira,T., Shinozaki,K. and Sugiura,M. (1985) *Nucleic Acids Res.*, **13**, 1045–1056.
Ohme,M., Tanaka,M., Chunwongse,J., Shinozaki,K. and Sugiura,M. (1986) *FEBS Lett.*, **200**, 87–90.
Ohtani,T., Uchimiya,H., Kato,A., Harada,H., Sugita,M. and Sugiura,M. (1984) *Mol. Gen. Genet.*, **195**, 1–4.
Pillay,D.T.N., Guillemaut,P. and Weil,J.H. (1984) *Nucleic Acids Res.*, **12**, 2997–3001.
Quigley,F. and Weil,J.H. (1985) *Curr. Genet.*, **9**, 495–503.
Samuelsson,T., Elias,P., Lustig,F., Axberg,T., Folsch,G., Akesson,B. and Lagerkvist,U. (1980) *J. Biol. Chem.*, **255**, 4583–4588.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Shinozaki,K. and Sugiura,M. (1982a) *Nucleic Acids Res.*, **10**, 4923–4934.
Shinozaki,K. and Sugiura,M. (1982b) *Gene*, **20**, 91–102.
Shinozaki,K., Deno,H., Kato,A. and Sugiura,M. (1983) *Gene*, **24**, 147–155.
Shinozaki,K., Deno,H., Sugita,M., Kuramitsu,S. and Sugiura,M. (1986a) *Mol. Gen. Genet.*, **202**, 1–5.
Shinozaki,K., Deno,H., Wakasugi,T. and Sugiura,M. (1986b) *Curr. Genet.*, **10**, 421–423.
Sijben-Müller,G., Hallick,R., Alt,J., Westhoff,P. and Herrmann,R.G. (1986) *Nucleic Acids Res.*, **14**, 1029–1044.
Smith,H.H. (1974) In King,R.C. (ed), *Handbook of Genetics 2*. Plenum Press, New York, pp. 281–314.
Staden,R. (1980) *Nucleic Acids Res.*, **8**, 817–825.
Stern,D.B. and Lonsdale,D.M. (1982) *Nature*, **299**, 698–702.
Sugita,M. and Sugiura,M. (1983) *Nucleic Acids Res.*, **11**, 1913–1918.
Sugita,M. and Sugiura,M. (1984) *Mol. Gen. Genet.*, **195**, 308–313.
Sugita,M., Kato,A., Shimada,H. and Sugiura,M. (1984) *Mol. Gen. Genet.*, **194**, 200–205.
Sugita,M., Shinozaki,K. and Sugiura,M. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 3557–3561.
Sugiura,M. and Kusuda,J. (1979) *Mol. Gen. Genet.*, **172**, 137–141.
Sugiura,M., Shinozaki,K., Zaita,N., Kusuda,M. and Kumano,M. (1986) *Plant Sci.*, **44**, 211–216.
Takaiwa,F. and Sugiura,M. (1980) *Mol. Gen. Genet.*, **180**, 1–4.
Takaiwa,F. and Sugiura,M. (1982a) *Nucleic Acids Res.*, **10**, 2665–2676.
Takaiwa,F. and Sugiura,M. (1982b) *Eur. J. Biochem.*, **124**, 13–19.
Tanaka,M., Wakasugi,T., Sugita,M., Shinozaki,K. and Sugiura,M. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, in press.
Tohdoh,N. and Sugiura,M. (1982) *Gene*, **17**, 213–218.
Tohdoh,N., Shinozaki,K. and Sugiura,M. (1981) *Nucleic Acids Res.*, **9**, 5399–5406.
Tomioka,N. and Sugiura,M. (1983) *Mol. Gen. Genet.*, **191**, 46–50.
Torazawa,K., Hayashida,N., Obokata,J., Shinozaki,K. and Sugiura,M. (1986) *Nucleic Acids Res.*, **14**, 3143.
Wilbur,W.J. and Lipman,D.J. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 726–730.
Willey,D.L., Auffret,A.D. and Gray,J.C. (1984) *Cell*, **36**, 555–562.
Yamada,K., Shinozaki,K. and Sugiura,M. (1986) *Plant Mol. Biol.*, **6**, 193–199.
Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) *Gene*, **33**, 103–119.

Received on 6 May 1986; revised 12 June 1986