GENE ORGANIZATION AND EXPRESSION OF CHLOROPLAST DNA FROM A LIVERWORT, *MARCHANTIA POLYMORPHA*

TAKAYUKI KOHCHI

1989

GENE ORGANIZATION AND EXPRESSION OF CHLOROPLAST DNA FROM A LIVERWORT, *MARCHANTIA POLYMORPHA*

TAKAYUKI KOHCHI

1989

ABBREVIATIONS

bp	base pair
BSA	bovine serum albumin
DNA	deoxyribonucleic acid
kb	kilobase
kDa	kilodalton
MOPS	3-(N-morpholino)propanesulfonic acid
M_r	molecular weight
nt	nucleotide
ORF	open reading frame
PIPES	piperazine-N,N'-bis(2-ethanesulfonic acid)
Pu	purine nucleotide
Py	pyrimidine nucleotide
RNA	ribonucleic acid
mRNA	messenger RNA
rRNA	ribosomal RNA
tRNA	transfer RNA
SDS	sodium dodecyl sulfate
A. nidulans	*Anacystis nidulans*
As. nidulans	*Aspergillus nidulans*
Az. vinelandii	*Azotobacter vinelandii*
Az. chroococcum	*Azotobacter chroococcum*
C. reinhardtii	*Chlamydomonas reinhardtii*
Cy. paradoxa	*Cyanophora paradoxa*
E. coli	*Escherichia coli*
Eu. gracilis	*Euglena gracilis*
L. tarentolae	*Leishmania tarentolae*
M. polymorpha	*Marchantia polymorpha*
N. tabacum	*Nicotiana tabacum*
P. sativum	*Pisum sativum*
R. capsulata	*Rhodopseudomonas capsulata*
S. typhimurium	*Salmonella typhimurium*

CONTENTS

INTRODUCTION

Chloroplasts are photosynthetic organelles which have their own genetic system. The non-Mendelian inheritance of chloroplast characters suggested that there is cytoplasmic transmission of the genetic information. Sager & Ishida (1963) first demonstrated that *C. reinhardtii* chloroplasts contain DNA molecules distinct from those of the nuclear genome. Chloroplast genomes are double-stranded, circular DNA molecules, ranging 120-160 kb in size (depending on the species). Chloroplast DNA has a set of large inverted repeat sequences (IR) separated by a large single-copy (LSC) region and a small single-copy (SSC) region. There are now several reports dealing with the physical structure of chloroplast DNA. Expression of chloroplast genome is genetically controlled, not only by their own genes but also by nuclear genes. Nuclear-encode proteins are synthesized in cytoplasm as precursor molecules with transit sequences and then transported into chloroplast as processed molecules. The chloroplast-encoded proteins are expressed within the chloroplasts *via* their own machinery, although the machinery is also composed partly of nuclear-encode proteins (Ellis, 1981; Whitfeld & Bottomley, 1983; Crouse *et al.*, 1985; Palmer, 1985).

The author has focused on the chloroplast gene organization of a liverwort, *M. polymorpha*, a bryophyte, for several reasons. (i) Established cell cultures are available that are highly active in photosynthesis (Ohta *et al.*, 1977). (ii) The cells are haploid (n = 9; Ono, 1976), and can be reversibly converted into protoplasts (Ono *et al.*, 1979), which makes them suitable for experiments in genetics and cell biology. (iii) Chloroplast DNA and RNA molecules can be highly purified from cultured cells (Ohyama *et al.*, 1982; Ohyama *et al.*, 1983; Yamano *et al.*, 1984, 1985). (iv) Chloroplast DNA is about 120 kb long (Ohyama *et al.*, 1983), one of the smallest chloroplast genomes so far examined (Palmer, 1985). A liverwort, *M. polymorpha*, is a model organism to study chloroplast genetic system.

Considerable progress has been made in understanding molecular aspect of chloroplast genetic system. Studies on DNA synthesis in liverwort chloroplast were performed by Dr. Akira Tanaka (Tanaka, A., PhD thesis, 1984). Construction of the physical map and studies on chloroplast small RNAs (4.5S rRNA, 5S rRNA, and tRNA(Ser) in liverwort chloroplasts were performed by Dr. Yosiaki Yamano (Yamano, Y., PhD thesis, 1985). Studies on the transcriptional promoters and the structural analysis of the chloroplast genome were by Dr. Hideya Fukuzawa (Fukuzawa, H., PhD thesis, 1986).

The aim of this study

To improve our understanding of the chloroplast genetic system, the best strategies are to determine the entire nucleotide sequence of the chloroplast DNA and to deduce the gene organization. Although the nucleotide sequences of many chloroplast genes have been determined, the entire gene organization of chloroplast genome had not been elucidated for any species of plants. In this study, entire gene organization of chloroplast genome in liverwort is clarified (Chapter I). Based on the deduced gene organization, some unique feature and problem have arisen about gene expression, especially RNA processing. A presence of divergently overlapping transcript in chloroplasts and its control are shown (Chapter II). The processing order of polycistronic transcript containing two introns is analyzed for ORF203-*rps12'* region as an example (Chapter III). It had been shown that the ribosomal protein S12 gene in chloroplast is *trans*-split (Fukuzawa *et al.*, 1986). The author analyzed transcripts of the S12 gene and deduced possible mechanism of *trans*-splicing depending on the secondary structure of introns (Chapter IV). These studies described above will give us the basic information on the genetic system and molecular evolution in chloroplasts.

CHAPTER I Gene organization of the liverwort chloroplast genome

INTRODUCTION

Chloroplasts are the photosynthetic organelles in green plants and contain their own unique DNA. Genetic control of chloroplasts is performed not only by their own genes, but also by the nuclear-encoded genes. The chloroplast genes are expressed by transcription-translation machinery in chloroplasts.

In chapter I-1, the author focuses genetic system of chloroplasts for their gene expression. The nucleotide sequences of the chloroplast DNA in a liverwort, M. polymorpha, was determined to clarify the genetic contribution of chloroplast genome in plant cells.

In chapter I-2, the author describes genes and their characteristics deduced from the complete nucleotide sequence of the inverted repeat (IR) and small single copy (SSC) regions in the liverwort M. polymorpha chloroplast genome. The large inverted repeat sequences are identical by mutation analysis of C. reinhardii chloroplast DNA (Myers et al., 1982). In the SSC region, the leucine tRNA gene only has been mapped in land plants (Bergmann et al., 1984; Kato et al., 1985).

I-1 Structure and organization of Marchantia polymorpha chloroplast genome — Cloning and gene identification

MATERIALS AND METHODS

Cells and culture medium for M. polymorpha

Callus tissue of the liverwort M. polymorpha was originally derived from female gemma cultures (Ono, 1973). Cells were grown in liquid medium of 1M51C (Gamborg et al., 1968) on a gyratory shaker (150 rpm) under continuous illumination (3,000 lux). Cells from 7- to 10-day-old cultures were used for preparation of chloroplasts and its DNA (Ohyama et al., 1982).

Phages, plasmids, bacterial strains, and biochemicals

The fragments of liverwort chloroplast DNA were cloned into E. coli vectors of plasmids pBR322, pKC7, pUC13, pUC18, and pUC19. Phages M13mp8, M13mp9, M13mp10, MP13mp11, M13mp18, and M13mp19 were used for the cloning and sequencing of fragments with their

Table 1

Clones and chloroplast DNA fragments used for DNA sequencing

Clones	Chloroplast DNA fragments	Vectors	Position	
			from	to
pMP389	Bg 4	pKC7	117691	6548
pMP589	Bg16	pKC7	6549	8772
pMp594	Bg18	pKC7	8773	10405
pMP561	Bg17	pKC7	34137	35861
pMP591	Bg 9	pKC7	35918	40988
pMP314	Bg14	pKC7	44526	47294
pMP593	Bg10	pKC7	51612	55380
pMP452	Bg 8	pKC7	55381	61569
pMP713	Bg13	pUC18*	61570	64474
pMP310	Bg 5	pKC7	65514	73011
pMP376	Bg 3	pKC7	73012	84425
pMP321	Bg21b	pKC7	85138	85860
pMP318	Bg 7	pKC7	85861	92236
pMP323	Bg 6	pKC7	109353	116255
pMP238	Ba11	pBR322	119934	1135
pMP222	Ba 9	pBR322	29633	32566
pMP209	Ba12	pBR322	32567	34489
pMP217	Ba13	pBR322	34490	35740
pMP220	Ba 8	pBR322	35741	39778
pMP227	Ba 7	pBR322	39779	44648
pMP228	Ba 5	pBR322	47156	57496
pMP206	Ba 6	pBR322	76398	82182
pMP055	Ba 4	pBR322	82183	96252
pMP768	Ps 8	pBR322	12390	18106
pMP773	Ps 7	pBR322	18107	23831
pMP795	Ps10	pBR322	23832	28628
pMP708	Ps 9	pBR322	57087	62372
pMP727	Ps 6	pBR322	62373	69319
pMP710	Ps11	pBR322	69809	73525
pMP703	Ps17	pBR322	97653	98117
pMP102	Bc22	pUC18*	98780	99915
pMP101	Bc27	pUC18*	99916	100470
pMP103	Bc15	pUC18*	100471	104291
pMP151	Cl15	pUC18*	103976	105049
pMP152	Cl11	pUC18*	105050	107462
pMPX520	Xh 7	pUCX8	129	6196
pMPX501	Xh 8	pUCX8	6197	12217
pMP603	Hd16b(IR_B)	pBR322	83734	86627
pMP699	Hd 9	pUC18*	106418	110391
pMP802	Hd16a(IR_A)	pBR322	115489	118382
pMP781	Bg 1-Ps 1	pUC13**	10824	12389
pMPX601	Xh 2-Kp 3	pUCX8	12217	17178
pMP801	Ps 4-Ba 1	pUC13**	28629	29632
pMP721	Bg 2-Ps 2	pBR322	94522	97652
	Ps 1-Bg 2		108989	109352
pMP171	Ps 5-Bc26	pUC18	98118	98779
pMP172	Cl 3-Ps 5	pUC18	107463	108988
pMP173	Bc18-Xb 4	pUC18	104292	105104
pMPm179	Al179	M13mp18	97648	97866
pMPm120	Al120	M13mp18	98477	98815
pMPm 46	Al 46	M13mp18	99693	100278

The clones pMP pMPm indicate recombinant plasmid and M13 phage, respectively. The restricted fragments obtained with BglII, BamHI, PstI, BclI, ClaI, XhoI, XbaI, and AluI restriction endonucleases are shown as Bg, Ba, Ps, Cl, Xh, Xb and Al, respectively. Transformation has been done with E. coli strain HB101. Single and double asterisks indicate host strain of E. coli JM109 and JM105, respectively. The 1st position number was given to the 1st nucleotide of the LSC region in the junction (J_{LA}, see Fig. 2) and the numbering proceeds counterclockwise to the last nucleotide of IR_A region.

bacterial hosts, *E. coli* strains JM105 and JM109.

Nucleotide sequencing

Nucleotide sequences were determined for chloroplast DNA fragments subcloned from the recombinant DNA plasmids described above. Occasionally, chloroplast DNA fragments were directly cloned from chloroplast DNA into M13mp18 and M13mp19 vectors (Table 1). Nucleotide sequencing was done as follows. (i) Fragments of chloroplast DNA cloned in pBR322 or pKC7 were prepared by sonication (Messing *et al.*, 1981) and treated with nuclease P1 followed by DNA polymerase (Klenow fragment) to form blunt ends. Fragments larger than 600 bp were collected by electrophoresis through disc agarose gels, ligated into M13mp18 or M13mp19 vectors, cut at the *Hinc*II site, and introduced into cells of *E. coli* strain JM109 treated with calcium. M13 phages containing chloroplast fragments were identified by plaque hybridization with chloroplast DNA probes. Single-stranded DNA of M13 phages, precipitated by the polyethyleneglycol method, was used for sequencing of nucleotides by the dideoxy-chain termination technique (Sanger *et al.*, 1977a). At times, chloroplast DNA fragments recovered from recombinant DNA were self-ligated to form circular molecules and sonicated as described above. Enzymatically selected chloroplast fragments were used instead of shotgun-clones to close gaps. (ii) Selected chloroplast DNA fragments were directly cloned into appropriate cloning sites of the pUC18 or pUC19 vector. Deletion mutants were obtained as described by Yanisch-Perron *et al.* (1985). DNA from these deletion mutants was sequenced by the procedures described above. With small DNA fragments, double-stranded DNA of M13mp18 or M13mp19 was used instead of a pUC vector to generate the deletion mutants.

Nomenclature of chloroplast genes

The nomenclature for the chloroplast genes identified followed that used in the review article of Crouse *et al.* (1985). Unidentified open reading frames (ORFs) are shown as ORF(n) depending on the number (n) of amino acids in the ORFs. We tentatively designated chloroplast ORFs with homology to those found in human mitochondria as *ndh*, ORFs with homology to Fe-proteins as *frx*, inner membrane permease protein found in *E. coli* as *mbp*, initiation factor 1 as *infA*, and a functionally unidentified protein expressed in *E. coli* as *secX* (X gene, Cerretti *et al.*, 1983), protein of which has been identified as a ribosomal protein of the 50S subunit (Wada & Sako, 1987). The symbol *lhcA* was given to the ORF with much homology to antenna proteins of the light harvesting complex (LHC) found in photosynthetic bacteria.

RESULTS

Physical maps of chloroplast DNA

Chloroplast DNA was prepared from purified chloroplasts from cell suspensions of *M. polymorpha* cultures (Ohyama *et al.*, 1982). Physical

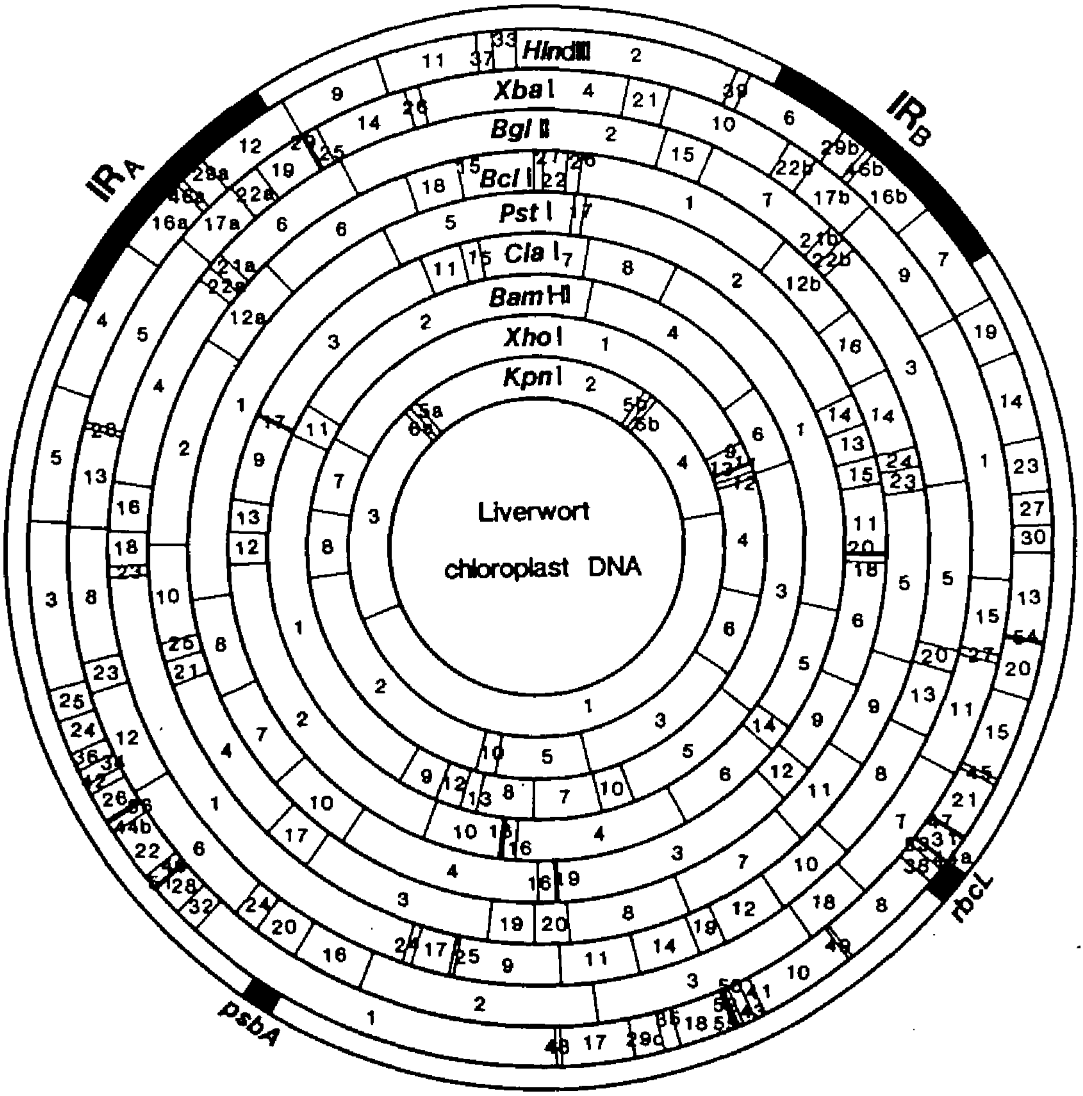


Fig. 1. Physical map of liverwort chloroplast DNA.

IR_A and IR_B indicate a set of large inverted repeats. The genes rbcL and psbA are given as landmarks on the liverwort chloroplast genome. Restriction fragments are numbered in decreasing size; symbols a, b, and c are given to fragments of the same size.

maps of the DNA previously constructed with the restriction endonucleases BamHI, SmaI, KpnI, and XhoI showed it to be a circular molecule of about 121.0 kb (Ohyama et al., 1983; Umesono et al., 1984). Physical maps constructed for DNA cloning and sequencing coincided well with those obtained from the complete nucleotide sequence (Fig. 1).

A set of large inverted repeats predicts the presence of two isomeric molecules in the chloroplast DNA formed by recombination across the repeated sequences (Palmer, 1985). A physical map of one of the isomeric molecules was shown although two isomeric molecules were dtected by digestion with the restriction endonuclease PstI, which does not cleave within the large inverted repeat (IR) regions. The nucleotide sequence was determined using independent clones from each IR_A and IR_B region. Nucleotide sequences were identical in the IR_A and IR_B regions (Fig. 1).

Gene detection strategy for protein coding sequence

Significant ORFs were identified with use of the universal codon table with the initiation codons ATG or GTG and with consideration that the ORFs are framed as long as possible (Crick, 1966a). ORFs can be considered to have introns if there are a 5'-consensus sequence (GTGPyG) and 3'-consensus secondary structure of introns in their coding regions. Proteins homologous to ORFs satisfying these conditions were found in a search by a FACOM M-160 computer (National Institute of Genetics, Mishima, Japan) with a program by Wilbur & Lipman (1983) from an NBRF (Release 8 - 12) database. Extensive computer analysis of the complete nucleotide sequence was done on a FACOM M-380 computer (Institute of Chemical Research, Kyoto University) and FACOM M-382 (Data Processing Center, Kyoto University) using the program IDEAS and on a personal computer PC-9801 (NEC Corp.) with the Hitachi DNASIS program (Hitachi SK. Ltd.).

The entire gene organization and the information obtained from it are shown in Fig. 2 and Tables 2, 3, and 4. The numbering of the

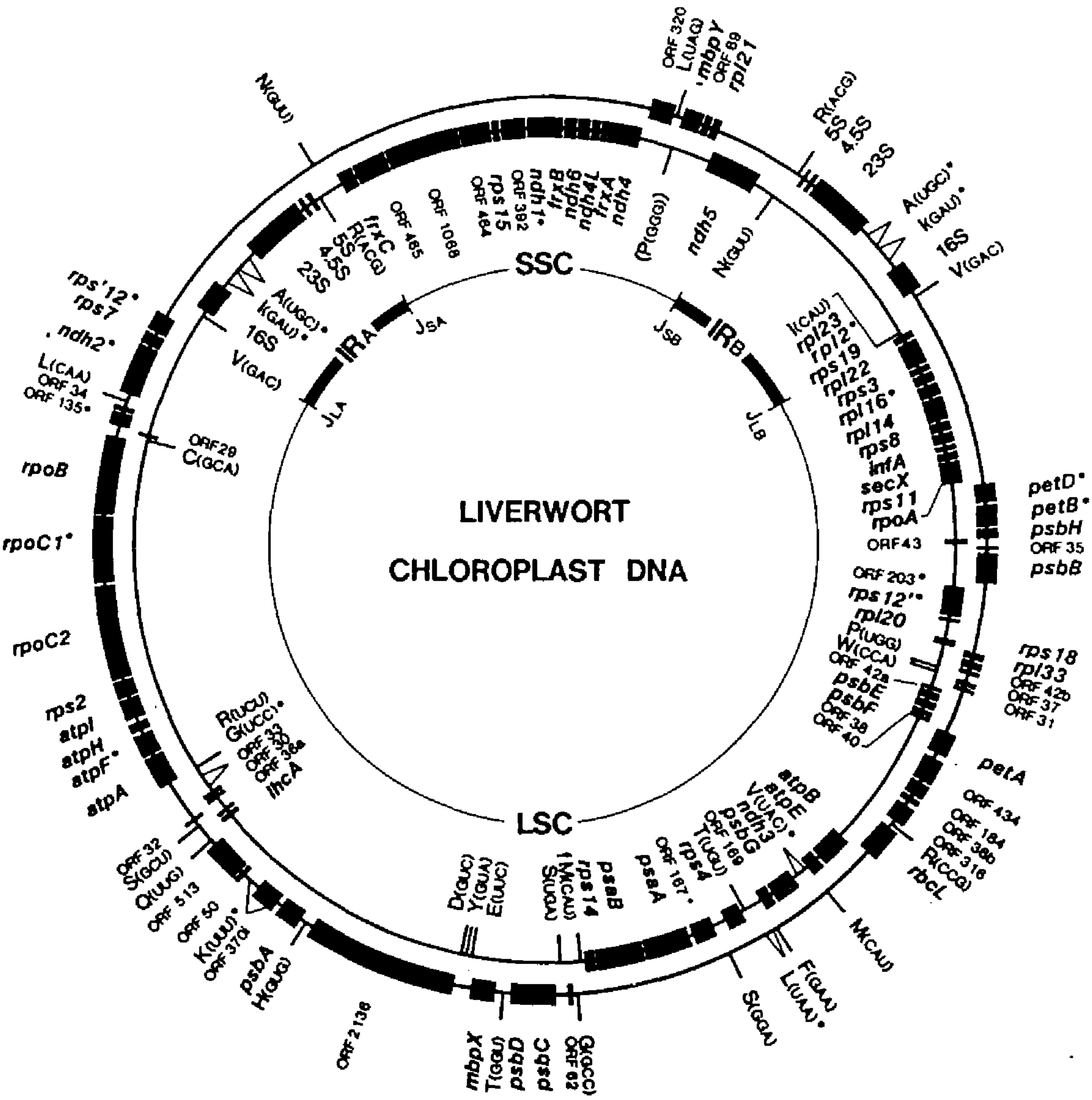


Fig. 2. Gene organization of the liverwort chloroplast genome.
Genes shown on the outside map are transcribed counterclockwise; those inside are transcribed clockwise. Thick lines inside the map indicate a set of large inverted repeats (IR_A and IR_B). SSC and LSC indicate the small single copy region and large single copy region, respectively. Asterisks indicate genes having introns in their coding sequences. The tRNA genes are shown by the one-letter code for amino acids with their anticodons in parentheses. Abbreviations for genes follow the nomenclature of Crouse et al. (1985). We designate chloroplast ORFs that have homology to those found in human mitochondria as ndh, ORFs homologous to Fe-proteins as frx, and ORFs homologous to inner membrane permease protein found in E. coli as mbp. ORFs that show homology to genes for initiation factor 1 and a functionally unknown protein expressed in E. coli are designated infA and secX, respectively. The genetic symbol lhcA is given to the ORF that has homology to antenna proteins in bacterial light harvesting complexes (LHC).

Table 2

Ribosomal and transfer RNA genes on the liverwort chloroplast genome

Genes	DNA	Position		Length	Promoter	Remarks	Genes	DNA	Position		Length	Promotor	Remarks
	strand	From	To	(bp)	(-35, -10)			strand	From	To	(bp)	(-35, -10)	
(a) *Ribosomal RNA genes*							*trnL(UAA)	+	50522	50921	400	+	intron 315 bp
16S rRNA	+	82109	83604	1,496	+		(exon 1)		50522	50556	35		
23S rRNA	+	85912	88722	2,811	-		(exon 2)		50872	50921	50		
4.5S rRNA	+	88830	88932	103	-		*trnF*(GAA)	+	50998	51070	73	-	
5S rRNA	+	89159	89277	119	-		*trnV(UAC)	-	53652	53051	602	+	intron 530 bp
5S rRNA	-	112961	112843	119	-		(exon 1)		53652	53616	37		
4.5S rRNA	-	113290	113188	103	-		(exon 2)		53085	53051	35		
23S rRNA	-	116208	113398	2,811	-		*trnM*(CAU)	+	53801	53874	74	+	
16S rRNA	-	120011	118516	1,496	+		*trnR*(CCG)	+	57877	57950	74	-	
							trnW(CCA)	-	64626	64553	74	-	
(b) *Transfer RNA genes*							*trnP*(UGG)	-	64788	64715	74	+	
trnL(CAA)	+	3679	3758	80	-		*trnI*(CAU)	-	81057	80984	74	+	
trnC(GCA)	-	5720	5650	71	+		*trnV*(GAC)	+	81814	81885	72	+	
trnR(UCU)	-	21321	21250	72	-		*trnI(GAU)	+	83878	84835	958	-	intron 886 bp
*trnG(UCC)	-	22047	21385	663	-	intron 593 bp	(exon 1)		83878	83914	37		
(exon 1)		22047	22025	23			(exon 2)		84801	84835	35		
(exon 2)		21431	21385	47			*trnA(UGC)	+	84912	85752	841	-	intron 768 bp
trnS(GCU)	+	22892	22979	88	+		(exon 1)		84912	84949	38		
trnQ(UUG)	+	23804	23875	72	+		(exon 2)		85718	85752	35		
*trnK(UUU)	+	26040	28222	2,183	+	intron 2111 bp	*trnR*(ACG)	+	89482	89555	74	-	
(exon 1)		26040	26076	37			*trnN*(GUU)	-	90332	90261	72	+	
(exon 2)		28188	28222	35			*trnP*(GGG)	-	95213	95145	69	--	pseudogene
trnH(GUG)	+	29595	29669	75	-		*trnL*(UAG)	+	95274	95353	80	+	
trnD(GUC)	-	36484	36411	74	-		*trnN*(GUU)	+	111788	111859	72	+	
trnY(GUA)	-	36643	36562	82	-		*trnR*(ACG)	-	112638	112565	74	-	
trnE(UUC)	-	36787	36715	73	+		*trnA(UGC)	-	117208	116368	841	-	intron 768 bp
trnT(GGU)	+	38367	38438	72	+		(exon 1)		117208	117171	38		
trnS(UGA)	-	41494	41407	88	+		(exon 2)		116402	116368	35		
trnG(GCC)	+	42035	42105	71	+		*trnI(GAU)	-	118242	117285	958	-	intron 886 bp
trnfM(CAU)	-	42229	42156	74	+		(exon 1)		118242	118206	37		
trnS(GGA)	+	48845	48932	88	+		(exon 2)		117319	117285	35		
trnT(UGU)	-	50333	50261	73	+		*trnV*(GAC)	-	120306	120235	72	+	

DNA strand (+/-) shows normal and reverse DNA strands, respectively, in our file. Promoter (+) indicates the presence of typical promoter sequences (-35 and -10) upstream from the coding sequences. Asterisks indicate the presence of introns in coding sequences. Position numbers are the same as in Table 1.

Table 3
Genes for coding proteins on the liverwort chloroplast genome

Genes	DNA strand	From	To	Length (bp)	Amino acid residues	Molecular weight	Promoter (-35, -10)	SD sequence	Remarks
(a) *Ribosomal protein genes*									
*_rps'12_	+	-	842	-	123	(13,797.0)	+	-	trans-split gene
(exon 2)		85	316	232	77				(see *_rps12'_)
(exon 3)		817	842	26	8				intron 2:500 bp
rps7	+	892	1359	468	155	17,821.9	-	-	
rps2	+	16055	16762	708	235	26,776.8	-	-	double stop TAA
rps14	-	42635	42333	303	100	11,879.8	+	-	
rps4	-	50033	49425	609	202	23,639.6	+	+	
rpl33	+	65273	65470	198	65	7,782.1	-	-	
rps18	+	65498	65725	228	75	8,879.5	-	-	
rpl20	-	66157	65807	351	116	12,773.0	-	-	
*_rps'12_	-	67057	-	-	123	(13,797.0)	-	-	trans-split gene
(exon 1)		67057	66944	114	38				(see *_rps'12_)
rps11	-	75249	74857	393	130	14,172.5	-	-	
rps8	-	76171	75773	399	132	14,921.4	-	+	
rpl14	-	76621	76253	369	122	13,496.6	-	-	
*_rpl16_	-	77685	76719	967	143	16,149.8	-	-	intron 535 bp
(exon 1)		77685	77677	9					
(exon 2)		77141	76719	423					
rps3	-	78396	77743	654	217	25,055.0	-	-	
rpl22	-	78804	78445	360	119	13,580.8	-	+	
rps19	-	79100	78822	279	92	10,553.3	-	+	
*_rpl2_	-	80514	79137	1,378	277	31,162.9	-	-	intron 544 bp
(exon 1)		80514	80118	397					
(exon 2)		79573	79137	437					
rpl23	-	80825	80550	276	91	10,768.5	-	+	
rpl21	+	93469	93819	351	116	13,626.1	+	-	
rps15	-	103699	103433	267	88	10,428.2	+	+	
(b) *RNA polymerase subunit genes*									
rpoB	+	5859	9056	3,198	1,065	120,445.6	+	+	
*_rpoC1_	+	9087	11737	2,651	684	78,959.6	-	+	intron 596 bp
(exon 1)		9087	9518	432					
(exon 2)		10115	11737	1,623					
rpoC2	+	11811	15971	4,161	1,386	160,153.4	-	+	
rpoA	-	74824	73802	1,023	340	39,240.2	-	-	
(c) *Genes for photosystems I and II*									
psbA	+	28368	29429	1,062	353	38,764.8	+	-	
psbD	+	38855	39916	1,062	353	39,388.6	+	+	
psbC	+	39864	41285	1,422	473	51,784.8	-	-	overlap to psbD
psaB	-	44928	42724	2,205	734	82,413.7	-	+	
psaA	-	47207	44955	2,253	750	83,212.4	+	+	
psbG	-	52524	51793	732	243	27,609.6	-	-	
psbF	-	63293	63174	120	39	4,468.3	-	+	
psbK	-	63554	63303	252	83	9,493.5	-	+	TAG stop
psbH	+	69026	70552	1,527	508	56,191.5	+	+	
psbI	+	71092	71316	225	74	7,928.3	-	-	double stop TAA

Table 3 (continued)

(d) Genes for H^+-ATPase subunits									
atpI	+	16890	17636	747	248	27,742.0	-	+	double stop TAA
atpH	+	18014	18259	246	81	8,004.4	+	+	
*atpF	+	18468	19609	1,142	184	21,080.8	+	+	intron 587 bp
(exon 1)		18468	18612	145					
(exon 2)		19200	19609	410					
atpA	+	19654	21177	1,524	507	55,311.7	-	-	
atpE	-	54362	53955	408	135	15,054.3	-	+	
atpB	-	55846	54368	1,479	492	53,137.2	+	-	
(e) Genes for proteins in the photoelectron transport									
petA	+	61641	62603	963	320	33,482.6	-	-	
*petB	+	71424	72566	1,143	215	24,306.5	-	+	intron 495 bp
(exon 1)		71424	71429	6					
(exon 2)		71925	72566	642					
*petD	+	72715	73690	976	.160	17,413.5	-	+	intron 493 bp
(exon 1)		72715	72722	8					
(exon 2)		73216	73690	475					
(f) Gene for ribulose bisphosphate carboxylase large subunit									
rbcL	+	56355	57782	1,428	475	5,2790.0	+	+	
(g) Genes for NADH-PQ oxidoreductase subunits									
*ndh2	+	1514	3555	2,042	501	5,6189.3	+	+	intron 536 bp
(exon 1)		1514	2239	726					TAG stop
(exon 2)		2776	3555	780					
ndh3	-	52877	52515	363	120	1,4188.7	+	+	overlap to psbG
ndh5	-	93179	91101	2,079	692	79,277.9	+	-	
ndh4	-	98164	96665	1,500	499	56,665.9	-	-	
ndh4L	-	99059	98757	303	100	11,197.3	-	+	double stop TAA.TAG
ndh6	-	99688	99113	576	191	21,606.7	-	-	URFG, URFC, TGA stop
*ndh1	-	102200	100382	1,819	368	41,509.6	-	+	intron 712 bp
(exon 1)		102200	101645	556					
(exon 2)		100932	100382	551					
(h) Genes for ferredoxin proteins and others									
frxA	-	98534	98289	246	81	8,941.2	+	+	
frxB	-	100330	99779	552	183	21,196.2	-	+	
frxC	-	110973	110104	870	289	31,944.5	+	-	TAG stop
lhcA	-	23605	23438	168	55	6,441.5	+	+	light harveting complex
mbpX	+	37012	38124	1,113	370	42,798.8	+	+	double stop TAA
mbpY	+	94183	95049	867	288	32,563.6	-	+	
secX	-	75413	75300	114	37	4,521.5	-	+	
infA	-	75686	75450	237	78	8,978.4	-	+	

Asterisks indicate the presence of introns in the coding sequences. Gene symbols are the same as in Fig. 2. Marks (+/-) in the columns of promoter and SD sequences indicate the presence and absence of typical prokaryotic promoter and SD sequences as seen in E. coli upstream from the coding sequences. The position number are the same as in Table 1.

Table 4

Unidentified open reading frames in the liverwort chloroplast genome

Genes	DNA strand	From	To	Length (bp)	Amino acid residues	Molecular weight	Promoter (-35, -10)	SD sequence	Remarks
ORF34	+	4001	4105	105	34	3,740.5	-	-	
ORF135	+	4236	5128	893	135	16,043.9	+	-	intron 485 bp
(exon 1)		4236	4341	106					TAC stop
(exon 2)		4827	5128	302					
ORF29	-	5257	5168	90	29	3,225.0	-	+	
ORF33	-	22263	22162	102	33	3,386.2	-	+	
ORF30	-	22425	22333	93	30	3,841.7	-	-	
ORF32	+	22516	22614	99	32	3,566.3	+	+	
ORF36a	-	23107	22997	111	36	4,152.9	-	+	
ORF513	+	24053	25594	1,542	513	58,326.9	-	+	
ORF50	+	25769	25921	153	50	6,382.5	-	+	
ORF370i	+	26976	28088	1,113	370	45,407.0	-	-	in trnK intron
ORF2136	+	29909	36319	6,411	2,136	259,908.8	-	-	
ORF62	+	41647	41835	189	62	6,538.8	+	+	
ORF167	-	48599	47488	1,112	167	19,509.5	+	-	intron 608 bp
(exon 1)		48599	48476	124					
(exon 2)		47867	47488	380					
ORF169	-	51742	51233	510	169	20,084.8	-	-	
ORF316	+	58065	59015	951	316	35,826.3	-	+	
ORF36b	+	59193	59303	111	36	4,017.8	-	+	
ORF184	+	59525	60079	555	184	21,533.1	-	+	
ORF434	+	60151	61455	1,305	434	51,866.2	-	-	
ORF40	-	62916	62794	123	40	4,101.8	-	+	
ORF38	-	63152	63036	117	38	4,479.1	-	-	
ORF42a	-	63684	63556	129	42	5,101.9	+	-	
ORF31	+	64152	64247	96	31	3,466.4	+	+	
ORF37	+	64370	64483	114	37	4,075.9	-	-	
ORF42b	+	65027	65155	129	42	4,746.6	+	+	TAG stop
ORF203	-	68640	67130	1,511	203	22,685.0	+	+	
(exon 1)		68640	68570	71					intron 1: 518 bp
(exon 2)		68051	67760	292					intron 2: 380 bp
(exon 3)		67378	67130	249					
ORF35	+	70669	70776	108	35	3,958.8	-	-	
ORF43	-	70994	70863	132	43	4,875.0	+	+	
ORF69	+	93886	94095	210	69	7,838.1	-	+	ribosomal protein
ORF320	+	95482	96444	963	320	37,169.7	-	-	
ORF392	-	103380	102202	1,179	392	45,371.4	-	+	
ORF464	-	105267	103873	1,395	464	57,164.9	-	-	
ORF1068	-	108535	105329	3,207	1,068	127,654.8	-	+	
ORF465	-	110064	108667	1,398	465	53,097.0	-	+	

DNA strand (+/-) indicates the coding sequences on either the normal or reverse strand in our file. The presence of typical prokaryotic promoter and SD sequences are shown as (+) when it can be seen upstream from the coding sequences. The position numbers given are the same as in Table 1.

Table 5

Sequences of transfer RNA genes in the liverwort chloroplast genome

	AMINOACYL-STEM		D-STEM	D-LOOP	D-STEM		ANT.-STEM	ANT.-LOOP	ANT.-STEM	EXTRA-ARM		Tψ-STEM	Tψ-LOOP	Tψ-STEM	AMINOACYL-STEM	
1. Ala(UGC)	GGGGATA	TA	GCTC	AGTT GGT A	GAGC	G	CCGCC	CTTGCAA▼	GGCGG	ATGT	C	AGCGG	TTCGAGT	CCGCT	TATCTCC	A
2. Arg(ACG)	GGGTTTG	TA	GCTC	AGA GGATTA	GAGC	A	CGTGG	CTACGAA	CCACG	GTGT	C	GGGGG	TTCGAAT	CCCTC	CTTGCCC	A
3. Arg(CCG)	GGGTTTG	TA	GCTC	AGT GGATTA	GAGC	T	CATGG	TTCCGAA	TCATG	AAGT	C	AAGGG	TTCGAAT	CCCTT	CTAACCC	T
4. Arg(UCU)	GCGTCCA	TC	GTCT	AAA GGAT A	GGAC	A	GAGGT	TTTCTAA	ACCTC	CAG	T	ATAGG	TTCGAAT	CCTAT	TGGACGT	A
5. Asn(GUU)	TCCTTAG	TA	GCTC	AGT GGT A	GAGC	G	GTCGG	CTGTTAA	CCGAT	TGGT	C	GTAGG	TTCAAAT	CCTAC	CTGAGGA	G
6. Asp(GUC)	GGGATTG	TA	GTTC	AATT GGTT A	GAGT	A	CCGCC	CTGTCAA	GACGG	AAGT	T	GCGGG	TTCGAGC	CCCGT	CAATCCC	G
7. Cys(GCA)	GGCGACA	TG	GCCA	AGT GGT A	AGGC	A	GAGGA	CTGCAAA	TCCTT	TAT	C	CCCAG	TTCAAAT	CTGGG	TGTCGCT	T
8. Gly(GCC)	GCGGGTA	TA	GTTT	AAT GGT A	AAAT	T	CCTCC	TTGCCAA	GGAGA	ATA	T	GCGGG	TTCGATT	CCCGC	TACCCGC	C
9. Gly(UCC)	GCGGGTA	TA	GTTT	AGT GGT A	AAA▼C	C	TTAGC	CTTCCAA	GCTAA	CGA	T	GCGGG	TTCGATT	CCCGC	TACCCGC	T
10. Gln(UUG)	TGGGGCG	TC	GCCA	AGT GGT A	AGGC	T	GCAGG	TTTTGGT	CCTGT	TATT	C	GGAGG	TTCGAAT	CCTTC	CGTCCCA	G
11. Glu(UUC)	GCCCCCA	TC	GTCT	AGT GGCCTA	GGAC	A	CCTCT	CTTTCAA	GGAGG	CGA	C	GGGGA	TTCGAAT	TCCCC	TGGGGGT	A
12. His(GUG)	GGCGGACG	TA	GCCA	AGT GGATTA	AGGC	A	GTGGA	TTGTGGA	TCCTC	TACG	C	GCGGG	TTCAATT	CCCGT	CGTTCGC	C
13. Ile(CAU)	GCATCCA	TG	GCTG	AAT GGTT A	AAGC	A	CCCAA	CTCATAA	TTGGC	GAATT	C	ACAGG	TTCAATT	CCTGT	TGGATGC	A
14. Ile(GAU)	GGGCTAT	TA	GCTC	AGT GGT A	GAGC	G	CGCCC	CTGATAA▼	GGGCG	AGGT	C	TCTGG	TTCAAGT	CCAGG	ATAGCCC	A
15. Leu(CAA)	GCCTTGA	TG	GTGA	AAT GGTA G	ACAC	G	CGAGA	TTCAAAA	TTTCG	TGCTTAAAGCA	T	GGAGG	TTCGAGT	CCTCT	TCAAGGC	A
16. Leu(UAA)	GGGGGTA	TG	GCGA	AATT GGTA G	ACGC	T	GCGGA	CTT▼AAAA	TCCGT	TGGCTTTAAAGACCG	T	GAGGG	TTCAAGT	CCCTC	TACCCCC	A
17. Leu(UAG)	GCCGCTA	TG	GTGA	AATT GGTA G	ACAC	G	CTGCT	CTTAGGA	AGCAG	TGCTAAGGCT	T	CTCGG	TTCGAAT	CCGAG	TAGCGGC	A
18. Lys(UUU)	GGGTTGC	TA	ACTC	AAT GGT A	GAGT	A	CTCGG	CTTTTAA▼	CCGAC	GAGT	T	CCGGG	TTCGAGC	CCCGG	GCAACCC	A
19. Met(CAU)	ACCTACT	TA	ACTC	AGT GGTTTA	GAGT	A	TCGCT	TTCATAC	GGCGA	GAGT	C	ATTGG	TTCAAAT	CCAAT	AGTAGGT	A
20. fMet(CAU)	CGCGGAG	TA	GAGC	AGTCTGGT A	GCTC	G	CAAGG	CTCATAA	CCTTG	AGGT	C	ATAGG	TTCAAAT	CCTGT	CTCCGGC	A
21. Phe(GAA)	GCCGGGA	TA	GCTC	AGTT GGT A	GAGC	A	GAGGA	CTGAAAA	TCCTC	GTGT	C	ACCAG	TTCAAAT	CTGGT	TTCTGGC	A
22. Pro(GGG)+	CGGAGTA	TA	GT	TTGGT A	GTGT	A	TCATC	TTGGGGT	GATGA	AAGT	C	GTGGG	TTCAAAT	CCCGC	TACTCAA	A
23. Pro(UGG)	AGGGATG	TA	GCGC	AGTTTGGT A	GCGC	G	TTTGT	TTTGGGT	ACAAA	ATGT	C	GCAGG	TTCGAAT	CCTGT	CATCCCT	A
24. Ser(GCU)	GGAGAGA	TG	GCCG	AGT GGACGA	AAGC	G	GCGGA	TTGCTAA	TCCGT	TGTACAAGCTTTTTGTACC		GAGGG	TTCGAAT	CCCTC	TCTCTCC	G
25. Ser(GGA)	GGAAAGA	TG	GTTG	AGT GGTTTA	AGGC	G	TAGCA	TTGGAAA	TGCTA	TGTAGGCTTTTGGTCTATC		GAGGG	TTCGAAT	CCCTC	TCTTTCC	G
26. Ser(UGA)	GGAGAGA	TG	GCCG	AGT GGTTTA	TGGC	G	TCGGT	CTTGAAA	ACCGA	TATAGTTTTTAAGATTATC		GAGGG	TTCAAAT	CCCTC	TCTCTCC	T
27. Thr(GGU)	GCCCTTT	TA	ACTC	AGT GGT A	GAGT	A	ACGCC	ATGGTAG	GGCGT	AAGT	C	ATCGG	TTCAAAT	CTGAT	AAAGGGC	T
28. Thr(UGU)	GCCTGTT	TA	GCTC	AGA GGTC A	GAGC	G	TCGCA	CTTGTAA	TGCGA	TGGT	C	ATCGG	TTCGACT	CCGAT	AGCGGGC	T
29. Trp(CCA)	GCGCTTT	TA	GTTC	AGTTCGGT A	GAAC	G	TAGGT	CTCCAAA	ACCTA	ATGT	C	GTAGG	TTCAAAT	CCTAC	AGAGCGT	G
30. Tyr(GUA)	GGGTCGA	TG	CTCG	AGT GGTTAA	TGGG	G	ACGGA	CTGTAAA	TCCGC	TGGCAATGCCTA	C	GCTGG	TTCAAAT	CCAGC	TCGACCC	A
31. Val(GAC)	AGGGATA	TA	ACTC	AGC GGT A	GAGT	A	TCACC	TTGACGT	GGTGG	AAGT	C	ATCAG	TTCGAAC	CTGAT	TATCCCT	A
32. Val(UAC)	AGGGCTA	TA	GCTC	AGC GGT A	GAGC	G	CCTCG	TTTACAC▼	CGAGA	ATGT	C	TACGG	TTCAAAT	CCGTA	TAGCCCT	A

The marks *, **, ---, and == show the position of stem-loop structure conformation. Pro(GGG)+ has an incomplete aminoacyl stem structure, which was not detected by Northern hybridization. Arrowheads indicate the splicing sites of introns.

complete liverwort nucleotide sequence starts at the first nucleotide of the large single-copy (LSC) region in the junction (J_{LA}). The numbering proceeds counterclockwise around the chloroplast genome and ends at the last nucleotide in the IR_A region (Fig. 2).

Identification of coding sequences for RNA and protein genes

The coding sequences for four kinds of rRNA genes (16S, 23S, 4.5S, and 5S), 31 species of tRNA genes, and a proline $tRNA_{GGG}$-like sequence in which the amino-acyl stem structure is incomplete were identified (Table 2). Coding sequences for tRNAs were searched for using the T ψ loop sequence (GTTCRA) and identified by construction of the familiar clover-leaf structure. The coding sequences of the 31 species of tRNAs and the proline $tRNA_{GGG}$-like sequence are listed in Table 5. None of them had a CCA sequence at the 3'-end of their coding sequences. Six species, alanine $tRNA_{UGC}$, isoleucine $tRNA_{GAU}$, glycine $tRNA_{UCC}$, lysine $tRNA_{UUU}$, leucine $tRNA_{UAA}$, and valine $tRNA_{UAC}$, had introns in their coding sequences. Their codon-anticodon table is shown in Table 6. The tRNAs encoded by the chloroplast genome were sufficient to read all codons taking into account the expanded wobble and modification in the anticodons (Crick, 1966b).

As shown in Table 3, there are coding sequences for 19 ribosomal proteins (large subunit proteins L33, L20, L14, L16, L22, L2, L23, and L21, and small subunit proteins S7, S2, S14, S4, S18, S12, S11, S8, S3, S19, and S15; Post & Nomura, 1980; Zurawski & Zurawski, 1985; Posno et al., 1986). In addition, the genes infA (Pon et al., 1979) and secX (X gene, Cerretti et al., 1983) and the rpo genes for α , β , and β ' subunits of RNA polymerase were identified by correspondence to E. coli genes (Bedwell et al., 1985). Two ORFs (mbpX and mbpY) showed homology to inner membrane permease proteins found in S. typhimurium (Higgins et al., 1982, 1986).

Genes for photosynthesis were identified by comparison of the amino-acid sequences deduced to those deduced from chloroplast genes of

Table 6

Codon table and unmodified anticidons of tRNAs coded by liverwort chloroplast

Codon	Anticodon		Codon	Anticodon		Codon	Anticodon		Codon	Anticodon	
UUU	Phe		UCU	Ser		UAU	Tyr		UGU	Cys	
UUC		GAA	UCC		GGA	UAC		GUA	UGC		GCA
UUA	Leu	UAA*	UCA		UGA	UAA	Ter		UGA	Ter	
UUG		CAA	UCG			UAG	Ter		UGG	Trp	CCA
CUU	Leu		CCU	Pro		CAU	His		CGU	Arg	ACG
CUC			CCC		GGG	CAC		GUG	CGC		
CUA		UAG	CCA		UGG	CAA	Gln	UUG	CGA		
CUG			CCG			CAG			CGG		CCG
AUU	Ile		ACU	Thr		AAU	Asn		AGU	Ser	
AUC		GAU*	ACC		GGU	AAC		GUU	AGC		GCU
AUA		CAU	ACA		UGU	AAA	Lys	UUU*	AGA	Arg	UCU
AUG	Met fMet	CAU CAU	ACG			AAG			AGG		
GUU	Val		GCU	Ala		GAU	Asp		GGU	Gly	
GUC		GAC	GCC			GAC		GUC	GGC		GCC
GUA		UAC*	GCA		UGC*	GAA	Glu	UUC	GGA		UCC*
GUG			GCG			GAG			GGG		

The AUG codon is an initiation codon. Termination codon (UAA, UAG and UGA) are indicated by Ter. Asterisks indicate the tRNA genes with introns in their coding sequences. Amino acids are shown by 3-letter symbols.

various sources as reported previously (Crouse et al., 1985; Hird et al., 1986; Westhoff et al., 1986). Seven of the identified ORFs had high homology to NADH dehydrogenase genes (ND1, ND2, ND3, ND4, ND4L, ND5, and ND6) found in human mitochondria (Chomyn et al., 1985, 1986), and two ORFs (*frxA* and *frxB*) had high homology to bacterial 4Fe-4S-type ferredoxin proteins (Yasunobu & Tanaka, 1980). The *frxA* gene product has been identified as a component of the photosystem I particles (Oh-oka et al., 1987). An ORF (*frxC*) consisting of 289 amino acids showed extensive homology to the *nifH* gene of *Az. vinelandii* (Howard et al., 1983) and ORF202 of *R. capsulata* (Hearst et al., 1985). The ORF55 (*lhcA*) had homology to antenna proteins of LHC of photosynthetic bacteria (Youvan et al., 1984).

Physicochemical characteristics

Liverwort chloroplast DNA contained genetic information encoding 136 possible genes with short spacer regions very rich in A + T. The total G + C content of the liverwort chloroplast genome was 28.8% (G, 17,556; C, 17,308; A, 42,898; and T, 43,263 in normal (+)-strand DNA (Tables 2 - 4). However, the coding sequences of rRNA and tRNA genes had higher G + C contents, 52.6% and 52.1%, respectively, and that in the coding sequences for proteins was 28.5%. The spacer regions between the coding sequences had much proportions of G + C content (19.5%; Fig. 3).

Conservation of high free energy, ΔG, calculated for stem-loop structures deduced from a nucleotide sequence may give information on the termination sites of transcription and sites of processing. Free energy conservation in stem-loop structures was greater at the positions between the genes of ends transcribed convergently from opposite strands of the liverwort chloroplast genome (Table 7).

Thirty-three ORFs remain unidentified with use of the NBRF database (Table 4). Membrane spanning analysis was done using the program of Klein et al. (1985). This program gives precise trans-

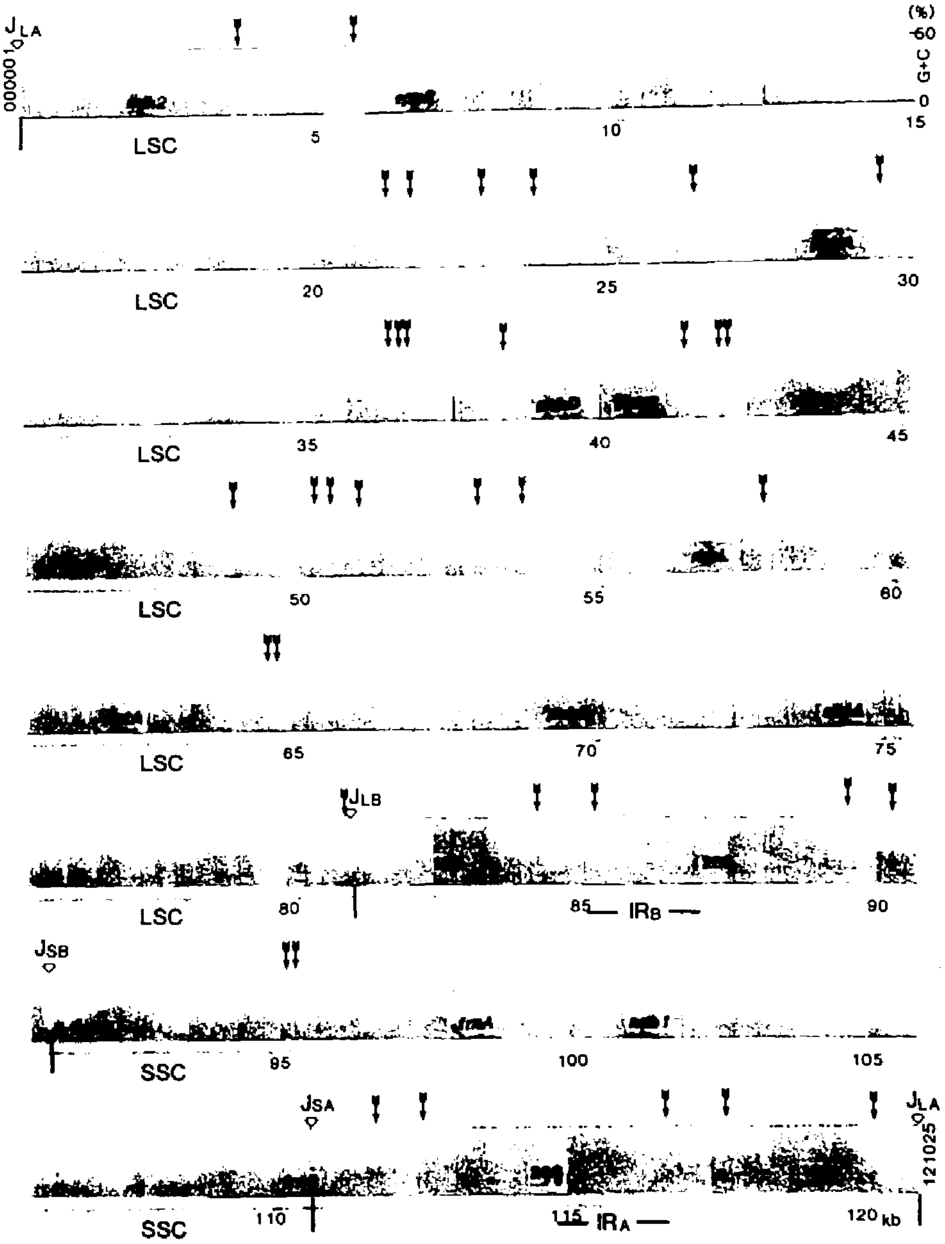


Fig. 3. G + C content throughout the liverwort chloroplast genome.
The G + C content in every 25 nucleotides (1 to 121025) are shown as percentages. Major genes are drawn as landmarks. The abbreviations IR, LSC, SSC are the same in Fig. 2. J_{LA}, J_{LB}, J_{SA}, and J_{SB} indicate junctions between IR_A and LSC, IR_B and LSC, IR_A and SSC, and IR_B and SSC, respectively. Open regions in the bottom line indicate the spacer (non-coding) regions. Arrows indicate coding regions for transfer RNA genes, which in most cases have higher G + C content.

Table 7
stem-loop structures at the end of convergent transcriptional units

Location	Number of stem-loop structures (no. base-pairs)	Free energy (Δ*G*)
ORF135–ORF29	1 (17)	−20·5
atpA–*trnR*(UCU)	1 (12)	−5·0
trnS(GCU)–ORF36a	1 (7)	−11·4
ORF2136–*trnD*(GUC)	1 (23)	−35·4
psbC–*trnS*(UGA)	1 (43)	−50·4
trnG(GCC)–*trnfM*(CAU)	1 (11)	−8·0
trnS(GGA)–*rps4*	5 (42, 7, 20, 17, 7)	−66·3, −6·3 −20·9, −16·8, −6·3
trnF(GAA)–ORF169	2 (35, 10)	−45·8, −6·9
trnM(CAU)–*atpB*	1 (18)	−28·0
petA–ORF40	3 (12, 20, 9)	−7·2, −30·1, −3·3
ORF37–*trnW*(CCA)	1 (10)	−13·2
rps18–*rpl20*	1 (13)	−23·6
ORF35–ORF43	2 (13, 26)	−7·0, −30·9
petD–*rpoA*	2 (7, 38)	−9·0, −55·4
trnR(ACG)–*trnN*(GUU)	1 (36)	−21·7
ORF320–*ndh4*	1 (30)	−36·6
trnN(GUU)–*trnR*(ACG)	1 (36)	−21·7

Free energy conserved in the stem-loop structure was calculated as described by Tinoco et al. (1973) and Salser (1977) and expressed as G (kcal:1kcal=4.184kJ).

Table 8
Membrane-spanning analysis of all liverwort chloroplast genes

Peripheral location			Integral location		
Confirmed		Predicted	Confirmed	Predicted	
atpA	*rpl2*	*frxA*	*atpF*	*frxB*	ORF36a
atpB	*rpl14*	*psbG*	*atpH*	*frxC*	ORF50
atpE	*rpl16*	*rpoA*	*atpI*	*mbpX*	ORF2136
rbcL	*rpl20*	*rpoC1*	*petA*	*mbpY*	ORF62
rps2	*rpl21*	*rpoC2*	*petB*	*ndh1*	ORF36b
rps3	*rpl22*	ORF513	*petD*	*ndh2*	ORF184
rps4	*rpl23*	ORF370i	*psaA*	*ndh3*	ORF434
rps7	*rpl33*	ORF167	*psaB*	*ndh4L*	ORF40
rps8	*secX*	ORF169	*psbA*	*ndh4*	ORF38
rps11	*infA*	ORF316	*psbB*	*ndh5*	ORF31
rps12		ORF42a	*psbC*	*ndh6*	ORF37
rps14		ORF60	*psbD*	*lhcA*	ORF42b
rps15		ORF302	*psbE*	*rpoB*	ORF203
rps18		ORF464	*psbF*	ORF29	ORF35
rps19			*psbH*	ORF34	ORF43
				ORF135	ORF320
				ORF33	ORF1068
				ORF30	ORF465
				ORF32	

Computer analysis was done with the program of Klein et al. (1985). Peripheral and integral in the columns Confirmed indicate the location of stromal proteins and proteins bound to the thylakoid membrane, respectively.

membrane characteristics to the known gene products in the thylakoid membrane except for the psbG gene product, which appears as a peripheral protein. Therefore, this analysis may indicate whether unidentified proteins are present in the membrane complex of the chloroplasts (Table 8). All ndh genes were categorized into this group of proteins with transmembrane characteristics.

DISCUSSION

Inverted repeats and gene rearrangements

The liverwort chloroplast DNA was composed of 121,025 bp, which is comparatively smaller than chloroplast genomes from other sources. The inverted repeats (IR_A and IR_B, each 10,058 bp) were smaller than those (some 26 kb) of tobacco (Shinozaki et al., 1986b). This suggests that there were rearrangements in the chloroplast genome in the vicinity of the IR regions. The IR regions in the liverwort chloroplast DNA did not have the clustered ribosomal protein operons present in both IR regions in tobacco (Shinozaki et al., 1986b). Palmer & Stein (1986) have reported that a fern chloroplast genome, which has IRs of similar size with that of the liverwort, has lost one of the the clusters of ribosomal protein genes. There are no ORF within the IR regions of the liverwort chloroplast genome.

Introns

Introns are present in the genes of eukaryotic cells. Chloroplasts are organelles in photosynthetic eukaryotic cells with a prokaryotic type of gene expression mechanism and introns in their coding sequences were first reported in chloroplasts of C. reinhardtii by Rochaix & Malnoe (1978). The leucine $tRNA_{UAA}$ gene had a group I intron that forms the secondary structure typical of group I introns (Bonnard et al., 1984). Introns in coding sequences were predicted using the 5'-consensus sequence (GTGPyG) and 3'-consensus secondary structures (PuAGCCGNATGAANNGA AANNTTCATGTNCGGTTPy and CTAPyPyNYNAPy) that are characteristic of group II introns found in fungal mitochondrial genes (Michel & Dujon, 1983). There were 22 introns in the chloroplast genome (Table 9; genes for alanine $tRNA_{UGC}$ and isoleucine $tRNA_{GAU}$ with introns were duplicated in IR regions). Ribosomal protein S12 gene (rps12) has two introns, one of which is far away on the opposite DNA strand (Fukuzawa et al., 1986). The ORF203 also had two introns in its coding sequence just upstream from the rps12' gene. The 5'-exon boundary sequences in the ORF203 introns (ATGCG in the first and TTGTG in the second) were different from the typical 5'-consensus sequence of GTGPyG. However, Northern hybridization experiments with synthetic probes that have sequences of the two connecting exons showed that the splicing site was exactly at the predicted junctions (Kohchi et al., 1988d).

Group II introns also were found in the coding sequences for tRNA genes (isoleucine $tRNA_{GAU}$, alanine $tRNA_{UGC}$, glycine $tRNA_{UCC}$, lysine $tRNA_{UUU}$, and valine $tRNA_{UAC}$). Introns in alanine $tRNA_{UGC}$ and isoleucine $tRNA_{GAU}$ had different 5'-exon boundary sequences (TTGGG and TTGCG, respectively); however, the introns belonged to the category of group II introns because of the presence of consensus secondary structures in the introns. Group II introns in tRNA genes were characteristic of the chloroplast genome of land plants.

Coding sequences for protein in photosystems I and II did not contain introns, although genes for a subunit of H^+-ATPase (atpF) and for proteins in the photoelectron transport system (petB and petD) had introns in their coding sequences. Interestingly, the first exons in both (petB and petD) genes consist of only a few amino acids and these introns are precisely spliced out (Fukuzawa et al., 1987).

Codon usage

As described above, there are coding sequences for 31 species of tRNA that can read all sense codons by expanded wobble codon-anticodon recognition, as seen in yeasts. Therefore, it is not needed to assume that there is transport of tRNA from the cytoplasm. The codon usages

Table 9

The 5'- and 3'-consensus sequences of introns found in liverwort chloroplast genome

Group I intron		
	/ 5' intron 3' intron /	
Leu(UAA)	ACTT/AATTTAATTGAGCTTTAGTTGAGAAATTTACTAAATGATT.......TACAAGTTAA[illegible]TTTAACAACAATGCAAATTGTAGTAAAATG/AAAA	315
Group II introns		
	/ 5' intron 3' intron /	
Gly(UCC)	AAAA/GTGCGAT-TCGT......AAGGAGCCGAATGAAAG-AAAACTTTCACGTTCGGTTTTGAATTAGAGGC..(20)..GTCGACTATAAC/CCTT	593
Lys(UUU)	TTAA/GTGCGAC-TTGG......AGAAAGCCGTATGCAGT-AAAAATTGCAAGTACGGTTTGGGAAGAGATGA..(29)..ATCTACT-TCAT/CCGA	2,111
Val(UAC)	ACAC/GTGCGCCAATGC......AACGAGCCCAATGCATA-AAAACATGCATGTTGGGTTCTTAAAGCAGTTC..(12)..AACTGTT-TTAC/CGAG	530
Ile(GAU)	ATAA/TTGCGTCGTTGT......GGAGAGCGCAGTACAACGGAAAGTTGTATGCTGCGTTCGGGAAGGATGAA..(64)..ATTTACT-TCAC/GGGC	886
Ala(UGC)	GCAA/TTGGGTCGTTGC......GGAGAGCACAGTACGAT-GAAAGTTGTAAGCTGTGTTTGGGGGGGAGTTA..(55)..GCTTACC-CTGT/GGCG	768
ndh2	AGGA/GTGCGAT-TCGT......TAGGAGCCGTGTGAATT-GAAAATCTCATGCACGGTTTTGAATGAGAGAA..(16)..TTCGACTCTAAC/TCAC	536
ORF135	AGAG/GTGTGAT-TTAA......TTTAAGCCATACAGAGTTGAAAATATCATATATGGTTTTCAAGGGGGGAA..(20)..ACCTATCCTAAT/ATTA	485
rpoCI	CGAT/GTGTGAC-TTGA......TTTGAGCCGGATGACGG-AAAACTTTCATGTCCGATTCTTAGGGGGGGAA..(11)..ACCTATCCCAAT/CTCT	596
atpF	GTGT/GTGCGG-GTTGA......AGAAAGCCGAATGAATT-GAAAAGTTCATGTTCGGTTTGGGAAGAGATTA..(15)..ATCTACTTTCAT/TAAG	587
ORF167	GATG/GTGTGAT-TTGA......GAGGAGCCGTATGAAGT-TTAAACTTCATGTACGGTTTTGAAACGGAGTT..(15)..AACAACCGTAAC/GAAT	608
petB	GGGT/GTGCGTC-TTGT......TTTAAGCTGTAAGATTA-TAAATAATCATTTACGGTTTTTCGAGGGGGAA..(8)..ACCTATCTCAAT/AAAG	495
petD	GAGT/GTGTGACTTTAT......TTGGAGCCGGATGATAT-TAAATTATCATGTCCGATTCTTTGGGGGGACT..(6)..ATCTACCTTAAT/AACA	493
rpl16	TAGT/GTGTGAC-TCGT......GAGGAGCCGGATGAATC-AAAA-TTTCATGTCCGGTTTTGAAGTAGCGAT..(3)..ATCGACTATAAC/CCTA	535
rpl2	TTGA/GTGCGGT-TTGA......GAAAAGCTGTATGC-TT-GAAAAAAGCTTGTACAGTTTGGGAAGAGATTT..(18)..ATCTACT-TCAA/CCAA	544
ndhI	CTAC/GTGTGAT-TCGT......GAGGAGCCGTATGAAAT-GAAAATTTCATGTACGGTTTTGCAATAGAGAT..(18)..ATCGACTATAAT/TATC	712
rps12-1	GTAT/GTGCGAC-TTGT..//..GAGAAGCCGTATGAAAT-GAAAATATCAAGTACGGTTTTGTAAAGTGACA..(17)..GTCAACTTTTCC/ACTA	???
rps12-2	TCTA/GTGCG---TTGT......AAAAAGCCGTATTCGTT-GAAAATCGGATGTACGGTTTGGAGGGAGATAA..(4)..ATCCACC-CTAC/AATA	500
ORF203-1	TATA/ATGCGCC-TTAT......CTTAAGCTGTATGCGCTTAAAAAGTGCTTGTACAGTTTTATAAGAAAAAA..(7)..AATTATCTTAAT/CAAT	518
ORF203-2	CGCT/TTGTGCCAATGA......ATAGAGCTGTATGCAAC-TAAAAATGCATGTACAGTTCGTTTCATTTATT..(27)..ATATTTATTAAT/AGGG	380
CONSENSUS	/GTGYG RAGCCGNATGAANN-GAAANNTTCATGTNCGGTTY CTAYYNYNAY	

Numbers in the right column show the length of the introns. The letters Y and R indicate pyrimidine and purine nucleotide, respectively. Numbers in the paretheses indicate the length of nucleotides deleted in this Table. The abbreviations Leu(UAA), Gly(UCC), Lys(UUU), Val(UAC), Ile(GAU) and Ala(UGC) are used for genes of leucine tRNA$_{UAA}$, glycine tRNA$_{UCC}$, lysine tRNA$_{UUU}$, valine tRNA$_{UAC}$, isoleucine tRNA$_{GAU}$, and alanine tRNA$_{UGC}$, respectively.

of all of the genes, including the unidentified ORFs, were analyzed. The results showed that 88.1% of the third position of codons was either A or T (Table 10). This A or T preference could also be seen in the nonsense codon usage in the liverwort chloroplast genome. Of the 91 stop codons in the entire genome, 84 TAA codons were used. It was of interest that a TGA stop codon was used only twice (ORF135 and *ndh6* genes) (Table 10). In the mitochondrial genome of yeast (Bonitz *et al.*, 1980) and humans (Anderson *et al.*, 1981), the TGA codon is used for the recognition of tryptophan tRNA. From the evolutionary point of view, the rare usage of the TGA stop codon in the liverwort genome indicates the tendency of the TGA stop codon to convert to a codon for tryptophan tRNA in the organelle genome (Jukes *et al.*, 1987; Osawa & Jukes, 1988). This A or T preference pressure in the third position of codons with the exception of the *psbA* gene also coincided with the overall high A + T content in the liverwort chloroplast genome and has facilitated ORFs throughout the genome.

Membrane spanning analysis

The membrane spanning analysis on a computer showed that the gene products of *atpH*, *atpI*, and *atpF* had sequences with the transmembrane character, but that those of *atpA*, *atpB*, and *atpE* did not. These results coincided well with the identification of the location of the gene products of *atpH*, *atpI*, and *atpF* as being the thylakoid membrane. The liverwort chloroplast gene organization clearly had clustering of similar functional genes and we can use this together with the properties of the sequences to suggest possible functions of unidentified ORFs. For example, ORF38 and ORF40 may be gene products for photosystem II because of their transmembrane characteristics and their location in the identified *psb* gene cluster (*psbE* and *psbF*) of *C. paradoxa* cyanelle (Cantrell & Bryant, 1988).

Among the unidentified ORFs, two large ORFs (ORF1068 and ORF2136) in the category of membrane association have a high basic

Table 10

Total codon usage including ORFs of liverwort chloroplast genome

Codon		Total	Codon		Total	Codon		Total	Codon		Total
Phe	UUU	1,547	Ser	UCU	628	Tyr	UAU	826	Cys	UGU	219
	UUC	97		UCC	71		UAC	95		UGC	41
Leu	UUA	1,867		UCA	359	Ter	UAA	84	Ter	UGA	2
	UUG	203		UCG	48		UAG	5	Trp	UGG	441
Leu	CUU	524	Pro	CCU	477	His	CAU	388	Arg	CGU	357
	CUC	25		CCC	39		CAC	62		CGC	47
	CUA	143		CCA	367	Gln	CAA	887		CGA	258
	CUG	25		CCG	49		CAG	53		CGG	22
Ile	AUU	1,519	Thr	ACU	616	Asn	AAU	1,256	Ser	AGU	414
	AUC	101		ACC	57		AAC	175		AGC	46
	AUA	708		ACA	499	Lys	AAA	1,764	Arg	AGA	382
Met	AUG	521		ACG	42		AAG	78		AGG	24
Val	GUU	648	Ala	GCU	779	Asp	GAU	735	Gly	GGU	627
	GUC	47		GCC	66		GAC	72		GGC	82
	GUA	450		GCA	452	Glu	GAA	1,133		GGA	678
	GUG	48		GCG	50		GAG	85		GGG	88

Numbers indicate frequency of codons used. Amino acids are shown by 3-letter symbols.

amino acid composition. ORFs of similar size in tobacco have also been described (Shinozaki et al., 1986b), and to some extent, the nucleotide sequences of liverwort and tobacco are similar (data not shown), so they may be functional genes. However, protein molecules corresponding to such large ORFs have not been detected in chloroplasts, indicating the possibility of either the presence of introns or protein processing if these genes are transcribed and translated.

Gene clusters and transcriptional units

Typical prokaryotic promoter sequences were found upstream of gene clusters although the rest of the genes in the cluster had no promoter sequences but sometimes had SD sequences (Table 11).

Gene clusters observed in the liverwort chloroplast genome are very similar to those in *E. coli* (*rif*, *unc*, *str*, *spc*, S10, α, and *rrn* operon (Umesono *et al.*, 1988; Fukuzawa *et al.*, 1988; Kohchi *et al.*, 1988a). On the other hand, tRNA genes were scattered throughout the liverwort chloroplast genome. In addition to this structural similarity of gene organization, sequence homologies of the genes also indicate the prokaryotic characteristics of the machinery of chloroplast protein synthesis.

Conserved gene contents in chloroplasts

The gene organization of the liverwort chloroplast genome was very compact. The liverwort chloroplast genome contains all of the chloroplast genes described for other species except for the genes for ribosomal protein S16 of tobacco (Shinozaki *et al.*, 1986a) and elongation factor (*tufA*) of *Eu. gracilis* (Montandon & Stutz, 1983), although most higher plant chloroplast genomes are much larger (about 160 kb). The nucleotide sequence of the *N. tabacum* chloroplast genome has been described (Shinozaki *et al.*, 1986b). The genome size (approximately 121 kb) of the liverwort chloroplasts is smaller than that (approximately 155 kb) of tobacco chloroplasts, but the number of gene species coded on their genome is almost the same (127 = 136 minus 9 duplicates in liverwort; 128 = 156 minus 26 duplicates in tobacco). This indicates that liverwort chloroplast genome has at least an indispensable gene composition among the green plants. Therefore, in addition to well-established cultured cells being available (Ohta *et al.*, 1977; Ono *et al.*, 1979), the liverwort chloroplast has other excellent features for use as a model system for studying the molecular basis of chloroplast gene expression and photosynthesis of green plants.

Table 11

Promoters (-35 and -10 sequences) and SD sequences detected in the liverwort chloroplast genome

Genes	Promoters	SD sequences
	-35 sequence -10 sequence	
(a) *Ribosomal RNA genes*		
rrnA/rrnB	TTGACA---(18)---TATACT	(IR_A and IR_B)
(b) *Transfer RNA genes*		
trnC(GCA)	TTGAAT---(17)---TATTAT	
trnS(GCU)	TTGAAA---(18)---TATAAT	
trnQ(UUG)	TTGACA---(18)---TTTACT	
trnK(UUU)	TTGACA---(17)---TATAAT	
trnE(UUC)	TTGACA---(17)---TAACAT	
trnT(GGU)	TTGACA---(18)---CACAAT	
trnS(UGA)	TTGTAA---(18)---TATAAT	
trnG(GCC)	TTATAT---(17)---TATAAT	
trnfM(CAU)	TTGTTT---(18)---TATACT	
trnS(GGA)	TTGCTA---(18)---TACAAT	
trnT(UGU)	TTGCTT---(17)---TATAAT	
trnL(UAA)	TTGCAT---(17)---TATAAT	
trnV(UAC)	TTGTAT---(15)---TATAAC	
trnM(CAU)	TTGCAT---(18)---TATAAT	
trnP(UGG)	TAGGCA---(17)---TACAAT	
trnI(CAU)	TTTAAG---(18)---TATAAT	
trnV(GAC)	TTGTTT---(18)---TATAAT* TTGAAT---(16)---TATTAT*	(IR_A and IR_B)
trnN(GUU)	TTGCTA---(17)---TATAAT	(IR_A and IR_B)
trnL(UAG)	TTAAAA---(19)---TATAAT	
(c) *ribosomal protein genes and others*		
rps'12	TTTAAG---(18)---TATAAT	- (IR_A)
rps14	TTGCAA---(17)---TTTAAT	-
rps4	TTGATT---(19)---TATAAT* TTGTTT---(16)---TATAGT*	 AGGAG
rps8	-	AGGAG
rpl22	-	AGGAG
rpl19	-	AGGA
rpl23	-	AGGAG
rpl21	TTGAAA---(16)---TCTACT	-
rpl15	TTGTAT---(18)---TTTACT	AGGAG
infA	-	AGGAG
secX	-	GAC
rpoB	TTGAAC---(17)---TATAAC	GAGG
rpoC1	-	GGA
rpoC2	-	GAGG
mbpX	TTGAAT---(20)---TGTTAT	AGGAG
mbpY	-	AGG

Genes	Promoters	SD sequences
	-35 sequence -10 sequence	
(d) *Genes for photosynthesis*		
rbcL	TTGCAT---(18)---TACAAT	GGAGG
psbA	TTGACA---(20)---TACTAT	-
psbD	TTGAAA---(19)---TACAAT	AGGAG
psaB	-	AGGA
psaA	TTGAAG---(20)---TATAAT	AGGAG
psbC	-	AGGAG
psbF	-	AGGAGG
psbE	-	GGAG
psbB	TCTCCA---(15)---TAGAAA	AGG
petB	-	AGGA
petD	-	AGGA
atpI	-	AGGAG
atpH	TTGTTT---(18)---TATAAT	AGGAG
atpF	TTGAGT---(20)---TTCAAT	GAGG
atpE	-	GGAG
atpB	TTGTCA---(17)---TTTAAT	-
ndh2	TTGTTT---(17)---TTTAAT	GGA
ndh3	TTGACC---(17)---TATAAA	AGGAGG
ndh5	TTATAA---(17)---TATAAA	-
ndh4L	-	AGGA
ndh1	-	GGAGG
frxA	TTGATA---(17)---TATAAA	AGGAG
frxB	-	AGGA
frxC	TTGATG---(18)---TAGACT	-
lhcA	TTGTAG---(19)---TATACT	GGAG
(e) *Unidentified ORFs*		
ORF29	-	AGGA
ORF135	TTGAAA---(18)---TATATT	-
ORF33	-	GAGG
ORF32	TTGGAT---(20)---TATATT	AGGAG
ORF36A	-	GGAG
ORF513	-	AGGAG
ORF50	-	AGGAG
ORF62	TTGATA---(17)---TTTGAT	GGAG
ORF167	TTGAAT---(17)---TTGAAT	-
ORF316	-	GAGG
ORF36b	-	AGG
ORF184	-	AGGAG
ORF40	-	AGGAG
ORF42a	TTGAAC---(18)---TATATT	-
ORF31	TTGTCA---(17)---TAATAT	AGGA
ORF42b	TTGTTT---(14)---CATAAT	AGGAG
ORF201	TTGTAT---(22)---TATAAT	GGA
ORF43	TTGAGA---(15)---TACTAT	GGA
ORF69	-	AGG
ORF392	-	AGGAG
ORF1068	-	GAGG
ORF465	-	AGGA

Genes have either promoter or SD sequences are listed. Numbers in parentheses indicate the nucleotide base-pairs between -35 and -10 sequences. The sequence of the 3' end of liverwort 16S rRNA, CCTCCT, is complementary to SD sequences listed above. Asterisks indicate 2 possible promoter sequences.

I-2 Structure and organization of Marchantia polymorpha chloroplast genome — Inverted repeat and small single copy regions

MATERIALS AND METHODS

Procedures for chloroplast DNA preparation and cloning and sequencing methods were as described in the chapter I-1. The open reading frames (ORF) were identified by the program IDEAS with the protein database NBRF (release 8 to 10).

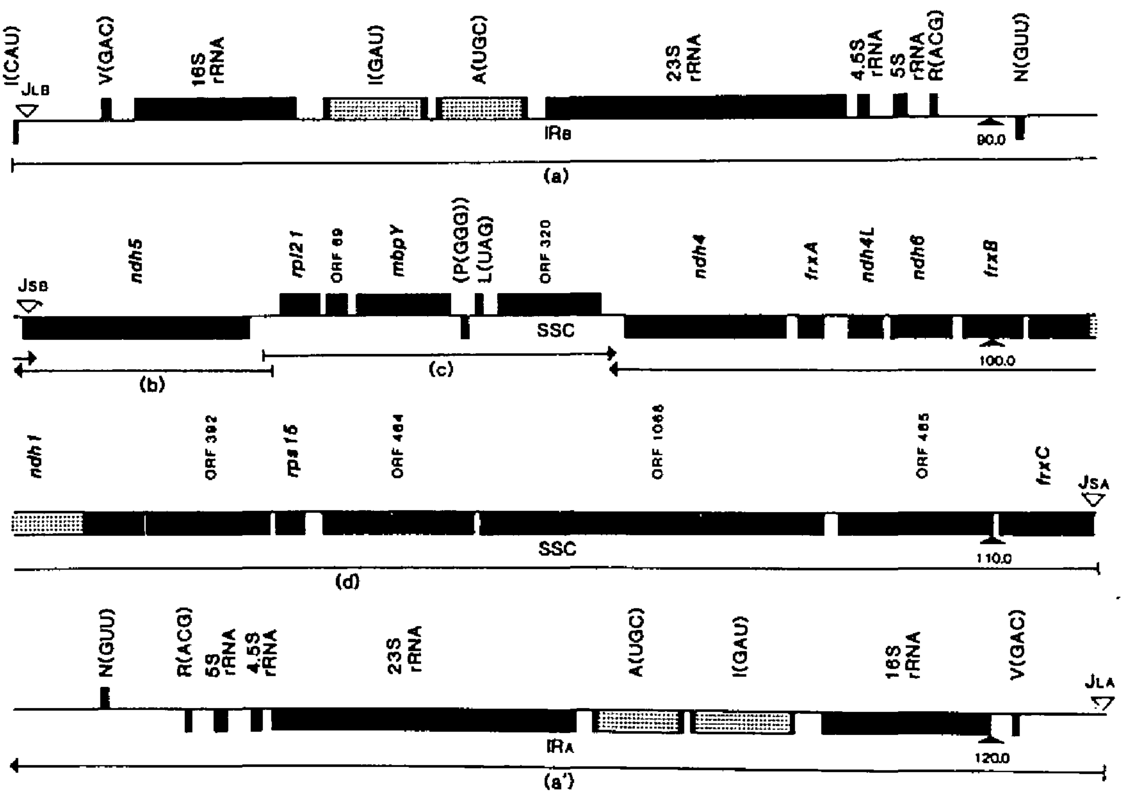


Fig. 1. Gene organization in the IR_A, SSC, and IR_B regions.
The coding region of genes are shown as boldface boxes. Introns are indicated by hatched boxes. Genes shown on the lines are transcribed to the right, and those under the lines are transcribed to the left. Blocks a to d shown by the arrows correspond to the nucleotide sequence files in Fig. 2a to d. Abbreviations for the genes are the same as described in the preceding chapter.

RESULTS

Gene organization of the IR region

A schematic diagram of the gene organization of the IR regions is given in Fig. 1a and a'. The DNA sequences of both the IR_A and IR_B regions were identical (10,058 bp; Fig. 2a). There were genes for nine stable RNA species in each IR region. The nucleotide sequence and deduced information (genes, introns, promoters, and stem-loop structures) are also presented in Fig. 2.

It has been mapped that the region of the liverwort chloroplast genome coding for rRNAs by Southern hybridization with ^{32}P-labelled rRNAs (23S, 16S, 4.5S, and 5S; Ohyama et al., 1983). The 5'- and 3'-termini of 16S and 23S rDNA were determined by comparison with the 16S and 23S rDNA sequences of maize (Schwarz & Kössel, 1980; Edwards & Kössel, 1981) and tobacco (Tohdoh & Sugiura, 1982; Takaiwa & Sugiura, 1982b). The 16S rRNA (1496 bases) of liverwort chloroplasts shows high homology with that in E. coli (78.7%, Brosius et al., 1978), A. nidulans (84.8%, Tomioka & Sugiura, 1983), maize chloroplasts (94.2%, Schwarz & Kössel, 1980), and tobacco chloroplasts (95.5%, Tohdoh & Sugiura, 1982). The 3'- terminal region of the liverwort 16S rRNA gene also contains the sequence of CCTCCT that is complementary to prokaryotic SD sequences (Shine & Dalgarno, 1974). The existence of a sequence complementary to the SD sequence in 16S rRNA is a feature of prokaryote, and this may influence the translation efficiency of mRNA. The 23S rRNA (2811 bases) of liverwort chloroplasts also shows high homology with that in E. coli (67.9%, Brosius et al., 1980), A. nidulans (81.4%, Kumano et al., 1983), maize chloroplasts (92.5%, Edwards & Kössel, 1981), and tobacco chloroplasts (94.5%, Takaiwa & Sugiura, 1982a), although the 3'– terminal regions of the E. coli and A. nidulans 23S rRNAs include the coding sequence for the 4.5S rRNA found in land plant chloroplasts. The RNA sequences of 5S rRNA (Yamano et al., 1984) and 4.5S rRNA (Yamano et al., 1985) have been reported, and these are confirmed by the DNA sequence. Replacements, insertions, and deletions of nucleotides in the

(a) *trnV*(GAC) – *trnN*(GUU)

```
gaacctgtgaattcgccaattatgagttgggtgctttaaccattcagccatggatgctttttatatattatttattaataatagatattaattggATGATTATAGTTTATAACATTTTAT   81120
                                                                                                <TAATAT                  (121000)

TTCTTAAATTCAATCAATTATTTTTCTTTTTTTCTTCTATATGTATTTTTTTAACATAAAAATTTGACGGGTTAGCGTGAGCTTATCTGTGTGATTATACATTTTAGTGGAATTTTTTAT   81240
 <GAATTT                                                                                                                 (120880)

TTTTTAAAAATTTTTGCTTTCGTTCTGTGAGTTTATGTCTTTTTTTTCTATTTTTGAATATAGTGTTTTTTTTATTTACCTACGTATAAAAAACTAAAAATATAGGTATAAAACAAATAT   81360
                                                                                                                         (120760)

GTAAAAAATTGTCTATTTTAATGTTAAAAAAGGAAAAAACGTATATAACTTTTTGAAGTTAAGGTGTAGCTTTTTTTACAAACTGTTTGCATAAAAAAAATTGAATTTCAATTTATGATC   81480
                                                                                                                         (120640)

GTAGATTAGGAAAAAAAATAAGTATTTATTTAGAAGACAATTTTATCATGAGACATTTCAGAGGTCAACTAATTTTTTTTATTATAAAATATAAATTTTTCTTTATATTTTTTTGTCAAG   81600
                                                                                                                         (120520)

TAAACAAAAGTATATTAAATACTAAATAGTATTTTTTGTTTGGATTGGTAAGGTGGTAATATAATTATATGAAATAATTAATGGTTGCATAAGTTTCTTAAATTTTTGAGATTTAAGTTT   81720
                                      TTGTTT>                    TATAAT>                                                 (120400)

AAAAAATTTGAATGAAAATAGAAAAAATATATTATCTATAAATAATATATTTTTTAGAAGTCAGATTTTAAGTTCTCTTTTTTTTCTGAGAATAGGGATATAACTCAGCGGTAGAGTATC   81840
     TTGAAT>                TATTAT>                                                  Val-GAC>  5'-AGGGAUAUAACUCAGCGGUAGAGUAUC (120280)

ACCTTGACGTGGTGGAAGTCATCAGTTCGAACCTGATTATCCCTAAAAATAGTTAAAGCATTTTGATTTGTTGTTTTCTTGTTATTGAAAGAGGCTTGTGGGATTGACATAATAGGGTAG   81960
ACCUUGACGUGGUGGAAGUCAUCAGUUCGAACCUGAUUAUCCCUA-3'                                                  TTGACA>                (120160)

GTATGGGTATACTAGAAATGAGCTTCAAGCTAATATGAAGTGAATGAAAAATAAACATAAGTTATCTATCTCTTGGGATGGAAGACGATTTGAAATCTGCTTTGTTTACGAAAAAGGAAG   82080
      TATACT>                                                                                                            (120040)

CTATAAGTAAAAGTAATATAATTATGAATCTCATGGAGAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGCATGCTTAACACATGCAAGTCGTACGGAAGGATCCTAGTGGTGTTTCC    82200
              16S rRNA>   5'-UCUCAUGGAGAGUUUGAUCCUGGCUCAGGAUGAACGCUGGCGGCAUGCUUAACACAUGCAAGUCGUACGGAAGGAUCCUAGUGGUGUUUCC (119920)

AGTGGCGGACGGGTGAGTAACGCGTAAGAACCTGCCCTTGGGAGGGGGACAACAGCTGGAAACGGTTGCTAATACCCCATAGGCTGAGGAGCAAAAGGAGGAATCCGCCTAAGGAGGGGC   82320
AGUGGCGGACGGGUGAGUAACGCGUAAGAACCUGCCCUUGGGAGGGGGACAACAGCUGGAAACGGUUGCUAAUACCCCAUAGGCUGAGGAGCAAAAGGAGGAAUCCGCCUAAGGAGGGGC (119800)

TTGCGTCTGATTAGCTAGTTGGTGAGGTAATAGCTTACCAAGGCGACGATCAGTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGACTGAGACACGGCCCAGACTCTTACGGGAGGCA   82440
UUGCGUCUGAUUAGCUAGUUGGUGAGGUAAUAGCUUACCAAGGCGACGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUGGGACUGAGACACGGCCCAGACUCUUACGGGAGGCA (119680)

GCAGTGGGGAATTTTCCGCAATGGGCGAAACGTGACGGAGCAATGCCGCGTGGAGGTAGAAGGCTCACGGGTCGTAAACTCCTTTTCTCAGAGAAGATGCAATGACGGTATCTGAGGAAT   82560
GCAGUGGGGAAUUUUCCGCAAUGGGCGAAACGUGACGGAGCAAUGCCGCGUGGAGGUAGAAGGCUCACGGGUCGUAAACUCCUUUUCUCAGAGAAGAUGCAAUGACGGUAUCUGAGGAAU (119560)

AAGCATCGGCTAACTCTGTGCCAGCAGCCGCGGTAAGACAGAGGATGCAAGCGTTATCCGGAATGATTGGGCGTAAAGCGTCTGTAGGTGGCTTTTTAAGTCCGCCGTCAAATCCCAGGG   82680
AAGCAUCGGCUAACUCUGUGCCAGCAGCCGCGGUAAGACAGAGGAUGCAAGCGUUAUCCGGAAUGAUUGGGCGUAAAGCGUCUGUAGGUGGCUUUUUAAGUCCGCCGUCAAAUCCCAGGG (119440)

CTCAACCCTGGACAGGCGGTGGAAACTACCAAGCTGGAGTACGGTAGGGGCAGAGGGAATTTCCGGTGGAGCGGTGAAATGCGTAGAGATCGGAAAGAACACCAATGGCGAAAGCACTCT   82800
CUCAACCCUGGACAGGCGGUGGAAACUACCAAGCUGGAGUACGGUAGGGGCAGAGGGAAUUUCCGGUGGAGCGGUGAAAUGCGUAGAGAUCGGAAAGAACACCAAUGGCGAAAGCACUCU (119320)

TCTGGGCCGACACTGACACTGAGAGACGAAAGCTAGGGGAGCAAATGGGATTAGATACCCCAGTAGTCCTAGCCGTAAACGATGGATACTAAGCGCTGTGCTATCGACCCGTGCAGTGCT   82920
UCUGGGCCGACACUGACACUGAGAGACGAAAGCUAGGGGAGCAAAUGGGAUUAGAUACCCCAGUAGUCCUAGCCGUAAACGAUGGAUACUAAGCGCUGUGCUAUCGACCCGUGCAGUGCU (119200)

GTAGCTAACGCGTTAAGTATCCCGCCTGGGGAGTACGTTCGCAAGAATGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGA   83040
GUAGCUAACGCGUUAAGUAUCCCGCCUGGGGAGUACGUUCGCAAGAAUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAAGCGGUGGAGCAUGUGGUUUAAUUCGAUGCAACGCGAAGA (119080)

ACCTTACCAGGGCTTGACATGCCGTGAATCTTTTTGAAAGAAAAGAGTGCCTTCGGGAACGCGGACACAGGTGGTGCATGGCTGTCGTCAGCTCGTGCCGTAAGGTGTTGGGTTAAGTCC   83160
ACCUUACCAGGGCUUGACAUGCCGUGAAUCUUUUUGAAAGAAAAGAGUGCCUUCGGGAACGCGGACACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGCCGUAAGGUGUUGGGUUAAGUCC (118960)

CGCAACGAGCGCAACCCTCTTGTTTAGTTGCCATCATTAAGTTTGGAACCCTAAACAGACTGCCGGTGATAAGCCGGAGGAAGGTGAGGATGACGTCAAGTCAGCATGCCCCTTACGCCC   83280
CGCAACGAGCGCAACCCUCUUGUUUAGUUGCCAUCAUUAAGUUUGGAACCCUAAACAGACUGCCGGUGAUAAGCCGGAGGAAGGUGAGGAUGACGUCAAGUCAGCAUGCCCCUUACGCCC (118840)

TGGGCGACACACGTGCTACAATGGCCGGGACAAAGGGTCGCGACCTCGCGAGAGAAAGCTAACCTCAAAAACCCGGCCTCAGTTCGGATTGCAGGCTGCAACTCGCCTGCATGAAGCCGG   83400
UGGGCGACACACGUGCUACAAUGGCCGGGACAAAGGGUCGCGACCUCGCGAGAGAAAGCUAACCUCAAAAACCCGGCCUCAGUUCGGAUUGCAGGCUGCAACUCGCCUGCAUGAAGCCGG (118720)

AATCGCTAGTAATCGCCGGTCAGCCATACGGCGGTGAATCCGTTCCCGGGCCTTGTACACACCGCCCGTCACACTATGGGAGCTGGCCATGCCCGAAGTCGTTACTCTAACCGTAAGGAG   83520
AAUCGCUAGUAAUCGCCGGUCAGCCAUACGGCGGUGAAUCCGUUCCCGGGCCUUGUACACACCGCCCGUCACACUAUGGGAGCUGGCCAUGCCCGAAGUCGUUACUCUAACCGUAAGGAG (118600)

GGGGGTGCCGAACAGGGGCTAGTGACTGGAGTGAAGTCGTAACAAGGTAGCCGTACTGGAAGGTGCGGCTGGATCACCTCCTTTTTAGGGAGAGCTAATGCTTGTTGAACTTTTTCATTT   83640
GGGGGUGCCGAACAGGGGCUAGUGACUGGAGUGAAGUCGUAACAAGGUAGCCGUACUGGAAGGUGCGGCUGGAUCACCUCCUUU-3'                                  (118480)

AACGTTTTTTCGCAAAAAAAGTGAGTTATTTCATTTGAAAAAAAGTCATTTTTCACGTTTTTTTCTTGATACTTAAATAAAATTAAGTTCATAAGCTTATTATCCTAGGTCGGAACAAGT   83760
                                                                                                                         (118360)

TGATAAAAACCCATTAAATTATCCTTAGCATGGCAGTAACGTCATCAGGTAAATATGCAAATGGGATTGGTTTTTTTCGCCCTTGGTATTGCAGGTCTCCTAGGAGACCTGCACGACGGG   83880
                                                                                                          Ile-GAU>  5'-GGG (118240)

CTATTAGCTCAGTGGTAGAGCGCGCCCCTGATAATTGCGTCGTTGTGCTTGGGCTGTGAAGGTTTTCAGCCACATAAATAGTTCAATGTGCTCATCAGCGTCTGACCTGAAGATGTTAAT   84000
CUAUUAGCUCAGUGGUAGAGCGCGCCCCUGAUAAgugyg.................(intron)......................................................... (118120)

CATTTAAGGCACCTTAGCATGGCGTATTCCTTTTCTTTCAATTTGAAAGAGGATAGATGGGCGATTCAGGTAGATCAAATGGAGATTCAATTGCACTCACTCGTGGGATCTGGGCCGTCC   84120
.........................................................(intron)........................................................ (118000)

AGGGAGGACCCATTGAGCTCCTCTCTTCTCGAAAAAATCAATACATGCCTTATCAGTGTATGGATGACTATCTTTCGAGCACAGGTTGAAGTTCAACCTAGATGTGAAAAATGGAGCACC   84240
.........................................................(intron)........................................................ (117880)

TAATAACGCATCTTCACAGACCAAGAACTACGAGATCACCTATAGAGATTTTTATTCTAGGGTGACGGAGGGATCATATTATTCGAGCCTTTTTCTGCTTTTCTTGGAGGTTCGGAGAAA   84360
.........................................................(intron)........................................................ (117760)
```

Fig. 2, cont.

```
GCAGCAATCAATATTTTTTTTAGGTTAGTTTAGGATTAGAGAAGGATGTCAAATTGTTTAAAAAAGATCTTAGGTCCTAAAATATTAGATTCAGTCATAAAAATACTTGGTATAAGTAAC  84480
..........................................................(intron)......................................................  (117640)

GCTACGACTTTTTTAGTCATTACAGGCCGAGGGTCACAATGAATGTTTTTTTTCTCTATTCTCTAATGATGGATGCAGGTTCGAAAAAGGATCTTAGAGTGTTTAGTTGAGTTAAGAGAG  84600
..........................................................(intron)......................................................  (117520)

TGGATTTTCTAATGTTTTCTTTTCTCTTTTCATCAAAATTTTCTTCCAAAAACTTGTTAATGGCAAGAAAATAAATACACTTGGAGAGCGCAGTACAACGGAAAGTTGTATGCTGCGTTC  84720
..........................................................(intron)....................ragccg--augaa--gaaa—uucaugu-cgguuy  (117400)

GGGAAGGATGAATCGTTCCTAAAAAAAAAGAATTTATAGATTTTTTTTATTGAAATTGTAGGTGCGATGATTTACTTCACGGGCGAGGTCTCTGGTTCAAGTCCAGGATAGCCCAGCTGC  84840
..........................................................(intron)......cuayy-y-ayGGGCGAGGUCUCUGGUUCAAGTCCAGGAUAGCCCA-3'    (117280)
                                                                                                           +-    --    --- -----

GCCAAGATAAATAAAAAAAGCATAATGATTTATTTTTGCATGCTTTCCTTGCTTTCCTTGGCCTGGGTATAGGGGATATAGCTCAGTTGGTAGAGCGCCGCCCTTGCAATTGGGTCGTTG  84960
                                                               Ala-UGC>   5'-GGGGAUAUAGCUCAGUUGGUAGAGCGCCGCCCUUGCAAgugyg......  (117160)
------  --- --  ------>                     <------  --  ---  -------------  --+

CGCTTACGGGTTGGATGCTTAATTGTCTAGGCGGTAATGATAGTATCTCTTACCTAAACCGGTGGCTAACTTTTTCTTAGGAATGAGAAAGAGGACTGGAACATGCCACTGAAAAGTTTT  85080
..........................................................(intron)......................................................  (117040)

ACTAAGACAAAGATGAGTTGTTAAAAGTAAAAAAAAGGTAGGATGGGTAGTTGGTTAGATCTAATATGGATCGTACATGGACGATAGTTGGAGTCGGCGGCTCTCCTAGGGTTACCTCAT  85200
..........................................................(intron)......................................................  (116920)

CTAGAATCCCTGGGGAAGAGAATCAAGTTGGCCCTTGCGAACAGCTTGATGCACTATCTCTCTTTAACCCTTCAAGCCAAATGTGGCAAAAGGAACGAAAAGCCATGGACTGACCCCATC  85320
..........................................................(intron)......................................................  (116800)

GTTTCCACCCCGTAGGAACTACGAGATCGCCCCCAAGAATGTCGAATAAGGCATCAAGGGGTCACAGACCGACCATAAACTTAATTCAATAAGGCGAACGGATTAACCTCTTTTGTTCTT  85440
..........................................................(intron)......................................................  (116680)

ATTGGTAAGAAGGGTCGGAGAAGGGTAACAACTCAATATTAAGACTAATTAGTCAGGTGGAAAAAAAAAAGAATTTTAAATTCTTGTGTAGTTAGATATTTTCAAATTACAAAAGTTCCT  85560
..........................................................(intron)......................................................  (116560)

ATCATTCTTCAATTCGACGCCTTTTGAGTTAAGTAGCTCTTTGGAGAGCACAGTACGATGAAAGTTGTAAGCTGTGTTTGGGGGGGAGTTATTGTCTATCAAAGGCCTCTATGGTAAAAT  85680
............................................ragccg-augaa--gaaa-uucaugu-cgguuy..........................................  (116440)

AAATCAATAAAGTCTAAGAGACGATGGCTTACCCTGTGGCGGATGTCAGCGGTTCGAGTCCGCTTATCTCCAGTTGATGATCGGAATGAAGACAATATAGTTGCCTTGGATATAATAAAA  85800
..........................cuayy-y-ayGGCGGAUGUCAGCGGUUCGAGUCCGCUUAUCUCCA-3'                                                 (116320)

AAAATTTTAATCTTTATAACCAAGTTGACCTAATTTTTGATTATTTATGGACGTTGATAAGATCTTTTTTTTAGCACTTTATAATGGCATAGCCTTTAATTAATGGCGAGGTTCAAACGA  85920
                                                                                                    23S rRNA>   5'-UUCAAACGA  (116200)

AAAAGGGCTTACGGTGGATACCTAGGCACCCAGAGACGAGGAAGGGCGTAGCAAGCGACGAAATGCTTCGGGGAGCTGAAAATAAGTATAGATCCGGAGATTCCCGAATAGGTTAACCTT  86040
AAAAGGGCUUACGGUGGAUACCUAGGCACCCAGAGACGAGGAAGGGCGUAGCAAGCGACGAAAUGCUUCGGGGAGCUGAAAAUAAGUAUAGAUCCGGAGAUUCCCGAAUAGGUUAACCUU  (116080)

TGAAACTGCTGCTGAATTCATAGGCAGACAAGAGACAACCTGGCGAACTGAAACATCTTAGTAGCCAGAGGAAAAGAAAGCAAAAGCGATTCTCGTAGTAGCGGCGAGCGAAATGGGAAC  86160
UGAAACUGCUGCUGAAUUCAUAGGCAGACAAGAGACAACCUGGCGAACUGAAACAUCUUAGUAGCCAGAGGAAAAGAAAGCAAAAGCGAUUCUCGUAGUAGCGGCGAGCGAAAUGGGAAC  (115960)

AGCCTAAACCGTGAAAACGGGTTGTGGGGGAGCTAAATAAGTGTTGTGTTGCTAGGCGAAGCAGTTGAGTCCTGCACCCTAGATGGTGAAAGTCCAGTAACCGAAAGCAGCACTAGCTTA  86280
AGCCUAAACCGUGAAAACGGGUUGUGGGGGAGCUAAAUAAGUGUUGUGUUGCUAGGCGAAGCAGUUGAGUCCUGCACCCUAGAUGGUGAAAGUCCAGUAACCGAAAGCAGCACUAGCUUA  (115840)

GGCTCTAACCCGAGTAGCATGGGGCACGTGGAATCCCGTGTGAATCAGCAAGGACCACCTTGTAAGGCTAAATACTCCTGGGTGACCGATAGCGAAGTAGTACCGTGAGGGAAAGGTGAA  86400
GGCUCUAACCCGAGUAGCAUGGGGCACGUGGAAUCCCGUGUGAAUCAGCAAGGACCACCUUGUAAGGCUAAAUACUCCUGGGUGACCGAUAGCGAAGUAGUACCGUGAGGGAAAGGUGAA  (115720)

AAGAACCCCCATCGGGGAGTGAAATAGAACATGAAACCGTAAGCTCCCAAGCAGTGGGAGGAGAATTGAATCTCTGACCGCGTGCCTGTTGAAGAATGAGCCGGCGACTTATAGGCAGTG  86520
AAGAACCCCCAUCGGGGAGUGAAAUAGAACAUGAAACCGUAAGCUCCCAAGCAGUGGGAGGAGAAUUGAAUCUCUGACCGCGUGCCUGUUGAAGAAUGAGCCGGCGACUUAUAGGCAGUG  (115600)

GCCTGGTTAAGGGAGCCCACCGGAGCCGTAGCGAAAGCGAGTCTTCTTAGGGCAATTGTCACTGCTTATGGACCCGAACCTGGGTGATCTATCCATGACCAGGATGAAGCTTGGGTGAAA  86640
GCCUGGUUAAGGGAGCCCACCGGAGCCGUAGCGAAAGCGAGUCUUCUUAGGGCAAUUGUCACUGCUUAUGGACCCGAACCUGGGUGAUCUAUCCAUGACCAGGAUGAAGCUUGGGUGAAA  (115480)

CTAAGTGGAGGTCCGAACCGACTGATGTTGAAAAATCAGCGGATGAGTTGTGGTTAGGGGTGAAATGCCACTCGAACCCAGAGCTAGCTGGTTCTCCCCGAAATGCGTTGAGGCGCAGCA  86760
CUAAGUGGAGGUCCGAACCGACUGAUGUUGAAAAAUCAGCGGAUGAGUUGUGGUUAGGGGUGAAAUGCCACUCGAACCCAGAGCUAGCUGGUUCUCCCCGAAAUGCGUUGAGGCGCAGCA  (115360)

GTTGACTGGACTATCTAGGGGTAAAGCACTGTTTCGGTGCGGGCTGCGAGAGCGGTACCAAATCGAGGCAAACTCTGAATACTAGGTAGGACTTCCTATTAATAGGAAGTAAGGGTCAGC  86880
GUUGACUGGACUAUCUAGGGGUAAAGCACUGUUUCGGUGCGGGCUGCGAGAGCGGUACCAAAUCGAGGCAAACUCUGAAUACUAGGUAGGACUUCCUAUUAAUAGGAAGUAAGGGUCAGC  (115240)

CAGTGAGACAGTGGGGGATAAGCTTCATTGTCGAGAGGGGAACAGCCCAGATCACCAGCTAAGGCCCCTAAATGACCGCTCAGTGGTAAAGGAGGTAGGAGTGCAAAGACAGCCAGGAGG  87000
CAGUGAGACAGUGGGGGAUAAGCUUCAUUGUCGAGAGGGGAACAGCCCAGAUCACCAGCUAAGGCCCCUAAAUGACCGCUCAGUGGUAAAGGAGGUAGGAGUGCAAAGACAGCCAGGAGG  (115120)

TTTGCCTAGAAGCAGCCACCCTTGAAAGAGTGCGTAATAGCTCACTGATCAAGCGCTCTTGCGCCGAAGATGAATGGGACTAAGCGGTCTGCCGAAGCTGTGGGATGTCAAAATACATCG  87120
UUUGCCUAGAAGCAGCCACCCUUGAAAGAGUGCGUAAUAGCUCACUGAUCAAGCGCUCUUGCGCCGAAGAUGAAUGGGACUAAGCGGUCUGCCGAAGCUGUGGGAUGUCAAAAUACAUCG  (115000)

GTAGGGGAGCGTTCCGCCTTAGGGAGAAGCATCACGTGAGCAGGTGTGGACGAAGCGGAAGCGAGAATGTCGGCTTGAGTAACGCAAACATTGGTGAGAATCCAATGCCCCGAAAACCTA  87240
GUAGGGGAGCGUUCCGCCUUAGGGAGAAGCAUCACGUGAGCAGGUGUGGACGAAGCGGAAGCGAGAAUGUCGGCUUGAGUAACGCAAACAUUGGUGAGAAUCCAAUGCCCCGAAAACCUA  (114880)

AGGGTTCCTCCGCAAGGTTCGTCCACGGAGGGTGAGTCAGGGCCTAAGATCAGGCCGAAAGGCGTAGTCGATGGACAACAGGCAAATATTCCTGTACTACCCCTTGTTGGTCCCGAGGGA  87360
AGGGUUCCUCCGCAAGGUUCGUCCACGGAGGGUGAGUCAGGGCCUAAGAUCAGGCCGAAAGGCGUAGUCGAUGGACAACAGGCAAAUAUUCCUGUACUACCCCUUGUUGGUCCCGAGGGA  (114760)

CGAGGAGGCTAGGTTAGCCGAAAGATGGTTATCGGTTCAAGGATGCAAGGTGAATTCCCTTGAAATTTTCAAGGGTAAAAAAGAGGTAGTGAAAATGCTTCCAGCCAATGTCCGAGTACC  87480
CGAGGAGGCUAGGUUAGCCGAAAGAUGGUUAUCGGUUCAAGGAUGCAAGGUGAAUUCCCUUGAAAUUUUCAAGGGUAAAAAAGAGGUAGUGAAAAUGCUUCCAGCCAAUGUCCGAGUACC  (114640)

AAGCACTACGGTCTGAAGTAATTAATGCCACACTCCCAAGAAAAGCTCGAACGACCTTAAACAAGTGGGTACCTGTACCCGAAACCGACACAGGTAGGTAGGTAGAGAATACCTAGGGGC  87600
AAGCACUACGGUCUGAAGUAAUUAAUGCCACACUCCCAAGAAAAGCUCGAACGACCUUAAACAAGUGGGUACCUGUACCCGAAACCGACACAGGUAGGUAGGUAGAGAAUACCUAGGGGC  (114520)

GCGAGATAACTCTCTCTAAGGAACTCGGCAAAATAGCCCCGTAACTTCGGGAGAAGGGGTGCCTCCTCTAAAAGGAGGTCGCAGTGACCAGGCCCAGGCGACTGTTTACCAAAAACACAG  87720
GCGAGAUAACUCUCUCUAAGGAACUCGGCAAAAUAGCCCCGUAACUUCGGGAGAAGGGGUGCCUCCUCUAAAAGGAGGUCGCAGUGACCAGGCCCAGGCGACUGUUUACCAAAAACACAG  (114400)

GTCTCCGCAAAGTCGTAAGACCATGTATGGGGCTGACGCCTGCCCAGTGCCGGAAGGTTAAGGAAGTTGGTGACCTGATGACAGGGAAGCCAGCGACTGAAGCCCCGGTAAACGGCGGCC  87840
GUCUCCGCAAAGUCGUAAGACCAUGUAUGGGGCUGACGCCUGCCCAGUGCCGGAAGGUUAAGGAAGUUGGUGACCUGAUGACAGGGAAGCCAGCGACUGAAGCCCCGGUAAACGGCGGCC  (114280)
```

Fig. 2, cont.

```
GTAACTATAACGGTCCTAAGGTAGCGAAATTCCTTGTCGGGTAAGTTCCGACCCGCACGAAAGGCGTAACGATCTGGGCACTGTCTCGGAGAGAGACTCGGTGAAATAGACATGTCTGTG   87960
GUAACUAUAACGGUCCUAAGGUAGCGAAAUUCCUUGUCGGGUAAGUUCCGACCCGCACGAAAGGCGUAACGAUCUGGGCACUGUCUCGGAGAGAGACUCGGUGAAAUAGACAUGUCUGUG (114160)

AAGATGCGGACTACCTGCACCTGGACAGAAAGACCCTATGAAGCTTTACTGTTCCCTGGGATTGGCTTTGGGTTTTTCTTGCGCAGCTTAGGTGGAAGGCAAAGAAGGCCCCTTCTGGGC   88080
AAGAUGCGGACUACCUGCACCUGGACAGAAAGACCCUAUGAAGCUUUACUGUUCCCUGGGAUUGGCUUUGGGUUUUUCUUGCGCAGCUUAGGUGGAAGGCAAAGAAGGCCCCUUCUGGGC (114040)

GGTGGGAGCATCAGTGAAATACCACTCTAGAAGAGCTAGAATTCTAACCTTGTGTCAAAATTTACGGGCCAAGGGACATTCTCAGGTAGACAGTTTCTATGGGGCGTAGGCCTCCCAAAA   88200
GGUGGGAGCAUCAGUGAAAUACCACUCUAGAAGAGCUAGAAUUCUAACCUUGUGUCAAAAUUUACGGGCCAAGGGACAUUCUCAGGUAGACAGUUUCUAUGGGGCGUAGGCCUCCCAAAA (113920)

GGTAACGGAGGTGTGCAAAGGTTTCCTCAGGCTGGACGGAAATCAGCCTTCGAGTGCAAAGGCAGAAGGGAGCTTGACTGCAAGACATACCCGTCGAGCAGGGACGAAAGTCGGCCTTAG   88320
GGUAACGGAGGUGUGCAAAGGUUUCCUCAGGCUGGACGGAAAUCAGCCUUCGAGUGCAAAGGCAGAAGGGAGCUUGACUGCAAGACAUACCCGUCGAGCAGGGACGAAAGUCGGCCUUAG (113800)

TGATCCGACGGTACCAAGTGGAAGGGCCGTCGCTCAACGGATAAAAGTTACTCTAGGGATAACAGGCTGATCTTCCCCAAGAGTTCACATCGACGGGAAGGTTTGGCACCTCGATGTCGG   88440
UGAUCCGACGGUACCAAGUGGAAGGGCCGUCGCUCAACGGAUAAAAGUUACUCUAGGGAUAACAGGCUGAUCUUCCCCAAGAGUUCACAUCGACGGGAAGGUUUGGCACCUCGAUGUCGG (113680)

CTCTTCGCCACCTGGGGCGGTAGTACGTTCCAAGGGTTGGGCTGTTCGCCCATTAAAGCGGTACGTGAGCTGGGTTCAGAACGTCGTGAGACAGTTCGGTCCATATCCGGTGTGGGCGTT   88560
CUCUUCGCCACCUGGGGCGGUAGUACGUUCCAAGGGUUGGGCUGUUCGCCCAUUAAAGCGGUACGUGAGCUGGGUUCAGAACGUCGUGAGACAGUUCGGUCCAUAUCCGGUGUGGGCGUU (113560)

AGAGCATTGAGAGGACCTTTCCCTAGTACGAGAGGACCGGGAAGGACGCACCTCTGGGTTACCAGTTATCGTGCCCACGGTAAACGCTGGGTAGCCAAGTGCGGACGGATAACTGCTGAA   88680
AGAGCAUUGAGAGGACCUUUCCCUAGUACGAGAGGACCGGGAAGGACGCACCUCUGGGUUACCAGUUAUCGUGCCCACGGUAAACGCUGGGUAGCCAAGUGCGGACGGAUAACUGCUGAA (113440)

AGCATCTAAGTAGGAAGCCCACCTCAAGATGAGTGCTCTCCTATTCTTCTTCTCTTGAAGCAGTCTTTGGGTAATAAACATACTCAAGACACTGATAGATTTTCTGTCGTTGCAAGAAAT   88800
AGCAUCUAAGUAGGAAGCCCACCUCAAGAUGAGUGCUCUCCU-3'                                                    ←——— —— ———→           (113320)

GAAACGACAAAAGTCTTGAGAATCCAAGATAAGGTCACGGCAAGACTAGCCGTTTATTTTTACGATAGGTGCCAAGTGGAAGTGCAGTAATGTATGTAGCTGAGGCATCCTAACAGACCG   88920
               4.5S rRNA>   5'-UAAGGUCACGGCAAGACUAGCCGUUUAUUUUUACGAUAGGUGCCAAGUGGAAGUGCAGUAAUGUAUGUAGCUGAGGCAUCCUAACAGACCG (113200)
  <————————      —— —→

AGAGATTTGAACCTTGTTCCGCCATGACCTGATAAAAGTAATCAGGTATAGCCACCAACTTTCATTGTTCAATTGTTTGACAACATAAACCTAACAACTTTACCCTGCTCTTATTTTGGG   89040
AGAGAUUUGAAC-3'                                                                                                          (113080)
      ←—— ——    —— —— ———>    <——— — — —   ——→

CAGGGTTTCAAAGGGGTTTTTTTCCTGGAAGGGACACTTCTAGTGCCCTTTCCAGAATGAAAGACTCACAATTACTTGGTTTTTTTTTTATTATACTTTTCTTTGTTCATGGGTTGATATT   89160
             ←——   ————————— ——>  <———— ——   —— ——→                                                   5S rRNA>   5'-UU (112960)

CTGGTGTCTTAGGCGTAGAGGAACCACACCAATCCATCCCGAACTTGGTGGTGAAACTCTATTGCGGTGACAATACTTTAGGGGAAGCCCTATGGAAAAATAGCTCGACGCCAGGATGAA   89280
CUGGUGUCUUAGGCGUAGAGGAACCACACCAAUCCAUCCCGAACUUGGUGGUGAAACUCUAUUGCGGUGACAAUACUUUAGGGGAAGCCCUAUGGAAAAAUAGCUCGACGCCAGGAU-3' (112840)

AAAATTAATGTCTCCTATTATTAGTTCAAAATACCATACATACCAATTTTGACCTCCTTTATTTCCTACTCCACACTTCAAAATGCATATATTTTTTTTTGAATAACAATTCTTAAATTT   89400
                   ←———>             <———→                                                                  ←— ————— (112720)

CCGCGCATCTTCTTAGTCTTGAATGGCTAAAGAGAAAAGATTGCTTTTGGAAAAGGCTTCTAGAACAGATTAGTGGAGGCGGGGTTTGTAGCTCAGAGGATTAGAGCACGTGGCTACGAA   89520
——— ————————>      <——— —— ——— ——    ———    —— ———→ Arg-ACG>   5'-GGGUUUGUAGCUCAGAGGAUUAGAGCACGUGGCUACGAA (112600)

CCACGGTGTCGGGGGTTCGAATCCCTCCTTGCCCACAACAACCTTCAGAGGTTTTTTACATGGTTAGGAGGTTCCAACGATTATTGGAAGACCCAACGGCGGGACTTATGGTATTTTTTT   89640
CCACGGUGUCGGGGGUUCGAAUCCCUCCUUGCCCA-3'                                ←——— ——  ——>   <——— ————→                          (112480)

TAAGCAGGTCTTTTACTCAAAATTACTAAAAATAATACATTACCTACTCTTTATGTATAGTACACTTAATATTAATCAAACAACTTTTTGTTTTCCCTCTTGCAACTTTGATTTACCACT   89760
                                                                                       ←——>   <——→                         (112360)

GTCAGGATTGAGCAAAGTTTTAGTAATAATAAACTTCGCATAATTAAGTAGGTTTGTTTAGATAAGGCAATGAAATTGTGGTAATAATATTTTACTAAATTTTATGACTGCATTCTTGAT   89880
                                                                                                                 ←— —— —— (112240)

AAATTGCTGGTAAATTATTTGAGTAAGTTATCTATTAGTTAGTTGAAAAGAACTAATTAGAACATTCAATTTATAGCCAAATTTATAGTGGTACATCCGAGTAATTCTATTATCAATGCT   90000
——————  ————   ——— ——  ————>     <——— —————————— ——  ———————→                                                         (112120)

TTTTAAGCAAAAAACTTACATTGTTACTAGTTAGTCACAAGTCTCAAACAAATAGAAGCCTTCACTCAAATAAGTCTAATACTTTCATCAGAAAAAAAATAATCCATTTGCTTTTCTTAG   90120
                                                                                                                         (112000)

TTTTTCAGTACTCCACATAGATCATTGTTTCCATTTTTTTAGATTACGGATAATCATCAGTACATTTTTTTTTTGTTGCAATTTGACATTAGTATAAACAATGAAAAAGAACACCTAGTT   90240
                                                                            ←—— ——   ——>        <——— ——→                 (111880)

CTAAAAGTCAAAACAAGCATCTCCTCAGGTAGGATTTGAACCTACGACCAATCGGTTAACAGCCGACCGCTCTACCACTGAGCTACTAAGGAACAATGAGTTTAATTCTAAAAACATTCA   90360
                 3'-GAGGAGUCCAUCCUAAACUUGGAUGCUGGUUAGCCAAUUGUCGGCUGGCGAGAUGGUGACUCGAUGAUUCCU-5'   <Asn-GUU                 (111760)

AAAACTTTTCAACCTAAAATTAGCCCATAAACTGTTCAAAGAACCAAAAAATTCTTGGATTAAGAATGGAAATAACTTTCAGTACACTCTACCTTCTTTTATTATAGGGTAAAAAGATAA   90480
                                                                                                       <TAATAT           (111640)

CGATAGCAATCCCCTAAACTCTACATCGAAAAATTTTTAGACAAGGGGGAGGCGGTCAACCATCACTATGATCTTCTCCAGTGTCCTCCCCGAGATGCTTATTGATTAAGCAAGTTCAAT   90600
 <ATCGTT                    ←— ——><——— —→    ←———>                                   <———→          TTGATT>                (111520)

GATGCTACAATCTTAACGATTTGCTAAGTCAACTCATTCTCCCGAAGGGCTACAAATAACTCTTCTACAAAAAAAAGTATTCTTTTATCCTAAAAGTGAGCAGATCATGGTGAAACGGTT   90720
    TACAAT>                                                                              ←——>   <——→          ←——— —— —>  (111400)

TTAACCGACTTTACCTATTTTTCCGATTCCTTTCTTTTAATAAAACAAAGCAGATTTTATAGCACTTGGAAATTATTTTCAATCACAAAATCTCTTTCAAAATCCCTTTTCCTGTTTGTC   90840
   <—— ———— →              ←—— ———> <——— ——→                                                                          (111280)

CTTTGACCTCTTTGCTTACTTCATGTTTAGAGTCTTTGGCTTATAGACTAAATGTTAGAGTACGTAACATTCTATCTGATCTACTCCGGTTTTTTGTTAATCAAGTAGTAAGGTTGTGGA   90960
                     TTCATG>                  TAGACT>                                                                    (111160)

AAGTAGCGAAGTCAGAAAAACTCCATATTCACGATTGTATCCTTATTCTTGAAAGAATTTAAAAGAATTTTAAATTTACTTAATAAATTAATTCAAAAGCTCATAATAACATAAAATATG   91080
                                                       ←————>  <————→                                ←—————>   <—— (111040)

TTATCATAAATAAATATTATTTAAATAATATAGCTATATAAAAAAAACAAAAACATACAAAAAATTATGAAAAtaaataagaagaaattctacctccttctatatattttaaactctca   91200
——— →     ←————><————→
```

Fig. 2, cont.

(b) *ndh5*

```
AAACATATCTTTTTAGTTCAAAACTAAAAAAAGATATATAATATTATATTACTTTATATTATAAGTATCTTTTTTAGTTTTTTTTTATTACTTAAAAAGAAAAAATATTAATTATTTTTATA   93241

TTTTTGAATTATAAACGAAAATTTATAAAAATATAAAGTTTTATTTTAATAAAAATGTTTTATGGAACTTATATTTCAAAATGTTTGGTTTGTACCATTGTTTCCATTTTTAGCTTCTAT   93121
      TTATAA>                 TATAAA>                 ndh5>  M  E  L  I  F  Q  N  V  W  F  V  P  L  F  P  F  L  A  S  I

TTTATTAGGAATCGGATTATTTTTTTTCCCAAATTCTATAAAAAAATTTCGTCGTCTATCTTCTTTTATTAGTATTATGTTTTTAAACATAGCTATGTTACTCTCATTTCATTTTTTTTG   93001
 L  L  G  I  G  L  F  F  F  P  N  S  I  K  K  F  R  R  L  S  S  F  I  S  I  M  F  L  N  I  A  M  L  L  S  F  H  F  F  W

GCAACAAATTACAGGTAGTCCAATTCATAGATATTTATGGTCTTGGGTTCTTTATAAAAATTTTGTTTTAGAAATAGGCTATTTACTTGATCCACTTACTTCAATTATGTTAGTTTTAGT   92881
 Q  Q  I  T  G  S  P  I  H  R  Y  L  W  S  W  V  L  Y  K  N  F  V  L  E  I  G  Y  L  L  D  P  L  T  S  I  M  L  V  L  V

AACTACAGTAGCAGTTATGGTTATGATTTATAGTGATAGTTATATGTTTTATGATGAAGGATATATAAAATTTTTTTGTTATTTAAGTCTTTTTACTGCATCAATGTTAGGGTTAGTTCT   92761
 T  T  V  A  V  M  V  M  I  Y  S  D  S  Y  M  F  Y  D  E  G  Y  I  K  F  F  C  Y  L  S  L  F  T  A  S  M  L  G  L  V  L

TAGTCCTAATTTAATACAAGTTTATATTTTTTGGGAATTAGTTGGAATGTGTTCATATTTATTAATTGGTTTTTGGTTTACTAGACCAAGTGCAGCTAATGCGTGTCAAAAAGCTTTTGT   92641
 S  P  N  L  I  Q  V  Y  I  F  W  E  L  V  G  M  C  S  Y  L  L  I  G  F  W  F  T  R  P  S  A  A  N  A  C  Q  K  A  F  V

TACAAATCGCATTGGTGATTTTGGATTATTATTAGGCATTTTAGGATTTTATTGGATAACAGGTAGTTTTGATTTTCAACAATTATCAAAACGATTTTTTGAATTACTAAGCTATAATCA   92521
 T  N  R  I  G  D  F  G  L  L  L  G  I  L  G  F  Y  W  I  T  G  S  F  D  F  Q  Q  L  S  K  R  F  F  E  L  L  S  Y  N  Q

AATTAATTTAGTTTTTGCTACTTTGTGTGCTCTATTTTTGTTTTTAGGTCCAGTAGCTAAATCTGCACAATTTCCATTACATATATGGTTACCAGATGCTATGGAAGGACCTACACCCAT   92401
 I  N  L  V  F  A  T  L  C  A  L  F  L  F  L  G  P  V  A  K  S  A  Q  F  P  L  H  I  W  L  P  D  A  M  E  G  P  T  P  I

TTCAGCCCTTATTCATGCTGCAACTATGGTTGCAGCTGGTATTTTTCTAGTTGCTCGAATGTTTCCTCTTTTTCAAATGTTACCATTTGTCATGAGTATCATTTCTTGGACAGGTGCCAT   92281
 S  A  L  I  H  A  A  T  M  V  A  A  G  I  F  L  V  A  R  M  F  P  L  F  Q  M  L  P  F  V  M  S  I  I  S  W  T  G  A  I

TACAGCTTTATTAGGAGCTACTATTGCTTTAGCTCAAAAAGATCTTAAAAAAGGTTTAGCTTATTCAACAATGTCACAATTAGGATATATGATGTTAGCATTAGGCATCGGATCTTACAA   92161
 T  A  L  L  G  A  T  I  A  L  A  Q  K  D  L  K  K  G  L  A  Y  S  T  M  S  Q  L  G  Y  M  M  L  A  L  G  I  G  S  Y  K

AGCTGGTTTATTTCATCTTATTACACATGCTTATTCAAAAGCTTTACTATTTCTTGGTTCTGGTTCAGTTATTCATTCAATGGAACCTATTGTAGGTTATCATCCGAATAAAAGTCAAAA   92041
 A  G  L  F  H  L  I  T  H  A  Y  S  K  A  L  L  F  L  G  S  G  S  V  I  H  S  M  E  P  I  V  G  Y  H  P  N  K  S  Q  N

TATGATTTTTATGGGTGGTTTAAGACAATATATGCCAATAACTGCAATAACTTTTTTGTTTGGTACACTTTCTTTATGTGGAATTCCACCTTTTGCTTGTTTTTGGTCCAAAGATGAAAT   91921
 M  I  F  M  G  G  L  R  Q  Y  M  P  I  T  A  I  T  F  L  F  G  T  L  S  L  C  G  I  P  P  F  A  C  F  W  S  K  D  E  I

TTTAGTAAATAGTTGGTTACATTTTCCTATTTTAGGGTCTATTGCTTTTTTTACAGCTGGTTTAACTGCTTTTTATATGTTTCGTATATATTTTTTAACTTTTGAGGGAGATTTTCGTGG   91801
 L  V  N  S  W  L  H  F  P  I  L  G  S  I  A  F  F  T  A  G  L  T  A  F  Y  M  F  R  I  Y  F  L  T  F  E  G  D  F  R  G

TCATTTTTTTGATGACGTAAAAAAATTATCTTCTATTTCAATATGGGGAAGTTTAGAATTTAACAAAGAACAATTTAAACTAGACAAAAAATCTACATTATATCCTAAAGAAGCTAATAA   91681
 H  F  F  D  D  V  K  K  L  S  S  I  S  I  W  G  S  L  E  F  N  K  E  Q  F  K  L  D  K  K  S  T  L  Y  P  K  E  A  N  N

TATAATGTTATTTCCTTTAATAATATTAACAATACCTACTGTATTTATAGGTTTTATAGGAATTTTATTTGATGAAAATAAAATGAATGTTGATTCTTTATCCTATTGGCTTACTTTATC   91561
 I  M  L  F  P  L  I  I  L  T  I  P  T  V  F  I  G  F  I  G  I  L  F  D  E  N  K  M  N  V  D  S  L  S  Y  W  L  T  L  S

CATAAATTCTTTTAATTACAGTAATTCTGAAAAGTTTTTAGAATTTTTATTTAATGCAATTCCTTCTGTTAGTATAGCTTTTTTTGGAATATTAATTGCTTTTTATTTATATGGTCCTAA   91441
 I  N  S  F  N  Y  S  N  S  E  K  F  L  E  F  L  F  N  A  I  P  S  V  S  I  A  F  F  G  I  L  I  A  F  Y  L  Y  G  P  N

TTTTTCTTTTTTAAAAAAAGAAAAAAAAAAATTACAATTGAAATCTGAAATAGATATTGTTTTAAAAAGTTTTTCAAATTTTATTTATAATTGGTCTTATTATCGAGCTTATATAGATGG   91321
 F  S  F  L  K  K  E  K  K  K  L  Q  L  K  S  E  I  D  I  V  L  K  S  F  S  N  F  I  Y  N  W  S  Y  Y  R  A  Y  I  D  G

GTTTTATTCTTCTTTTTTTATTAAAGGTTTAAGGTTTTTAATTAAAATAGTTTCTTTTATTGATCGATGGATTATTGATGGAATTATAAATGGAATTGGCATTTTTAGTTTTTTTGGAGG   91201
 F  Y  S  S  F  F  I  K  G  L  R  F  L  I  K  I  V  S  F  I  D  R  W  I  I  D  G  I  I  N  G  I  G  I  F  S  F  F  G  G

TGAGAGTTTAAAATATATAGAAGGAGGTAGAATTTCTTCTTATTTATttttcataatttttgtatgtttttgtttttttatatagctatattatttaaataatatttatttatgataa   91081
 E  S  L  K  Y  I  E  G  G  R  I  S  S  Y  L  F  F  I  I  F  C  M  F  L  F  F  L  Y  S  Y  I  I  ▬
```

(c) *rpl21* – ORF320

```
TATAAAAATAATTAATATTTTTTCTTTTTAAGTAATAAAAAAAACTAAAAAAGATACTTATAATATAAAGTAATATAATATTATATATCTTTTTTTAGTTTTGAACTAAAAGATATGTTT   93360

TTTTTTATAAAAAAACTAATTGAAAAAATATTTTGTTCTATTCTAGTTTAAAAATTTAATTAAATTTTTATTTAAAACAATTAATAAAATTAAAACATTTATAACATAATGAGTAAATAC   93480
                      TTGAAA>               TCTAGT>                                                            rpl21>  M  S  K  Y

GCAATAATTGAAACCGGAGGGCAGCAACTCCGAGTAGAACCTGGAAGATTTTATAATATTCGTCATTTTGTCTCATTAACACCAAATGAATTAGAACAAAACACAAAAATATTAATTTAT   93600
 A  I  I  E  T  G  G  Q  Q  L  R  V  E  P  G  R  F  Y  N  I  R  H  F  V  S  L  T  P  N  E  L  E  Q  N  T  K  I  L  I  Y

CGAGTATTAATGATTCGTCAAGAGTCTACTATAAAAATGGGACATCCTTGGTTAAAAGGAGCGATAGTTAAAGGTAGAATTTTACATTCTTGTCTTGAAAAAAAAATTACAATTTATAAA   93720
 R  V  L  M  I  R  Q  E  S  T  I  K  M  G  H  P  W  L  K  G  A  I  V  K  G  R  I  L  H  S  C  L  E  K  K  I  T  I  Y  K

ATGATTTCAAAAAAAAAAACACGACGTAAATTAGGACATCGACAAAAATCAACTCGATTTATAGTTGATTCTATTTTTTTAAATGGAAAAGAAATTTAATTATAAAAAAATATATAATAT   93840
 M  I  S  K  K  K  T  R  R  K  L  G  H  R  Q  K  S  T  R  F  I  V  D  S  I  F  L  N  G  K  E  I  ▬

TTTTTTCAGCAATTTTTATAAATAAAAGGTAAGGTATTTTTTATTATGGCAGTTCCAAAAAAACGTACATCTAAATCTAAAACACGAATTCGTAAAGCTATTTGGAAAAATAAAGCTAAT   93960
                         ORF69>  AGG             M  A  V  P  K  K  R  T  S  K  S  K  T  R  I  R  K  A  I  W  K  N  K  A  N

AAAAGCGCTTTAAGAGCTTTTTCTTTAGCAAAATCTATTTTAACAAATCGTTCAAAAAGTTTTTATTATACAATAAATGATAAATTATTAAATTCATCTAAATCCATATCAACGTCTAAA   94080
 K  S  A  L  R  A  F  S  L  A  K  S  I  L  T  N  R  S  K  S  F  Y  Y  T  I  N  D  K  L  L  N  S  S  K  S  I  S  T  S  K

TTAGATGAATCATAAAAAAAAATGTATTTTGTCAATTTTTTGTTTTTATAAATAAAATAAGAAAGTTAATAAGTTTAACTACATTTTTTTTAGGTTATTAAAAAAATGATTCCACTTTTTTTT   94200
 L  D  E  S  ▬                                                                        mbpY>  AGG          M  I  P  L  F  F

ATTCCTCCTTTTATAATACTTTTCATTACTAAAGGAAAATTTCGATTTTTAACTAAATTTGAATTAGTCTTAGCTTGTGCATTGCATTATGGTACTTTTATCTTAGCTTTGCCGATTTTT   94320
 I  P  P  F  I  I  L  F  I  T  K  G  K  F  R  F  L  T  K  F  E  L  V  L  A  C  A  L  H  Y  G  T  F  I  L  A  L  P  I  F

TTTTTGTTATATAAAACTAAGCAACAACCTTGGAATATTTTATTACAAACAGCTCTTGAACCAGTTGTGTTATCTGCTTATGGTTTTACTTTTTTAACTGCTTTATTGGCTACAATAATT   94440
 F  L  L  Y  K  T  K  Q  Q  P  W  N  I  L  L  Q  T  A  L  E  P  V  V  L  S  A  Y  G  F  T  F  L  T  A  L  L  A  T  I  I
```

Fig. 2, cont.

```
AACGCAATCTTTGGCCTAATTCTTGCTTGGGTTTTGGTAAGATATGAATTTCCAGGAAAAAAACTTTTAGATGCTACAGTAGATCTTCCATTTGCTCTTCCAACTTCAGTTGGAGGATTA   94560
N  A  I  F  G  L  I  L  A  W  V  L  V  R  Y  E  F  P  G  K  K  L  L  D  A  T  V  D  L  P  F  A  L  P  T  S  V  G  G  L

ACTTTAATGACTGTATTTAATGATAAAGGATGGATAAAACCTATTTGTTCATGGTTAAATATAAAAATAGTTTTTAATCCTATAGGAGTGCTTTTAGCAATGATTTTTGTATCTTTACCT   94680
T  L  M  T  V  F  N  D  K  G  W  I  K  P  I  C  S  W  L  N  I  K  I  V  F  N  P  I  G  V  L  L  A  M  I  F  V  S  L  P

TTTGTAGTACGCACCATACAACCCGTTTTACAAAACATGGAAGAAGATTTAGAAGAAGCTGCATGGTGTTTAGGTGCATCACCATGGACAACTTTTTGGCATATTTTGTTTCCACCATTA   94800
F  V  V  R  T  I  Q  P  V  L  Q  N  M  E  E  D  L  E  E  A  A  W  C  L  G  A  S  P  W  T  T  F  W  H  I  L  F  P  P  L

ACTCCATCATTATTAACTGGAACTACTTTAGGTTTTTCTAGAGCTTTAGGTGAATATGGTTCAATAGTTTTAATAGCGTCTAATATTCCAATGAAAGATTTAGTAATTTCTGTACTTCTT   94920
T  P  S  L  L  T  G  T  T  L  G  F  S  R  A  L  G  E  Y  G  S  I  V  L  I  A  S  N  I  P  M  K  D  L  V  I  S  V  L  L

TTTCAAAAACTTGAACAATATGATTATAAAAGTGCTACTATTATTGCAAGTTTTGTTTTAATAATTTCATTTACTGCACTTTTTTTTATTAATAAAATTCAGTTATGGAAAAAAACTTTT   95040
F  Q  K  L  E  Q  Y  D  Y  K  S  A  T  I  I  A  S  F  V  L  I  I  S  F  T  A  L  F  F  I  N  K  I  Q  L  W  K  K  T  F

CATAAATAAATCTTTGAAAGAATTAAAATGGAAAAAGTGTATTTCATAAAAAACTATGAAATACACTTTTTCCATTTTAATTAGTAAATTAATTATAATTTTTTTTTGAGTAGCGGGATT   95160
H  K  ——            ←-------------------------→      ←--------------------------→                    3'-AAACUCAUCGCCCUAA

TGAACCCACGACTTTCATCACCCCAAGATGATACACTACCAAACTATACTCCGTATGACTTTTTTATAAAAAAGTTAAAATTTTTAAAAAATTATAGTATATAATTTAAATAAGCCGCTA   95280
ACUUGGGUGCUGAAAGUAGUGGGGUUCUACUAUGUGAUGGUUUGAUAUGAGGC-5'<Pro-GGG         TTAAAA>   Leu-UAG>       TATAAT>     5'-GCCGCUA

TGGTGAAATTGGTAGACACGCTGCTCTTAGGAAGCAGTGCTAAGGCTTCTCGGTTCGAATCCGAGTAGCGGCATAAATTTTTTTATTATAAAAATATGTTTTATAATATCTTTTGTCTTT   95400
UGGUGAAAUUGGUAGACACGCUGCUCUUAGGAAGCAGUGCUAAGGCUUCUCGGUUCGAAUCCGAGUAGCGGCA-3'      ←---------→       ←-------→

TTATTTATTTATAATCAGTATTATATATATTTTTTTTTAATTCAATTTGTATACAATCTAAAAAACTTTAAATTTTTTATTATGCCATTTATAACCTTAGAGCGTATTTTAGCACATACA   95520
                                               ORF320>  ←------ --→←-------→      M  P  F  I  T  L  E  R  I  L  A  H  T

TCTTTTTTCCTTCTTTTTTTTGTTACGTTTATTTATTGGGGAAAATTTCTTTATATAAATATTAAACCAATAACTATTTTAGGAGAAATAAGTATGAAAATTGCTTGTTTTTTTATAACA   95640
S  F  F  L  L  F  F  V  T  F  I  Y  W  G  K  F  L  Y  I  N  I  K  P  I  T  I  L  G  E  I  S  M  K  I  A  C  F  F  I  T

ACTTTTTTATTAATTCGTTGGAGTTCTTCAGGACATTTTCCTTTAAGTAATTTATACGAATCTTCTATGTTTCTTTCTTGGAGTTTTACATTAATTCATTTAATTTTAGAAAACAAAAGC   95760
T  F  L  L  I  R  W  S  S  S  G  H  F  P  L  S  N  L  Y  E  S  S  M  F  L  S  W  S  F  T  L  I  H  L  I  L  E  N  K  S

AAAAACACATGGTTAGGTATAATAACTGCACCAAGCGCAATGTTAACTCATGGATTTGCAACTTTAAGTCTCCCAAAAGAAATGCAAGAATCTGTTTTTTTAGTTCCAGCTTTACAATCT   95880
K  N  T  W  L  G  I  I  T  A  P  S  A  M  L  T  H  G  F  A  T  L  S  L  P  K  E  M  Q  E  S  V  F  L  V  P  A  L  Q  S

CATTGGTTAATGATGCATGTAACTATGATGATGTTAAGTTATTCTACTCTTTTATGCGGATCTTTATTAGCAATAACTATTTTAATTATTACATTAACAAAACAAAAAAATTTGCCAATA   96000
H  W  L  M  M  H  V  T  M  M  M  L  S  Y  S  T  L  L  C  G  S  L  L  A  I  T  I  L  I  I  T  L  T  K  Q  K  N  L  P  I

CTTACATCTTATTTTAATTTTCCTTTTAATTCTTTTATTTTTAAAAATCTTTTACAACCAATGGAAAATGAAATATTATCATATAAAACGCAAAAAGTTTTTTCTTTTATTAATTTTCGT   96120
L  T  S  Y  F  N  F  P  F  N  S  F  I  F  K  N  L  L  Q  P  M  E  N  E  I  L  S  Y  K  T  Q  K  V  F  S  F  I  N  F  R

AAATGGCAATTAATAAAAGAATTAGATAATTGGAGTTATAGAGTTATTAGTTTAGGATTTCCTCTCTTAACTATTGGTATTCTATCTGGAGCAGTATGGGCTAATGAAGCATGGGGCTCC   96240
K  W  Q  L  I  K  E  L  D  N  W  S  Y  R  V  I  S  L  G  F  P  L  L  T  I  G  I  L  S  G  A  V  W  A  N  E  A  W  G  S

TATTGGAATTGGGATCCGAAAGAAACTTGGGCTTTAATTACTTGGTTAATATTTGCTATTTATTTGCATACTCGAATGATTAAAGGTTGGCAAGGAAAAAAACCGGCAATTATAGCTTCG   96360
Y  W  N  W  D  P  K  E  T  W  A  L  I  T  W  L  I  F  A  I  Y  L  H  T  R  M  I  K  G  W  Q  G  K  K  P  A  I  I  A  S

TTAGGTTTTTTTATTGTTTGGATTTGTTATTTAGGAGTTAATTTATTAGGAAAAGGTTTACATAGCTATGGATGGTTAATTTAACATTAAATATAGATATTGAAATATATAGAAATATAT   96480
L  G  F  F  I  V  W  I  C  Y  L  G  V  N  L  L  G  K  G  L  H  S  Y  G  W  L  I  ——

ATTATTTCAAATAGTGGATTTATAGATAAGTTATTTATATAATCCACTATTTGAAATAATGAAAAACTAAGTTCACTTATTTTTAGTAAAATTATAACATTTTAAATGTTTTCAAATTTA   96600
←----------------------------→ ←-------- ------------------→
```

(d) *frxC - ndh4*

```
aaaaaaaaacaaaacatacaaaaaattatgaaaaTAGCAGTTTATGGGAAAGGTGGCATAGGAAAATCTACAACTAGTTGTAATATTTCTATTGCATTAGCAAGACGTGGGAAAAAAGTT   110881
                    frxC>  M  K  I  A  V  Y  G  K  G  G  I  G  K  S  T  T  S  C  N  I  S  I  A  L  A  R  R  G  K  K  V

TTACAAATCGGTTGTGATCCAAAACATGACAGTACATTCACACTTACAGGATTTTTAATTCCTACAATTATAGATACTTTACAATCAAAAGATTATCATTACGAAGATGTTTGGCCTGAA   110761
L  Q  I  G  C  D  P  K  H  D  S  T  F  T  L  T  G  F  L  I  P  T  I  I  D  T  L  Q  S  K  D  Y  H  Y  E  D  V  W  P  E

GATGTAATATATAAAGGTTATGGCCGGTGTGATTGTGTAGAAGCTGGAGGACCTCCTGCTGGAGCTGGATGCGGGGGTTATGTTGTCGGAGAAACTGTTAAATTATTAAAAGAATTAAAT   110641
D  V  I  Y  K  G  Y  G  R  C  D  C  V  E  A  G  G  P  P  A  G  A  G  C  G  G  Y  V  V  G  E  T  V  K  L  L  K  E  L  N

GCTTTTTATGAATATGATATTATTTTATTTGATGTTCTAGGGGATGTAGTATGCGGTGGCTTTGCTGCTCCATTAAATTATGCAGATTATTGTATTATTATTACAGATAATGGATTTGAT   110521
A  F  Y  E  Y  D  I  I  L  F  D  V  L  G  D  V  V  C  G  G  F  A  A  P  L  N  Y  A  D  Y  C  I  I  I  T  D  N  G  F  D

GCGTTATTTGCTGCTAATAGAATAGCAGCTTCAGTAAGAGAAAAAGCTCGTACACATCCTCTTAGATTAGCAGGCTTAGTTGGAAATCGTACATCAAAACGAGATTTAATTGATAAATAT   110401
A  L  F  A  A  N  R  I  A  A  S  V  R  E  K  A  R  T  H  P  L  R  L  A  G  L  V  G  N  R  T  S  K  R  D  L  I  D  K  Y

GTTGAAGCTTGTCCAATGCCAGTTCTAGAAGTATTACCTCTTATTGAAGATATTCGAGTTTCTAGAGTTAAAGGTAAAACTTTATTTGAAATGGTAGAATTACAACCCAGTCTTAAATAT   110281
V  E  A  C  P  M  P  V  L  E  V  L  P  L  I  E  D  I  R  V  S  R  V  K  G  K  T  L  F  E  M  V  E  L  Q  P  S  L  K  Y

GTTTGTGATTTTTATTTAAATATAGCAGATCAAATTTTATCGAAACCAGAAGGAATTATTCCAAAAGAAGTTCCAGATCGAGAATTATTTAGTTTACTTTCAGATTTTTATTTAAATCCT   110161
V  C  D  F  Y  L  N  I  A  D  Q  I  L  S  K  P  E  G  I  I  P  K  E  V  P  D  R  E  L  F  S  L  L  S  D  F  Y  L  N  P

GTAAACACTGTAAATGAAAAAAATAAACCAAATTTAATTGATTTTATGATAATTTAGAAAAAAATTTTTGTTTTTTAGTTAAGGAAATATATAAATATGTCAATAAAAATATCTGAAACT   110041
V  N  T  V  N  E  K  N  K  P  N  L  I  D  F  M  I  I  ——               ORF465>   AGGA           M  S  I  K  I  S  E  T

CTCACTTTTGAATGCGAAACAGGTAATTATCATACATTTTGTCCTATTAGTTGCGTAGCATGGTTATATCAAAAAATTGAAGATAGTTTTTTTTTAGTAGTTGGTACAAAAACATGTGGT   109921
L  T  F  E  C  E  T  G  N  Y  H  T  F  C  P  I  S  C  V  A  W  L  Y  Q  K  I  E  D  S  F  F  L  V  V  G  T  K  T  C  G

TATTTTTTACAAAATGCTCTTGGAGTTATGATTTTTGCTGAACCTCGTTATGCTATGGCAGAATTAGAAGAAGGTGATATTTCAGCTCAATTAAATGATTATGAAGAATTAAAACGATTA   109801
Y  F  L  Q  N  A  L  G  V  M  I  F  A  E  P  R  Y  A  M  A  E  L  E  E  G  D  I  S  A  Q  L  N  D  Y  E  E  L  K  R  L

TGTGTTCAAATAAAAAAAGATAGAAACCCTAGTGTAATTATTTGGATTGGTACTTGTACAACAGAAATAATTAAAATGGATTTAGAAGGAATGGCACCAAAATTGGAAAACGAAATCGAA   109681
C  V  Q  I  K  K  D  R  N  P  S  V  I  I  W  I  G  T  C  T  T  E  I  I  K  M  D  L  E  G  M  A  P  K  L  E  N  E  I  E
```

Fig. 2, cont.

```
ATTCCAATTGTTGTTGCAAGAGCTAATGGTCTAGATTATGCTTTTACTCAAGGAGAAGATACTGTTTTAGCTGCTATGGCACATCGTTGTCCTGAACAAAAAACTGAAATTGAAAAAAAA  109561
I P I V V A R A N G L D Y A F T Q G E D T V L A A M A H R C P E Q K T E I E K K

ATAGATGATAAATCTATACAAGAATTATTTTCTTTTCTTCCTCTTAAAACAAAAGAAAAATCCAATAAATCTTTTACTTTAAAAAATACATTTTCTTTAGTTCTTTTTGGTTCTTTACCT  109441
I D D K S I Q E L F S F L P L K T K E K S N K S F T L K N T F S L V L F G S L P

TCAACAGTAGCTTCTCAGCTTAGTTTAGAATTAAAACGTCAATCTATTCATGTTTCAGGTTGGCTACCTGCTCAAAGATATACAGATCTTCCGATATTGGGAGATAAAGTTTACGTTTGT  109321
S T V A S Q L S L E L K R Q S I H V S G W L P A Q R Y T D L P I L G D K V Y V C

GGTGTAAATCCTTTTTTAAGTCGAACAGCAACTACTTTAATGAGGCGTCGAAAATGTAAATTAATTGGAGCTCCTTTTCCAATTGGTCCTGATGGAACTCGTGCATGGATTGAAAAAATT  109201
G V N P F L S R T A T T L M R R R K C K L I G A P F P I G P D G T R A W I E K I

TGTTCAGTTTTTAATATTGAAACACAAGGTTTAGAAGAAAGAGAACAACAAGTGTGGGAAAGTTTAAAAAATTATCTTAATTTAGTACGTGGAAAATCTGTTTTTTTTATGGGAGATAAT  109081
C S V F N I E T Q G L E E R E Q Q V W E S L K N Y L N L V R G K S V F F M G D N

CTTTTAGAAATCTCTTTAGCAAGATTTTTAATCCGATGTGGTATGATTGTTTATGAAATTGGAATTCCATATATGGATAAAAGATATCAAGCTGCAGAATTAACACTTTTGCAAGAAACT  108961
L L E I S L A R F L I R C G M I V Y E I G I P Y M D K R Y Q A A E L T L L Q E T

TGTAAAAAAATGTGTATACCAATGCCTCGAATTGTTGAAAAACCTGATAATTATAATCAAATACAGCGTATGCGTGAATTACAACCGGATTTAGCTATTACAGGAATGGCTCATGCAAAT  108841
C K K M C I P M P R I V E K P D N Y N Q I Q R M R E L Q P D L A I T G M A H A N

CCACTTGAAGCGAGAGGTATTAATACAAAATGGTCTGTTGAATTTACTTTCGCTCAAATTCATGGATTTACAAATGCCAAAGATGTTCTTGAACTTGTTACACGTCCTTTACGTCGTAAT  108721
P L E A R G I N T K W S V E F T F A Q I H G F T N A K D V L E L V T R P L R R N

AATAATTTAGAAAATTTAGGTTGGACTAATTTAATAAAAATACAAAAAAGATAAAAAAAAAAAGACTTATTCATTAGTAGACTAAATTTATTATATTAATGAATAAGTCTTTTTTTACAAA  108601
N N L E N L G W T N L I K I Q K R ——→ ←———————————— — - ——>< —— - -- ——————————→

TATTCGGTTTTTTTCTAATAATTTTATTCTATTTTAATCCTATTTTATTGAGGAAGGTTTTTTATTATGATAACAAGCATTCCCTTATTGCTTTCTGTGCTATGGGTTCCAATATTATCAT  108481
                              ORF1068>   GAGG          M I T S I P L L L S V L W V P I L S W

GGATAAATTTTTCAAGTACATTTTTTTTTGTTTGGAATATATTATGGATTTCTGACTACTTTACCTATTGGTCCCTCTCAACTGTTATCTATAAGAGCTTTTCTGTTAGAAGGAAATTTTA  108361
 I N F S S T F F L F G I Y Y G F L T T L P I G P S Q L L S I R A F L L E G N F S

GTGGTATAGCAGCTGTTAGTGGACTAATTACAGGACAATTGTTGATATTTTTATCAATTTTTTATTCACCATTGTATGTTCTATTAATAAAACCTCACTTATTAACTTTGTTGGTTTTAC  108241
 G I A A V S G L I T G Q L L I F L S I F Y S P L Y V L L I K P H L L T L L V L P

CATATATTTTATTCTATTGGTATAAAATTAAAGATTTAATAGATTATCAATCTTTAAAACCTATAACATCTATTAAAGATACTCGAATTTCTAAAATATTTTTTGATAGTTTTATATTTC  108121
 Y I L F Y W Y K I K D L I D Y Q S L K P I T S I K D T R I S K I F F D S F I F Q

AATTATTTAATCCTGTTGTATTACCAAGTCCTGTATTAGCAAGATTACTAAATATTTTTCTTTTTCGTTATAGTAATAATTTTATTTTTTTACTAAGTAGTTTTTTAGGTTGGTGTTTTG  108001
 L F N P V V L P S P V L A R L L N I F L F R Y S N N F I F L L S S F L G W C F G

GTCAATTTTTATTTGTAAGTTTAGGCAAATTACTTCTATTTCGTATTGAATCGGATTCACCTATTCTTTATCTTTTAGTTAAACGTATTATTTATCGAACTTTTAGTATTATTATATTAA  107881
 Q F L F V S L G K L L L F R I E S D S P I L Y L L V K R I I Y R T F S I I I L S

GTTTTTCATTATTACATTTAAGTAGAGCTCCAGTACCTTTTATTACAAAAAAACTTAATGATAATTTGCAATTTAATTTATCAAAACCAGAGGATTCTTTTGTATTAACAAAATCTTGGC  107761
 F S L L H L S R A P V P F I T K K L N D N L Q F N L S K P E D S F V L T K S W P

CGACTCTTTTTTTTGATTATCGCAAATGGAATAGACCATTACGATATATAGAAAATAGTAGATTTAGTAGTCAAAGTCCTATAAAAAAAAAAGTATCTCAATATTTTTTTAATATTTCTT  107641
 T L F F D Y R K W N R P L R Y I E N S R F S S Q S P I K K K V S Q Y F F N I S L

TAAGTGATGGAAAACCAAGACTATCTTTTACATATTTACCAAGTTTGTATTATTTTGAAAAAAATTTGCAAAAATCCTCTATTAATTTCAATCTTTTTTCATCTAATGAAATTTATGAAA  107521
 S D G K P R L S F T Y L P S L Y Y F E K N L Q K S S I N F N L F S S N E I Y E K

AATGGATTAAAAATAAAAAAAATAAAAAATTAAAGATATATAAAGAATTTAAAAATCGATTTAAATTTTTAGATAATGGATTTTTTTTAGCAGAAATTATTGAAAAAAAAAACATATTGT  107401
 W I K N K K N K K L K I Y K E F K N R F K F L D N G F F L A E I I E K K N I L S

CAACTTTTGAAGGAAATATTTTTACAAAAATTTGTGATCCTTTGTTAATTAAACAATATGATAAAAAAATGATAGTATCAAAATCACCATGGCTTTTAACAGAAAAATCTTATAAATTAA  107281
 T F E G N I F T K I C D P L L I K Q Y D K K M I V S K S P W L L T E K S Y K L T

CAAAAACGCAAAAAACACTTACTTTTTCTAAAAAAGATAATAAACTGAAAAAGTGGATTTCTAATCAATGTCAAGAATTTGAAGATAAAAATTTCATTTTACCTTGGGAACCTTTAACTC  107161
 K T Q K T L T F S K K D N K L K K W I S N Q C Q E F E D K N F I L P W E P L T Q

AAGATGCTAGGCGCATTTTAAGTTTACTTATTAACAAATCAAAAAAAACAAAAATTGATACAAATTTGAAACAAATGAATTTTTTTGATGAAAATGCAACCCAATTATTGAATAAACAAA  107041
 D A R R I L S L L I N K S K K T K I D T N L K Q M N F F D E N A T Q L L N K Q N

ATCTTTCTTCAATTGAAAATACACGTAAAAAAATAAATAGAAAATCAAATCTAAATTGGGAACTTATTTTAAATTTATCTCCTCGCCAAAAAATTTTATTTTTGAATTATTTACAAAAAG  106921
 L S S I E N T R K K I N R K S N L N W E L I L N L S P R Q K I L F L N Y L Q K D

ATAAATGGAACACGCTTAAAATTTCTTGGAAAAATTTTTTTTTAGGTGATTTTACTCAAATAAAAAATATTCTTTTTTTATTAACAAAGATAATTAAACCTGATCAAAACTATCAATTTC  106801
 K W N T L K I S W K N F F L G D F T Q I K N I L F L L T K I I K P D Q N Y Q F Q

AAGAAATAAATAAAGAAATACCAAGATGGACTTCAAAATTAAAAAATGATAAATTTGACGTTATAGCTATCGGAGTTACTGATATTCGTCAAAGAAAAGTTAAAAATTTAGGATATTTAA  106681
 E I N K E I P R W T S K L K N D K F D V I A I G V T D I R Q R K V K N L G Y L I

TAAAAGGAAAAGATAAAAGAAGAAAAATAATAAGACGTTTTTCACAACAATCTGATTTTCGTCGAAAATTAGTAAAAGGATCTATGCGTGCTCGACGACGTAAAACTTTAATATGGAAAA  106561
 K G K D K R R K I I R R F S Q Q S D F R R K L V K G S M R A R R R K T L I W K I

TTTTTCAAGTTAAAATTAATTCTCCATTTTTTTTACGAATAATGGATAAACCGAATTTAAACGTTAACAATGTTTTGAAAATAAAACCAACATTTCAAAATATTTTAGAAAAAAAAAAAA  106441
 F Q V K I N S P F F L R I M D K P N L N V N N V L K I K P T F Q N I L E K K K K

AAGAATCGCTAAATCAAAAAGCTTTGTTTATAAAAAGAACAAAAGCAGATCGTTTTGCTATAGCAAACAGATGGGATTTTCCATTAGCTCAGTGGGGAAGAAGTTGGTTACTACTTATTC  106321
 E S L N Q K A L F I K R T K A D R F A I A N R W D F P L A Q W G R S W L L L I Q

AATCACATCTAAGAAAATATATATTATTACCTATATTAATAATATTTAAAAATGTCATTCGTTTGTTTTTATTTCAAATTCCGGAATGGAATCAAGATTGGTATGAATGGAATAAAGAAA  106201
 S H L R K Y I L L P I L I I F K N V I R L F L F Q I P E W N Q D W Y E W N K E I

TTCATATAAGATGTACATATGATGGGACTGAAGTTTCAGAAAAAGAATTACCAGAACAATGGCTTAGAGATGGTCTTCAAATAAAAATTATTTATCCTTTTTATTTAAAACCTTGGCATA  106081
 H I R C T Y D G T E V S E K E L P E Q W L R D G L Q I K I I Y P F Y L K P W H N
```

Fig. 2, cont.

```
ATATTCAAAATAGAAATAATTTACCTAATAAAAAAAATGAAAAATTGGATTTAATTTATGATGATACAAATTTTCTACAAAATTTAGTAAAAACTAAAGAACATTCTAATAATTTAGTTA  105961
 I  Q  N  R  N  N  L  P  N  K  K  N  E  K  L  D  L  I  Y  D  D  T  N  F  L  Q  H  L  V  K  T  K  E  H  S  N  N  L  V  K

AAAAGAAAAAATTAAATTATTGCTATTTAACTGCTTGGGGATTTCAAACTAATTTACCGTTTGGTAATATAAAAAAACAACCTTCTTTTTGGAAACCGATAAAAAAAAAATTAAAAAAAA  105841
 K  K  K  L  N  Y  C  Y  L  T  A  W  G  F  Q  T  N  L  P  F  G  N  I  K  K  Q  P  S  F  W  K  P  I  X  K  K  L  K  K  N

ATATGTTTTTTAAACCATATCAAAATTTAAAAAATATTTCAACGAAAAACAAATTATATAAAATTTCTGATGTTGAGAATTTAAAAAATTTTAACAGTAAAAAAAAGGAATATATTAATC  105721
 M  F  F  K  P  Y  Q  N  L  K  N  I  S  T  K  N  K  L  Y  K  I  S  D  V  E  N  L  K  N  F  N  S  K  K  K  E  Y  I  N  P

CTAATTTAAATAATCTAGATTTCAAAAAAAAAAATGTAGTAATTACTAAATTAAATAATCAAAATACATTATTAAAACAAAACACTGACAATGACATTTTTTCTATTAATTTTGAATATA  105601
 N  L  N  N  L  D  F  K  K  K  N  V  V  I  T  K  L  N  N  Q  N  T  L  L  K  Q  N  T  D  N  D  I  F  S  I  N  F  E  Y  K

AAACTTCAATATATATATAAATAATTTAGAAAATTTACTAAAAAAAAAACATATATCTATAGAAAAATATTTAAAAAAACAAAAAACATTAAATTTAAAAAAAAAATTAATAATGATAAAAC  105481
 T  S  I  Y  I  N  N  L  E  N  L  L  K  K  K  H  I  S  I  E  K  Y  L  K  K  Q  K  T  L  N  L  K  K  K  L  I  M  I  K  Q

AAAAAACTATTAAAATTTTAAAAAAAAAATGTTCAATTAATAAAAAAATTACCTAAAAATTTGAAAATAAATATTCAAAAAATTTATATAAATTTGAAAATAAATAAAACAAAATTGATTA  105361
 K  T  I  K  I  L  K  K  N  V  Q  L  I  K  K  L  P  K  N  L  K  I  N  I  Q  K  I  Y  I  N  L  X  I  N  K  T  K  L  I  N

ATAACATTAGTAAGTTTTTTCAAGATAATTAAAAAAAATATTCAATTGTAAATAATAAAAAAAGAAAAAATTAAAGTTAATTTGAAACAAATAATGATATTAATTTCTCAAGCATATGTA  105241
 N  I  S  K  F  F  Q  D  N  ---                                                    ORF464>  M  I  L  I  S  Q  A  Y  V

TTTAATAAAATATGGGAAATAAAAACAAAAAATAAGTCTTATTTAAAATGTTTATTAAAATATTGGACATCCCATTTATGGATAAAAAAAAATTTTCAAAGTTTTTTATCTAATCAAGGA  105121
 F  N  K  I  W  E  I  K  T  K  N  K  S  Y  L  K  C  L  L  K  Y  W  T  S  H  L  W  I  K  K  N  F  Q  S  F  L  S  N  Q  G

ATCGTTGGTTCTCTAGAATTACAAAATTTTAAAGAAGAAAATTGGAAAGAATGGTTAAAAGGTTTTAATCGATATAATTTTTCATCCAAAGAATGGTATAAAATAACACCTCAGCAATGG  105001
 I  V  G  S  L  E  L  Q  N  F  K  E  E  N  W  K  E  W  L  K  G  F  N  R  Y  N  F  S  S  K  E  W  Y  K  I  T  P  Q  Q  W

AGAAATAAAGTTAGTGAACATTGGAAAAACCAAGAAAATAAAAAATTAAATCCTAATCAACAAATTAGTAAAAATAATTTTTTTTATTAATACTTCTATTTTAGAACAAACTAAAAAACGT  104881
 R  N  K  V  S  E  H  W  K  N  Q  E  N  K  K  L  N  P  N  Q  Q  I  S  K  N  N  F  F  I  N  T  S  I  L  E  Q  T  K  K  R

AATAAAATATTTAAACAAAATTTATTAACTTATAGTTGTTTTGATTTTACAAAAAATTTAGCAATTAGAAACTTTTTGAATTTAAACAGAAAAAAAATATATAATAATATTATTATAAAT  104761
 N  K  I  F  K  Q  N  L  L  T  Y  S  C  F  D  F  T  K  N  L  A  I  R  N  F  L  N  L  N  R  K  K  I  Y  N  N  I  I  I  N

AAAATACAGAAATCTTATTTTATTTATAATAAAAAAGCAAAATATTTAGATTTTTTTTCTCAAAAACAAAATATTTTTTTTCGAATATAATCTATTATTATGGCTTATTCCAGAATTTATA  104641
 K  I  Q  K  S  Y  F  I  Y  N  K  K  A  K  Y  L  D  F  F  S  Q  K  Q  N  I  F  F  E  Y  N  L  L  L  W  L  I  P  E  F  I

GAAGAAAAAAATCAATATCAAAATAAAAGAATTTTAATTCTAAAAAATTCTATTATTAAAGAAAATAATAAAAAAATAATTCAAAATCAAAAATTATTTCGAAAAAGAGAACTTAATCAA  104521
 E  E  K  N  Q  Y  Q  N  K  R  I  L  I  L  K  N  S  I  I  K  E  N  N  K  K  I  I  Q  N  Q  K  L  F  R  K  R  E  L  N  Q

TCTATTCGCCAATGGAGATGGAAATCAAAAAGTTTAGAAAAAAAAATTTAAAAAAATTAGGAAATATGGCTTCTCTAATGACTTTTATGCAAAATCAAGAAAATATTATTTCTCTTTCCAGT  104401
 S  I  R  Q  W  R  W  K  S  K  S  L  E  K  K  F  K  K  L  G  N  M  A  S  L  M  T  F  M  Q  N  Q  E  N  I  I  S  L  S  S

AAAATGCGAGAAGATTTAAAATTATTTCATCTTTTTTTTCGTCGAAATACTACTATAAATCAATTAACTATTAATTCAGAACATCGTTTAGCACGGTTATTAGATGATCAAATTTTAATG  104281
 K  M  R  E  D  L  K  L  F  H  L  F  F  R  R  N  T  T  I  N  Q  L  T  I  N  S  E  H  R  L  A  R  L  L  D  D  Q  I  L  M

TATAAAATGGTAAGTACTTTTTTAAATATTAAATATAGATTTAAACGACTATCAAATTTAGATAATTTTGATGATTTTTTAGGAATACAATTTTTTGAAAATAAAGAAAAAAATAATTTC  104161
 Y  K  M  V  S  T  F  L  N  I  K  Y  R  F  K  R  L  S  N  L  D  N  F  D  D  F  L  G  I  Q  F  F  E  N  K  E  K  N  N  F

TTTTTTTTTAATTCGTTTAATCTTGAAGATATTTTACTTCCAAAACGTCGTAGAAAATTTCGAATTTTAAATTCTTTAACTTCAAAAAATAAAAAAAATACACAACTTAATCAAAAATTT  104041
 F  F  F  N  S  F  N  L  E  D  I  L  L  P  K  R  R  R  K  F  R  I  L  N  S  L  T  S  K  N  K  K  N  T  Q  L  N  Q  K  F

GTTCAAAAAAAATTTTCTAAAACAAAAATAAAAAAAATTAAACGTTTTATTTGGGCTAGTTATCGATTTGAAGATTTAGCTTGTATGAATAGATTTTGGTTTAATACAATAAACGGTAGT  103921
 V  Q  K  K  F  S  K  T  K  I  K  K  I  K  R  F  I  W  A  S  Y  R  F  E  D  L  A  C  M  N  R  F  W  F  N  T  I  N  G  S

AGATTTTCTATGCTTAGGTTTCGAATGTATCCTTCTTTATTAACTTAAAAAAATTGAATTTTATTTTTTCTTGTTAGAAAATAGAGTTTATAAACTCTATTTTCTAATTAATTTTTCAAC  103801
 R  F  S  M  L  R  F  R  M  Y  P  S  L  L  T  ------->        <----->    <--------------><----------->

ATAAAGTAATTTTATTATTGTATTAATTACATTTTTATTTTTTTACTATTTTTCTATCACAAATAAAGAAATTTATTAAATATTTTTTAGGAGAATTTTTTATGTCAAAAAATTTGTTTA  103681
                  TTGTAT>                   TTTACT>                                        rpml5>  AGGAG      M  S  K  N  L  F  M

TGGATTTATCTTCCATTTCTGAAAAAGAAAAAGGATCTGTTGAATTTCAAATATTTAGATTAACTAATCGAGTTGTAAAATTAACTTATCATTTTAAAAAACATGGTAAAGATTATTCAT  103561
   D  L  S  S  I  S  E  K  E  K  G  S  V  E  F  Q  I  F  R  L  T  N  R  V  V  K  L  T  Y  H  F  K  K  H  G  K  D  Y  S  S

CTCAAAGAGGTTTATGGAAAATTTTAGGAAAACGTAAACGTCTTTTAGCTTATTTATTTAAAACGAATTTTGTTAGTTACGAAAATTTAATTATTCAATTAGGGATTCGCGGATTAAAAA  103441
   Q  R  G  L  W  K  I  L  G  K  R  K  R  L  L  A  Y  L  F  K  T  N  F  V  S  Y  E  N  L  I  I  Q  L  G  I  R  G  L  K  X

AAAATTAAAAATTTTTTTTTAGTTTTTTATTATAAAAAAAAATGAAAAGGAGAGTTATATATGATGATACTTACTAAAAACAAACCAATGATAGTAAGTATGGGTCCTCATCATCCATCA  103321
   N  ---               +------> <------+ORF392> AGGAG       M  M  I  L  T  K  N  K  P  M  I  V  S  M  G  P  H  H  P  S

ATGCATGGTGTTCTTCGACTTATTGTTACTTTAGATGGAGAAGATGTTTTGGATTGCGAACCTGTCTTAGGTTATTTACATAGAGGAATGGAAAAAATAGCTGAAAATAGAACAATTGTA  103201
 M  H  G  V  L  R  L  I  V  T  L  D  G  E  D  V  L  D  C  E  P  V  L  G  Y  L  H  R  G  M  E  K  I  A  E  N  R  T  I  V

CAATATCTTCCTTACGTAACACGATGGGATTATTTAGCTACAATGTTTACAGAAGCAATAACTGTAAATGCACCAGAAAAACTAACAAATATTCAAGTTCCTAAAAGAGCGAGTTATATA  103081
 Q  Y  L  P  Y  V  T  R  W  D  Y  L  A  T  M  F  T  E  A  I  T  V  N  A  P  E  K  L  T  N  I  Q  V  P  K  R  A  S  Y  I

CGAATCATTATGCTGGAATTAAGTCGTATTGCATCCCATTTGCTATGGCTTGGACCTTTTATGGCTGATATTGGTGCACAAACTCCTTTTTTTTATATTTTTAGAGAAAGAGAAATGATT  102961
 R  I  I  M  L  E  L  S  R  I  A  S  H  L  L  W  L  G  P  F  M  A  D  I  G  A  Q  T  P  F  F  Y  I  F  R  E  R  E  M  I

TATGATTTATTTGAATCTGCTACTGGAATGCGAATGATGCATAATTATTTTCGAATTGGAGGAGTTGCAGTAGATTTACCTTATGGTTGGATAGACAAATGTTTAGATTTTTGTGATTAT  102841
 Y  D  L  F  E  S  A  T  G  M  R  M  M  H  N  Y  F  R  I  G  G  V  A  V  D  L  P  Y  G  W  I  D  K  C  L  D  F  C  D  Y

TTTTTACCTAAAATAAATGAATATGAAAGACTTATTACAAATAATCCTATTTTTTTGAAACGAGTAGAAGGGATAGGTACTGTTACTAGAGAAGAAGCTATTAATTGGGGATTATCAGGT  102721
 F  L  P  K  I  N  E  Y  E  R  L  I  T  N  N  P  I  F  L  K  R  V  E  G  I  G  T  V  T  R  E  E  A  I  N  W  G  L  S  G

CCTATGTTACGAGCTTCAGGAGTTCAATGGGACCTTCGAAAAGTAGATCATTATGAATGTTATGATGAGTTAGATTGGAAAATTCAATGGCAAAAAGAAGGGGATTCATTAGCTCGTTAT  102601
 P  M  L  R  A  S  G  V  Q  W  D  L  R  K  V  D  H  Y  E  C  Y  D  E  L  D  W  K  I  Q  W  Q  K  E  G  D  S  L  A  R  Y

TTAGTAAGAATTGGTGAAATGAAAGAATCTGTTAAAATAATTCAACAAGCTTTAAAAGCTATTCCGGGAGGACCTTTTGAAAATTTAGAAGCACGCCGGCTTAATCAAGGAAAAAATTCA  102481
 L  V  R  I  G  E  M  K  E  S  V  K  I  I  Q  Q  A  L  K  A  I  P  G  G  P  F  E  N  L  E  A  R  R  L  N  Q  G  K  N  S
```

Fig. 2, cont.

```
GAATGGAATTTATTTGAATATCAATTTATTAGTAAAAAACCTTCACCAACTTTTAAATTACCTAAACAAGAACATTATGTAAGAGTAGAAGCACCAAAAGGAGAATTAGGTATTTTTTTA  102361
E  W  N  L  F  E  Y  Q  F  I  S  K  X  P  S  P  T  F  K  L  P  K  Q  E  H  Y  V  R  V  E  A  P  K  G  E  L  G  I  F  L

ATTGGAGATGATAGTGTTTTTCCTTGGAGACTTAAAATTCGTTCACCTGGTTTTATAAATTTACAAATTCTTCCTCAATTAGTAAAAGGAATGAAATTAGCAGATATTATGACAATTTTA  102241
I  G  D  D  S  V  F  P  W  R  L  K  I  R  S  P  G  F  I  N  L  Q  I  L  P  Q  L  V  K  G  M  K  L  A  D  I  M  T  I  L

GGTAGTATAGACATAATTATGGGGGAGGTTGATCGTTAAAATGATTTCAAATATAAATTTAGAAGACAAATTTTTTTCCTTTTTTTTTACATTAGGTTTTTCTAAAGAATTTTTTAATTT  102121
G  S  I  D  I  I  M  G  E  V  D  R  —
         ndhI>  GGAGG          M  I  S  N  I  N  L  E  D  K  F  F  S  F  F  F  T  L  G  F  S  K  E  F  F  N  F

TTTATGGATTATTTTTTCTATTTTAATTCTTATGTTAGGAGTTACTATTGGAGTACTAGTACTTGTATGGCTTGAAAGAAAAATATCTGCTGCAATCCAGCAACGGATTGGACCAGAATA  102001
 L  W  I  I  F  S  I  L  I  L (M) L  G  V  T  I  G  V  L  V  L  V  W  L  E  R  K  I  S  A  A  I  Q  Q  R  I  G  P  E  Y

TGCTGGTCCTTTAGGAATAATCCAAGCTTTAGCGGATGGAATTAAACTTTTTTTAAAAGAAGATATTGTTCCAGCACAAGGAGATGTTTGGTTATTTAATATTGGACCTATTTTGGTTCT  101881
 A  G  P  L  G  I  I  Q  A  L  A  D  G  I  K  L  F  L  K  E  D  I  V  P  A  Q  G  D  V  W  L  F  N  I  G  P  I  L  V  L

TATACCAGTCTTTTTAAGTTATTTAGTAATTCCTTTTGAATATAATGTTATTTTAGCTAATTTTAGTATAGGGGTTTTTTTTTGGATTGCTGTTTCTAGTGTTGTTCCTCTTGGACTTCT  101761
 I  P  V  F  L  S  Y  L  V  I  P  F  E  Y  N  V  I  L  A  N  F  S  I  G  V  F  F  W  I  A  V  S  S  V  V  P  L  G  L  L

TATGGCTGGTTATGGATCAAATAATAAGTATTCTTTTTTAGGTGGTTTAAGAGCTGCTGCTCAATCTATTAGTTATGAAATTCCTTTAGCTTTAAGTGTTTTATCTATAGCTCTACGTGT  101641
 M  A  G  Y  G  S  N  N  K  Y  S  F  L  G  G  L  R  A  A  A  Q  S  I  S  Y  E  I  P  L  A  L  S  V  L  S  I  A  L  Lgugy

GATTCGTTAAAGTACTTAAAAAAATTTAGTAATAAATTTTTTAATTGTTTAATAAAAAACTAAATAGTCATATGGGTGAGATTAAACAGCATTTTAATTTGCAGTAAAAAAATCAAGTCC  101521
g..........................................(intron)...........................................................

CATCCTCTTTGTACAAGAGTGAAAGCTATATACAATCAATAAAAAAGTATATTTTCCAGGTAAAAGAGTAGCAATCTAGTCCTGACAGCGAAAAGAAAAAATCAATAATATAATTTTATT  101401
...........................................(intron)...........................................................

ATATATATTGAATAAGCAAGAGTAGAAGGTAAGAAAACCCAAAATACACAAAAAATTAGTGAAAAATATATTAAATATATTAAAATGAGAATAAGGACTTAATATTAGTAGAGATTTTTA  101281
...........................................(intron)...........................................................

AGTAGTACTTCCTTTGAATCTCAATTAAAAGTATAAACCTATCTATTACAAAAATGACTATTAGGAACGAAATAATTTTTTTACAATTCTTCAATTAAAAAAAATTGATATTGTTTTATT  101161
...........................................(intron)...........................................................

TATAAATAAAACAATATCAATTTTTTTTAATGTAAAGTTAGCGCTAGCCAAAATTGTAAAAATTAAGATAGATTCATAAGCTTTTATTTTATAGCAAACAGAATCTCGTTGGTAAAAAAC  101041
...........................................(intron)...........................................................

TATATATTTTTTATATTTAAATAACCACTGAGGAGCCGTATGAAATGAAAATTTCATGTACGGTTTTGCAATAGAGATATAAATAGTAATGTTCATATCGACTATAATTATCAAATAGTT  100921
.............................ragccg-augaa--gaaa--uucaugu-cgguuy...............................cuayy-y-ay  S  N  S  L

TAAGTACAGTTGATATAGTTGAAGCTCAATCTAAATATGGATTTTTAAGTTGGAATTTATGGCGTCAACCTATTGGTTTTATTGTTTTTTTTATTGCTTCTTTAGCAGAATGTGAAAGAC  100801
 S  T  V  D  I  V  E  A  Q  S  K  Y  G  F  L  S  W  N  L  W  R  Q  P  I  G  F  I  V  F  F  I  A  S  L  A  E  C  E  R  L

TCCCTTTTGATTTACCAGAAGCTGAAGAAGAGTTAGTTGCGGGTTATCAAACTGAATATTCAGGCATGAAATTTGCTTTTTTTTATCTAGCTTCTTATTTAAATTTGCTAGTTTCTTCAT  100681
 P  F  D  L  P  E  A  E  E  E  L  V  A  G  Y  Q  T  E  Y  S  G  M  K  F  A  F  F  Y  L  A  S  Y  L  N  L  L  V  S  S  L

TATTTGTAACAATTCTTTATTTAGGGGGGTGGCACTTTTCAATTCCATTTTTTTCACTTTTTAAAAATTTTGAATGGAATTTAATGAGTAATGGAATTAGTGAAGTTATTAGCATAATAA  100561
 F  V  T  I  L  Y  L  G  G  W  H  F  S  I  P  F  F  S  L  F  K  N  F  E  W  N  L  M  S  N  G  I  S  E  V  I  S  I  I  I

TAGGAATAGTTATTACATTAGTAAAATCTTATTTATTTTTATTTATTTCAATAATGACAAGATGGACTTTACCTAGAATACGAATTGATCAATTATTAAATCTTGGTTGGAAATTTCTTT  100441
 G  I  V  I  T  L  V  K  S  Y  L  F  L  F  I  S  I  M  T  R  W  T  L  P  R  I  R  I  D  Q  L  L  N  L  G  W  K  F  L  L

TACCTATTGCTTTAGGTAATTTATTATTAACAACGTCTTTTCAACTTTTTTTATTATAAATCTAAAAAAATACATAACTAAAACCATAAATTTATGAAGGAAGTTTTTATATGTTTTCTA  100321
 P  I  A  L  G  N  L  L  L  T  T  S  F  Q  L  F  L  L  —                                       frxB>  AGGA        M  F  S  I

TTATAAATGGCTTAAAAAATTATAATCAACAAGCAATACAAGCTGCAAGATATATTGGTCAAGGGTTTTTAGTTACTTTAGATCATATGAATCGTTTACCTACAACTATTCAATATCCTT  100201
 I  N  G  L  K  N  Y  N  Q  Q  A  I  Q  A  A  R  Y  I  G  Q  G  F  L  V  T  L  D  H  M  N  R  L  P  T  T  I  Q  Y  P  Y

ATGAAAAATTAATACCTTCTGAGCGATTTCGTGGTCGTATTCATTTTGAATTTGATAAGTGTATTGCTTGCGAAGTCTGTGTACGTGTATGCCCAATAAATCTACCAGTTGTAGATTGGG  100081
 E  K  L  I  P  S  E  R  F  R  G  R  I  H  F  E  F  D  K  C  I  A  C  E  V  C  V  R  V  C  P  I  N  L  P  V  V  D  W  E

AATTAAAAAAAACTATAAAAAAAAAAACAATTAAAAAATTATAGTATTGATTTTGGAGTTTGTATATTTTGTGGTAATTGTGTTGAGTATTGTCCTACAAATTGTTTATCTATGACTGAAG   99961
 L  K  K  T  I  K  K  K  Q  L  K  N  Y  S  I  D  F  G  V  C  I  F  C  G  N  C  V  E  Y  C  P  T  N  C  L  S  M  T  E  E

AATACGAACTTTCAACTTATAATCGTCATGAATTAAATTATGATCAAATTGCATTAGGACGTTTACCTATATCAATAATCGAAGATTCTACAATTGAAAATATTTTCAATTTAACTTCTT   99841
 Y  E  L  S  T  Y  N  R  H  E  L  N  Y  D  Q  I  A  L  G  R  L  P  I  S  I  I  E  D  S  T  I  E  N  I  F  N  L  T  S  L

TACCTAAAGGCAAAATAGAAGGGCATATTTATTCTCGAAATATTACAAATATTGTAAATTAATTAATTTTTTTTTAATTGTTCGTTTATTTTTTTTTTTTTTTTAACAAATTCATTTTGTT   99721
 P  K  G  K  I  E  G  H  I  Y  S  R  N  I  T  N  I  V  N  —                                          ←———————→    ←———

AAAAAATATTTGATTTTATTATTGTGAGCTTTATGAAATTACCAGAATCATTTTATGAAACTATTTTTCTTTTTTTTAGAGTCAGGTCTTATATTAGGAAGTTTAGGTGTAATATTATTAA   99601
———→                 ndh6>  M  K  L  P  E  S  F  Y  E  T  I  F  L  F  L  E  S  G  L  I  L  G  S  L  G  V  I  L  L  T

CTAATATAGTTTATTCTGCTCTTTTTTTAGGTTTTGTTTTTGTTTGTATATCGTTATTATATCTTTTATTAAATGCAGATTTTGTGGCTGCCGCACAAATTTTAATTTATGTAGGAGCTG   99481
 N  I  V  Y  S  A  L  F  L  G  F  V  F  V  C  I  S  L  L  Y  L  L  L  N  A  D  F  V  A  A  A  Q  I  L  I  Y  V  G  A  V

TTAATGTATTAATTATTTTTGCTGTTATGTTAATAAATAAAAAACAATATTCCAATTTTTTTGTCTATTGGACAATAGGTGATGGTATTACTTTAACTCTTTGTACAAGTATTTTTTTAT   99361
 N  V  L  I  I  F  A  V  M  L  I  N  K  K  Q  Y  S  N  F  F  V  Y  W  T  I  G  D  G  I  T  L  T  L  C  T  S  I  F  L  L

TATTAAATAATTTTATTTCTAATACATCATGGTCTAAAATTTTTTTAATGACAAAGCCTAATTTAGTAGTAAAAGATATTATTTTAATAAATACGGTTAGGCATATTGGTTCGGAATTAT   99241
 L  N  N  F  I  S  N  T  S  W  S  K  I  F  L  M  T  K  P  N  L  V  V  K  D  I  I  L  I  N  T  V  R  H  I  G  S  E  L  L

TAACTGAATTTTTACTTCCATTTGAACTTATGTCAATAATTCTTTTAGTTGCTTTAATAGGTGCTATTACTTTAGCTCGTCGTGAAAAAAAAATTGAATTAGAAAAAAATGATTTTTTTA   99121
 T  E  F  L  L  P  F  E  L  M  S  I  I  L  L  V  A  L  I  G  A  I  T  L  A  R  R  E  K  K  I  E  L  E  K  N  D  F  F  N

ATTTTTGATTACTATTTTAATAAACAATTTCATAAAATGAAAAATTTTAAGGAAGTTTTTTATGCTTGAACATATACTTACTTTAAGTGCTTTTTTATTTTGTATTGGGGTTTTTGGATT   99001
 F  —                                  ndh4L>  AGGA        M  L  E  H  I  L  T  L  S  A  F  L  F  C  I  G  V  F  G  L

AATTACAAGTCGAAATATGGTAAGAGCATTGATGTGTCTTGAATTAATTTTTAATGCTGTTAATATTAATTTAGTAGCTTTTTCAAATTTTTTAGATAGCTCACAAATTAAAGGAGAAAT   98881
 I  T  S  R  N  M  V  R  A  L  M  C  L  E  L  I  F  N  A  V  N  I  N  L  V  A  F  S  N  F  L  D  S  S  Q  I  K  G  E  I
```

Fig. 2, cont.

```
TTTTCTATTTTTATTATAGCCATCGCTGCTGCTGAAGCCACTATAGGATTAGCTATTGTTTTAGCTATTTATCGAAATAGAAAATCAACTCGTATTGATCAATTTAATTTGCTAAAATG   98761
 F  S  I  F  I  I  A  I  A  A  A  E  A  T  I  G  L  A  I  V  L  A  I  Y  R  N  R  K  S  T  R  I  D  Q  F  N  L  L  K  W

GTAATAGTCTTTATAAAATATTTAATATAGAAATATATGAATAAAAACAATTATATAGATTATGATATATAGTATTAATATATTTATTTATATATGTTTTTTTATATAGAAATAGAAAAA   98641
=====

AGTTTGTATGTATATAAATTTAATTATATAAAAAAATAAATAAAAAAAAATATATATATATATGTTTTATTTAATTTAAGGTTTTTTCCAAATTAGGAGTTAATTAAATGGCACATGCAGT   98521
  TTGATA>                    TATAAA>   <----------------><------ ---- ---- --->      frxA>  AGGAG        M  A  H  A  V

CAAAATTTATGATACATGTATTGGTTGTACTCAATGTGTAAGAGCTTGTCCGACAGATGTATTAGAAATGATACCTTGGGATGGATGTAAAGCTAATCAAATAGCTTCTGCTCCTCGAAC   98401
 K  I  Y  D  T  C  I  G  C  T  Q  C  V  R  A  C  P  T  D  V  L  E  M  I  P  W  D  G  C  K  A  N  Q  I  A  S  A  P  R  T

AGAAGATTGTGTAGGTTGTAAAAGATGTGAATCTCGTTGTCCAACAGATTTTTTAAGTGTACGTGTTTATTTAGGAAATGAGACTACTCGTAGTATGGGTCTAAGTTATTAATTTATACT   98281
 E  D  C  V  G  C  K  R  C  E  S  R  C  P  T  D  F  L  S  V  R  V  Y  L  G  N  E  T  T  R  S  M  G  L  S  Y  ==

ATTGAAAAATAGTTAATATAACTAAATAGTAAATACAAATTTTTTATTTTTTTTTAATAAGAATCTATATATATATATTATATACAGGTTCTTATTAACTTTTTTTTATCTTTATTATGA   98161
                                                          <------- ------->    <------- ------->                 ndh4>   M  N

ACCATTTTCCTTGGCTAACTATTATTGTTCTTTTTCCTATATCTGCAGGTTTAGTAATCCCGTTTTTACCTTCTACAGGGAACAAAATTATTCGATGGTATACTTTAGGTGTTTGTTTAT   98041
 H  F  P  W  L  T  I  I  V  L  F  P  I  S  A  G  L  V  I  P  F  L  P  S  T  G  N  K  I  I  R  W  Y  T  L  G  V  C  L  L

TAGAATTTCTTTTAATAACTTATATTTTTTGTTATCATTATCAATTTAATGATCATTTAATTCAATTAAAAGAAGATTATAATTGGATTAGTTTTATTAATTTTCATTGGAGGTTAGGAA   97921
 E  F  L  L  I  T  Y  I  F  C  Y  H  Y  Q  F  N  D  H  L  I  Q  L  K  E  D  Y  N  W  I  S  F  I  N  F  H  W  R  L  G  I

TTGATGGATTTTCAATAGGACTTATTTTATTAACAGGATTTATAACTACTTTAGCTACTCTCGCTGCTTGGCCCGTAACAAGAAATCCACGATTATTTTATTTTTTGATGTTAGCAATGT   97801
 D  G  F  S  I  G  L  I  L  L  T  G  F  I  T  T  L  A  T  L  A  A  W  P  V  T  R  N  P  R  L  F  Y  F  L  M  L  A  M  Y

ATAGTGGACAAATTGGACTATTTGCTTCTCAAGATATTTTACTCTTTTTTTTTATGTGGGAGTTAGAATTACTTCCTGTTTATTTGCTTTTAGCAATGTGGGGAGGAAAACGACGTTTAT   97681
 S  G  Q  I  G  L  F  A  S  Q  D  I  L  L  F  F  F  M  W  E  L  E  L  L  P  V  Y  L  L  L  A  M  W  G  G  K  R  R  L  Y

ACGCTGCGACAAAATTTATTTTGTATACTGCAGCTGGGTCCCTTTTTATTTTAATAGGGGGACTAATTATGGCATTTTATAATTCTAATGAATTTACTTTCGATTTTCAATTTTTTAATTA   97561
 A  A  T  K  F  I  L  Y  T  A  A  G  S  L  F  I  L  I  G  G  L  I  M  A  F  Y  N  S  N  E  F  T  F  D  F  Q  F  L  I  N

ATAAAAAATATCCTTTGGAATTAGAAATAATAATATATTTAAGTTTTTTAATAGCTTATGCAGTTAAATTACCAATAATACCTTTCCATACTTGGTTACCAGATACACATGGAGAAGCAC   97441
 K  K  Y  P  L  E  L  E  I  I  I  Y  L  S  F  L  I  A  Y  A  V  K  L  P  I  I  P  F  H  T  W  L  P  D  T  H  G  E  A  H

ACTATAGTACATGTATGCTTTTAGCTGGAATACTTTTAAAAATGGGAGCCTATGGATTAATTAGAATTAATATGGAATTACTTCCTCATGCACATTCTTTTTTTGCTCCATGGTTAGTAA   97321
 Y  S  T  C  M  L  L  A  G  I  L  L  K  M  G  A  Y  G  L  I  R  I  N  M  E  L  L  P  H  A  H  S  F  F  A  P  W  L  V  I

TTGTAGGTGCAATTCAAATAGTTTATGCAGCTTTAACTTCTCTAAGTCAACGCAATTTAAAAAGAAGAATTGCTTATTCTTCAGTATCACATATGGGATTTGTTCTTATCGGAATTGGAT   97201
 V  G  A  I  Q  I  V  Y  A  A  L  T  S  L  S  Q  R  N  L  K  R  R  I  A  Y  S  S  V  S  H  M  G  F  V  L  I  G  I  G  S

CGATCACAAATTTAGGGCTTAATGGTGCTATTTTACAAATGATTTCACATGGTTTAATTGGTGCTTCACTTTTTTTCTTAGCAGGAATAAGTTATGATCGAACACGGACTTTGGTTTTAG   97081
 I  T  N  L  G  L  N  G  A  I  L  Q  M  I  S  H  G  L  I  G  A  S  L  F  F  L  A  G  I  S  Y  D  R  T  R  T  L  V  L  D

ATCAAATGGGTGGAATAGGTAATTCTATGCCCAAAATATTCACATTATTTACTAGTTGTTCAATGGCATCTTTAGCTTTGCCAGGTATGAGTGGTTTTATAGCCGAATTAATGATTTTTT   96961
 Q  M  G  G  I  G  N  S  M  P  K  I  F  T  L  F  T  S  C  S  M  A  S  L  A  L  P  G  M  S  G  F  I  A  E  L  M  I  F  L

TAGGAGTAATTGATAATCCTAATTATTCTTCATTATTTAAAATAATAATTATTATAATTCAAGGAATTGGTATTATTTTGACTCCTATTTATTTATTATCCATGTTACGTCAAATGTTTT   96841
 G  V  I  D  N  P  N  Y  S  S  L  F  K  I  I  I  I  I  I  Q  G  I  G  I  I  L  T  P  I  Y  L  L  S  M  L  R  Q  M  F  Y

ATGGATATAAATTTTCAAATACTTTAGAACCATATTTTATGGATGCTGGACCACGAGAAATTTTTATTTTAATTTGTTTATTTTTTCCAATTATAAGTATTGGAATTTATCCTAACTTTG   96721
 G  Y  K  F  S  N  T  L  E  P  Y  F  M  D  A  G  P  R  E  I  F  I  L  I  C  L  F  F  P  I  I  S  I  G  I  Y  P  N  F  V

TTTTATCTATTTGGAATAGTAAAGTAAATTTTCTTTTATCTAATAATTTTTTCTAAAATCTAGAAAATTTAGTATTTCTTTTGAATTGCAAAAAATGAAAGTTCTATAATTTCTAATTAAA   96601
 L  S  I  W  N  S  K  V  N  F  L  L  S  N  N  F  F  ==

TAAATTTGAAAACATTTAAAATGTTATAATTTTACTAAAAATAAGTGAACTTAGTTTTTCATTATTTCAAATAGTGGATTATATAAATAACTTATCTATAAATCCACTATTTGAAATAAT   96481
         <-----> <----->                                               <------------------ -------> <--- -------------------->
```

Fig. 2. Nucleotide sequences and deduced information of IR_B (IR_A) and SSC regions. Nucleotide sequences are shown in the direction of transcription except for tRNA transcripts. RNA transcripts for ribosomal and transfer RNAs (unmodified) are given under the DNA sequences. Genes (bold letters) with arrowhead (> or < indicate the direction of transcription) are shown at the front of their coding sequences. Putative promoter sequences (-35 and -10 sequence) together with SD sequences, if any, are indicated by capitals with arrowheads. Dotted arrows indicate inverted repeated sequences. Double underlining indicates termination codons. Deduced amino acid sequences are shown by one-letter symbols. Transfer RNA genes are shown by the accepting amino acid in three-letter-symbols with anti-codons in parentheses. The 5' and 3' consensus sequences of introns are shown by small letters under the sequence with dotted lines. Gene abbreviations are the same as in the legend of Fig. 1. (a) Nucleotide sequence of the large inverted repeat region (IR_B). The nucleotide sequence of the (-) DNA strand in the IR_A region was the same as that of the (+) DNA strand of IR_B region. (b) Nucleotide sequence of ndh5 ((-) strand, on our file). (c) Nucleotide sequence from rpl21 to ORF320. (d) Nucleotide sequence from frxC to ndh4 ((-) strand, on our map).

rDNA sequence were observed by comparison with rRNA genes of other species, but the liverwort chloroplast rRNAs retain the characteristic secondary structures described by Glotz *et al.* (1981). Putative promoters for *rrn* operons (*rrnA* and *rrnB*) were located upstream from the 16S rRNA gene. There are several reports on the transcriptional unit of the chloroplast *rrn* operon. S1 mapping experiments were done to identify the initiation site of *rrn* gene transcripts using a 5'-labelled probe that contained genes for 16S rRNA and valine $tRNA_{GAC}$. The results showed only two signals. One of them was for the 5'-end of mature 16S rRNA and the other was for the 5'-end of the primary transcript of the *rrn* gene, at a position about 130 bases upstream from the 16S rRNA gene (Umesono *et al.*, unpublished result). This position is in very good agreement with the transcription starting point in maize (Strittmatter *et al.*, 1985). No precursor RNA molecule that would have indicated co-transcription of the *trn*V(GAC) and rRNA genes was detected; 16S rRNA and valine tRNA(GAC) primary transcripts were different, although transcription of the *rrn* operon *in vitro* by *E. coli* RNA polymerase was reported to initiate at the promoter upstream from *trn*V(GAC) in tobacco chloroplasts (Tohdoh *et al.*, 1981). Our results support the idea that the primary transcript of the *rrn* operon is synthesized from its own promoter just upstream from the 16S rRNA gene, as in maize (Strittmatter *et al.*, 1985).

Five tRNA genes were detected in the IR region. None of the tRNA genes encoded the mature CCA sequence at its 3'-terminus. The *trn*I(GAU) and *trn*A(UGC) genes have 886-bp and 768-bp group II introns, respectively (Michel & Dujon, 1983). Introns in these tRNA genes have been reported in maize chloroplasts (Koch *et al.*, 1981) and tobacco chloroplasts (Takaiwa & Sugiura, 1982b), but the splice junctions were not accurately identified. The exon-intron junctions of the liverwort genes were located from the deduced 5'- and 3'-consensus sequences of group II introns in liverwort chloroplasts (Ozeki *et al.*, 1987; Ohyama *et al.*, 1988a). For *trn*I(GAU), the liverwort DNA sequence was compared with the isoleucine tRNA sequences of maize and spinach chloroplast (Guillemaut & Weil, 1982). These tRNA genes are split in the anticodon-loop and anticodon-stem junction; that is, at A_{38}-G_{39} for trnA(UGC) and A_{37}-G_{38} for trnI(GAU). The trnA(UGC) gene is the only gene for alanine tRNA in the chloroplast genome, and alanine $tRNA_{UGC}$ would recognize four GCN-alanine codons by an expanded wobble mechanism. The gene trnR(ACG) contained two mismatched base pairings in its aminoacyl stem with T_5-T_{69} and T_6-T_{68}, and had a highly conserved sequence (80% identical) with respect to trnR(CCG) in the LSC region (Fukuzawa et al., 1988). The trnN(GUU) gene is the only gene for asparagine tRNA in the chloroplast genome, and asparagine $tRNA_{GUU}$ recognizes two codons (AAU and AAC) in the chloroplasts (Ohyama et al., 1988a). The gene trnV(GAC), one of two chloroplast valine tRNA genes, is located upstream from the rrn operon in the IR regions, and the other trnV(UAC) has been mapped in the LSC region (Fukuzawa et al., 1988).

Gene arrangement near the junction between inverted repeats (IR) and large and small single copy (LSC and SSC) regions

In most chloroplast genomes, the IR regions contain both the rps'12-rps7 and trnL(CAA) gene cluster and the trnI(CAU) and rpl2-rps19 gene cluster, although the site of the junction between the IR and single copy regions varies somewhat in the plant species (Zurawski et al., 1984; Sugita et al., 1984). However, the liverwort rps'12 gene cluster and trnI(CAU) gene cluster were located separately in the LSC region near the junctions, and each gene was present as a single copy gene. The promoter sequences of these genes were located in the IR region near the junction $J_{LA/LB}$ (Fig. 2a). Stein et al. (1986) reported that the chloroplast genome of a fern (Osmunda) has relatively short inverted repeats similar to the liverwort chloroplast genome. This may be a feature of the gene organization of lower land plant chloroplast DNA, although the location of the rps'12, rps7, and trnL(CAA) genes is not given.

Genes *ndh5* (Fig. 2b) and *frxC* (Fig. 2d) were located across the junction sites J_{SB} and J_{SA}, respectively. The characteristics of *frxC* and *ndh5* are described below. The junction (J_{SA}) was located at the third codon after the initiation codon of the *frxC* gene. The other junction (J_{SB}) was at the eighteenth codon from the termination codon of *ndh5*. Although the transcribable DNA sequences for each gene are physically separated, that portion of the IR next to the SSC region contained parts of both genes on the opposite strands. Preliminary Northern hybridization analysis indicated the expression of both *frxC* and *ndh5* genes in liverwort chloroplasts (data not shown).

Genes in the SSC region

In the SSC region, there were 19,813 bp that made up 17 ORFs, the *trnL*(UAG) gene, and a *trnP*(GGG)-like sequence. The ORFs have been identified by comparison with a protein database (NBRF release 8 - 10). These genes were classified into five categories of coding sequences for iron-sulfur proteins, proteins homologous to mammalian mitochondrial components of NADH dehydrogenase, ribosomal proteins, other ORFs, and tRNAs. A schematic diagram of the gene organization in the SSC region is given in Fig. 1b to d. The nucleotide sequence and deduced information are shown in Fig. 2b to d. Promoter sequences can be seen upstream from the individual genes *frxC*, *rps15*, *frxA*, *ndh5*, *rpl21*, and *trnL*(UAG).

(i) 4Fe-4S protein genes (frxA, frxB, and frxC)

The *frxA* (98543-98289) gene product would be a hydrophilic polypeptide of 81 amino acid residues, rich in cysteine (9 residues, 11.1% of the total amino acids). This protein can be aligned with bacterial 4Fe-4S-type ferredoxin. The amino acid sequence derived from the *frxA* gene shows local homology with 4Fe-4S proteins found in several microorganisms (Fig. 3a; Howard *et al.*, 1983; Minami *et al.*, 1985b; Tanaka *et*

```
(a) frxA and frxB

frxA          MAHA-VKIYDTCIGCT--QCVRACPTDVLEMIPWD---GCKANQ-IASAPRT-ED-CVGCK-----RCESRCPTDFLSVRVYLGNETTRSMGLSY        81
frxB            (60) F:K::A:E--V:::V::IN-:PVVD:ELKKTI:KK:-LKNYSIDFGV-:IF:G-----N:VEY:::NC::MTEEYELS:YNRHE:N: (46) 183   29.3%
Sulfolobus     (41) DF:L::ADG--S:IT:::VN:FQWYETP---:HP:SEKKADPVNQ-QA-:IF:M-----A:VNV::VAAID:KPP               103   19.8%
Clostridium         :-Y::A:S:VS:G--A:ASE::--:-NA:S-Q---:-DSIFV:-D:----DT-:ID:G-----N:ANV::VGAPVQE              55   23.5%
Chlorobium          :LY-ITEE:TY:G--A:EPE::--:-TA:S-A---:-DDIYV:-D:----NT-:NE:AGLDEQA:VAV::AECIVQG              60   18.5%
Azotobacter         :F:V-T:N::K:KYTD::EV::--:-DCFY-E---:-PNFLV:-HP----DE-:ID:A-----L::PE::AEAIFSEDEVPEDMQEFIQLN (36) 106   23.5%
Typical repeating units  C..C.  .C...CP                        C..C.      .C...CP

(b) frxC

frxC          MKIAVYGKGGIGKSTTSCNISIALARRGKKVLQIGCDPKHDST-FTLTGFLIPTIIDTLQSKDYHYEDVWPEDVIYKGYGRCDCVEAGGPPAGAGCG
F202         MSAARACSMA::::::::::S:L:A:FSLL::R:::::::::::::-:::::R:QE:V::I:KQVHF:P:ELR:::YVTE:FNGVM::::::::::T:::
nifH          AMRQCAI::::::::::TQ:LVA:::EM::::MIV:::::A:::RLI:HSKAQN::ME-MAAEAGTV::LEL:::LKA:::GVK:::S:::EP:V::A
nifH*        MALRQCAI::::::::::TQ:LVA:::EA::::MIV:::::A:::RLI:HSKAQN:VME-MAASAGSG::LEL:::LQI:::GVK:::S:::EP:V::A

             GYVVGETVKLLKELNAFYE-YDIILFDVLGDVVCGGFAAPL-NY-ADYCIIITQNGFDALFAANRIAASVREKARTHPLRLAGLV-GNRTSKR-D-LIDK
             :::::Q::::::QHHLLED-T:VVV:::::::::::::::::-QH-::RAL:V:A:D::SIY:NM::I:A:QA:SVNYKV::::C:-A::SRET-N-EV:R
             :RG:ITAINF:E:EG:YEDDL:FVFY:::::::::::::M:IRENK:QEIY:VCSGEMM:MY:::N:SKGIVKY:NSGSV::G::ICNS:NTD:E:E::IA
             :RG:ITAINF:E:EG:YSDDL:FVFY:::::::::::::M:IRENK:QEIY:VCSGEMM:MY:::N::KGIVKY:HSGSV::G::ICNS:KTD:E:E::MA

             YVEACPMPVLEVLPLIEDIRVSRVKGKTLFEMVELQPSLKYVCDFY-LNIADQILSKPEGIIPKEVPDRELFSLLSDFYLNPVNTVNEKNKPNLIDFMII    289
             :C::ANFKRIAHM:DLDS::R::L:KR:::::DDAEDVVMARAEYIR:AETLWRSTGEP:LT:EPLT::HI:E::GED                          271   44.3%
             LANKLGTQMIHFV:-RDNV-:Q:AEIRRM-TVI:YD:KA:QA-:E:RAL-:RKVVDNKLLV::NPITMDE:EE::ME:GIME:EDESIVG:TAEEV        289   34.6%
             LAAKIGTQMIHFV:-RDNV-:QHAEIRRM-TVI:YD:KA:QA-:E:RAL-:QK::NNKLLV::NPGTMED:EE::ME:GIMEAEDESIVG:AGAEG        290   35.3%
```

Fig. 3. Comparison of amino acid sequences of ORFs identified as iron-sulfur proteins (frxA, frxB, and frxC). (a) Homology of frxA and frxB gene products to bacterial 4Fe-4S ferredoxin proteins (Sulfolobus in Minami et al., 1985b; Clostridium in Tanaka et al., 1966; Chlorobium in Tanaka et al., 1974; Azotobacter in Howard et al., 1983). (b) Homology of the frxC gene product to ORF(F202) of R. capsulata (Hearst et al., 1985), nifH of Az. vinelandii (Hausinger & Howard 1982), and nifH* of Az. chroococcum (Robson et al., 1986).

al., 1966; Tanaka et al., 1974). The sequence Cys-X-X-Cys-X-X-Cys-X-X-X-Cys-Pro, which is a characteristic repeating unit of 4Fe-4S ferredoxin, was found in the amino acid sequence of the frxA gene product (Fig. 3a).

The frxB (100330-99779) product would be 183 amino acid residues long. This product would also contain 9 cysteine residues, as seen in the frxA product. This protein can also be aligned with bacterial 4Fe-4S-type ferredoxin. The amino acid sequence of frxB also shows localized homology to 4Fe-4S proteins in microorganisms (Fig. 3a) and also contains two copies of the typical repeating sequence. To some extent, homology in the amino acid sequences of frxA and frxB was seen, although the frxB gene product had more amino acids residue at the N- and C-termini.

The frxC (110973-110104) gene product would be a polypeptide of 289 amino acid residues. Significant homology with the bacterial nitrogenase component encoded by nifH, for example, in Az. vinelandii (31.5%; Hausinger & Howard, 1982), is observed. Much higher homology (44.3%) was seen with a putative protein (F202) encoded in the R. capsulata photosynthetic gene cluster (Youvan et al., 1984, Hearst et al., 1985). The alignment of these sequences is shown in Fig. 3b. The frxC gene product also had nine cysteine residues, four of which were located in significantly homologous regions of the nifH product or the R. capsulata F202 gene product. These cysteine residues may be important in holding four iron atoms and four sulfur atoms. The sequence Gly-X-X-X-X-Gly-Lys-Ser is located in the N-terminal region of these proteins. This sequence contains amino acid residues conserved in the nucleotide binding site of a variety of ATP-binding proteins (Higgins et al., 1986).

(ii) Liverwort chloroplast genes homologous to human mitochondrial NADH dehydrogenase components encoded in the SSC region (ndh1, ndh4, ndh4L, ndh5, and ndh6)

Human mitochondrial "URFs" have been identified as components of NADH dehydrogenase and have been named "NDs" (Chomyn et al., 1985; 1986). There were five ORFs named ndh1, ndh4, ndh4L, ndh5, and ndh6 homologous to human mitochondrial URFs in the SSC region of the liverwort chloroplast genome. Two related genes (ndh2 and ndh3) encoded in the LSC region are described in Umesono et al. (1988).

The product of ndh1 (102200-100382, interrupted by a 712-bp group II intron) would be a polypeptide of 368 amino acid residues. The amino acid sequence deduced from the ndh1 gene had homology with the mammalian mitochondrial component of NADH dehydrogenase ND1 (URF1). The alignment of liverwort ndh1 and human ND1 (Chomyn et al., 1985) is shown in Fig. 4a. In comparison with ND1, the liverwort ndh1 reading frame may start at the second methionine codon (102089) to produce a polypeptide of 331 amino acid residues of M_r 36,997.3, which coincides to

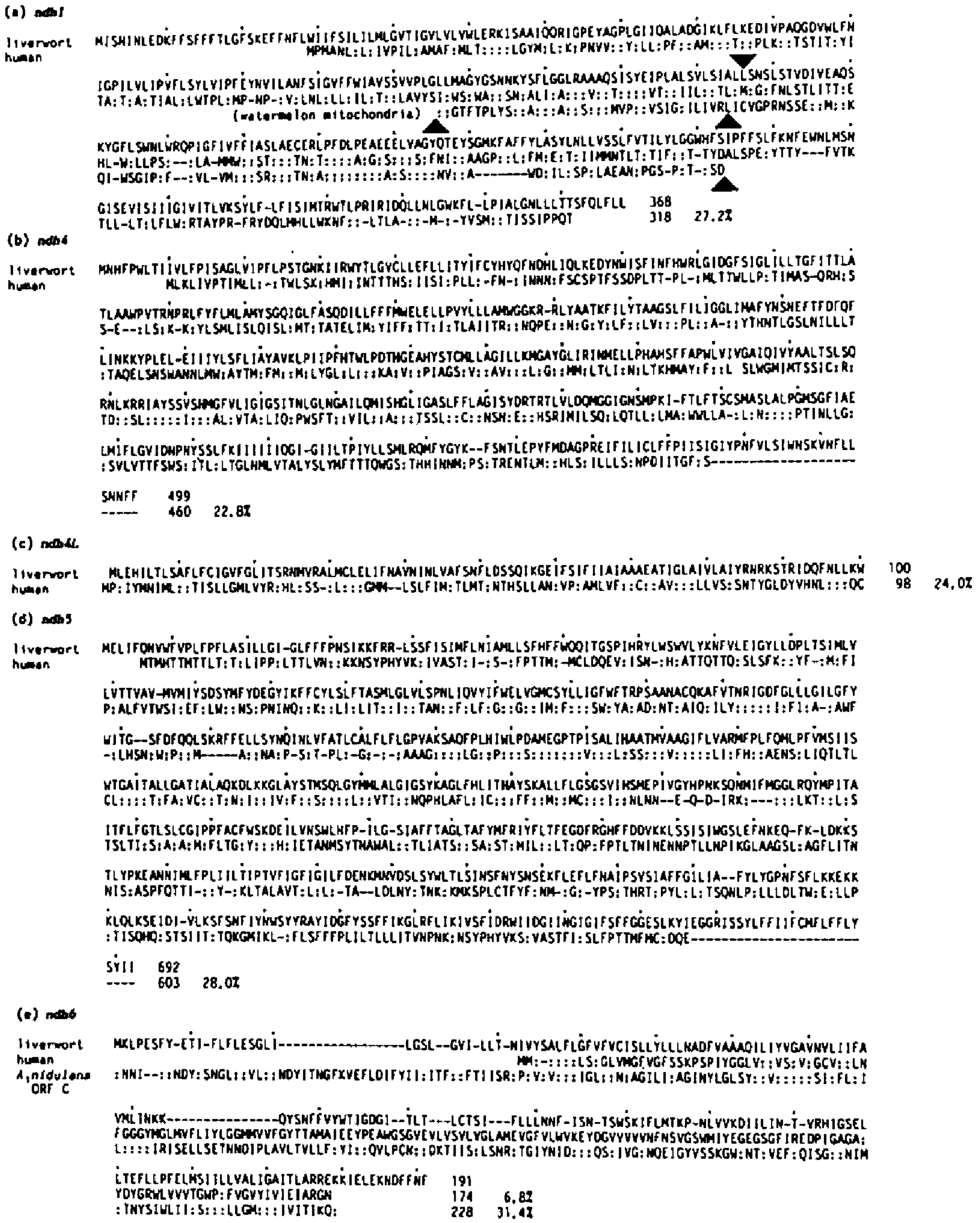

(a) ndh1
liverwort
human
MISHINLEDKFFSFFFTLGFSKEFFHFLWIIFSILILMLGVTIGVLVLVWLERKISAAIQQRIGPEYAGPLGIIQALADGIKLFLKEDIVPAQGDVWLFN
MPMANLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEPLKPATSTITLYI
IGPILVLIPVFLSYLVIPFEYNVILANFSIGVFFWIAVSSVVPLGLLMAGYGSNNKYSFLGGLRAAAQSISYEIPLALSVLSIALLSNSLSTVDIVEAQS
TA:T:A:TIAL:LWTPL:MP-MP-:V:LNL:LL:IL:T::LAVYSI:WS:MA::SM:ALI:A:::V::T::::VT::IIL::TL:M:G:FNLSTLITT:E
(watermelon mitochondria) ::GTFTPLYS::A:::A::S:::MVP::VSIG:ILIVRLICVGPRNSSE::M::K
KYGFLSWNLWRQPIGFIVFFIASLAECERLPFDLPEAEEELVAGYQTEYSGMKFAFFYLASYLNLLVSSLFVTILYLGGWHFSIPFFSLFKNFEWNLMSN
HL-W:LLPS:--:LA-MMW::ST:::TN:T::::A:G:S:::S:FNI::AAGP::L:FM:E:T:IIMMNTLT:TIF::T-TYDALSPE:YTTY---FVTK
QI-WSGIP:F--:VL-VM:::SR:::TN:A::::::::A:S::::NV::A-------WD:IL:SP:LAEAN:PGS-P:T-:SD
GISEVISIIIGIVITLVKSYLF-LFISIMTRWTLPRIRIDQLLNLGWKFL-LPIALGNLLLTTSFQLFLL 368
TLL-LT:LFLW:RTAYPR-FRYDQLMHLLWKNF::-LTLA-::-M-:-YVSM::TISSIPPQT 318 27.2%
(b) ndh4
liverwort
human
MNHFPWLTIIVLFPISAGLVIPFLPSTGNKIIRWYTLGVCLLEFLLITYIFCYHYQFNQHLIQLKEDYNWISFINFHWRLGIDGFSIGLILLTGFITTLA
MLKLIVPTIMLL:-:TWLSK:HMI:INTTTHS:IISI:PLL:-FN-:INNN:FSCSPTFSSDPLTT-PL-:MLTTWLLP:TIMAS-QRH:S
TLAAWPVTRNPRLFYFLMLAMYSGQIGLFASQDILLFFFMWELELLPVYLLLAMWGGKR-RLYAATKFILYTAAGSLFILIGGLIMAFYNSHEFTFDFQF
S-E--:LS:K-K:YLSMLISLQISL:MT:TATELIM:YIFF:TT:I:TLAIITR::NQPE::N:G:Y:LF::LV:::PL::A-::YTHNTLGSLNILLLT
LINKKYPLEL-EIIIYLSFLIAYAVKLPIIPFHTWLPDTHGEAHYSTCMLLAGILLKMGAYGLIRINMELLPHAHSFFAPWLVIVGAIQIVYAALTSLSQ
:TAQELSNSWANNLMW:AYTM:FM::M:LYGL:L:::KA:V::PIAGS:V::AV:::L:G::MM:LTLI:N:LTKHMAY:F::L SLWGMIMTSSIC:R:
RNLKRRIAYSSVSHMGFVLIGIGSITNLGLNGAILQMISHGLIGASLFFLAGISYDRTRTLVLDQMGGIGNSMPKI-FTLFTSCSMASLALPGMSGFIAE
TD::SL:::::I:::AL:VTA:LIQ:PWSFT::VIL::A:::TSSL::C::NSH:E::HSRIMILSQ:LQTLL:LMA:WWLLA-:L:N::::PTINLLG:
LMIFLGVIDNPHYSSLFKIIIIIIQGI-GIILTPIYLLSMLRQMFYGYK--FSNTLEPYFMDAGPREIFILICLFFPIISIGIYPNFVLSIWNSKVHFLL
:SVLVTTFSWS:ITL:LTGLHMLVTALYSLYMFTTTQWGS:THHINNM:PS:TRENTLM::HLS:ILLLS:NPDIITGF:S-----------------
SNNFF 499
----- 460 22.8%
(c) ndh4L
liverwort
human
MLEHILTLSAFLFCIGVFGLITSRNMVRALMCLELIFNAVNINLVAFSNFLDSSQIKGEIFSIFIIAIAAAEATIGLAIVLAIYRNRKSTRIDQFNLLKW 100
MP:IYMNIML::TISLLGMLVYR:HL:SS-:L:::GMM--LSLFIM:TLMT:NTHSLLAN:VP:AMLVF::C::AV:::LLVS:SNTYGLDYVHNL:::QC 98 24.0%
(d) ndh5
liverwort
human
MELIFQNVWFVPLFPFLASILLGI-GLFFFPNSIKKFRR-LSSFISIMFLNIAMLLSFHFFWQQITGSPIHRYLWSWVLYKNFVLEIGYLLDPLTSIMLV
MTMHTTMTTLT:T:LIPP:LTTLVN::KKNSYPHYVK:IVAST:I-:S-:FPTTM:-MCLDQEV:ISN-:H:ATTQTTQ:SLSFK::YF-:M:FI
LVTTVAV-MVMIYSDSYMFYDEGYIKFFCYLSLFTASMLGLVLSPNLIQVYIFWELVGMCSYLLIGFWFTRPSAANACQKAFVTNRIGDFGLLLGILGFY
P:ALFVTWSI:EF:LW::NS:PNINQ::K::LI:LIT::I::TAN::F:LF:G::G::IM:F:::SW:YA:AD:NT:AIQ:ILY:::::I:FI:A-:AWF
WITG--SFDFQQLSKRFFELLSYNQINLVFATLCALFLFLGPVAKSAQFPLHIWLPDAMEGPTPISALIHAATMVAAGIFLVARMFPLFQMLPFVMSIIS
-:LHSN:W:P::M-----A::NA:P-S:T-PL:-G:-:-:AAAG::::LG::P:::S:::::::V:::L:SS:::V:::::LI:FH::AENS:LIQTLTL
WTGAITALLGATIALAQKDLKKGLAYSTMSQLGYMMLALGIGSYKAGLFHLITHAYSKALLFLGSGSVIHSMEPIVGYHPNKSQNMIFMGGLRQYMPITA
CL:::::T:FA:VC::T:N:I:::IV:F::S::::L::VTI::NQPHLAFL::IC:::FF::M::MC:::I::NLNN--E-Q-D-IRK:---::::LKT::L:S
ITFLFGTLSLCGIPPFACFWSKDEILVNSWLHFP-ILG-SIAFFTAGLTAFYMFRIYFLTFEGDFRGHFFDDVKKLSSISIMGSLEFNKEQ-FK-LDKKS
TSLTI:S:A:A:M:FLTG:Y:::H:IETANMSYTNAWAL::TLIATS::SA:ST:MIL::LT:QP:FPTLTNINENNPTLLNPIKGLAAGSL:AGFLITN
TLYPKEANNIMLFPLIILTIPTVFIGFIGILFDENKMNVDSLSYWLTLSINSFNYSNSEKFLEFLFNAIPSVSIAFFGILIA--FYLYGPNFSFLKKEKK
NIS:ASPFQTTI-::Y-:KLTALAVT:L:L:-TA--LDLNY:TNK:KMKSPLCTFYF:NM-:G:-YPS:THRT:PYL:L:TSQNLP:LLLDLTW:E:LLP
KLQLKSEIDI-VLKSFSNFIYNWSYYRAYIDGFYSSFFIKGLRFLIKIVSFIDRWIIDGIINGIGIFSFFGGESLKYIEGGRISSYLFFIIFCMFLFFLY
:TISQMQ:STSIIT:TQKGMIKL-:FLSFFFPLILTLLLITVNPNK:NSYPHYVKS:VASTFI:SLFPTTMFMC:DQE--------------------
SYII 692
---- 603 28.0%
(e) ndh6
liverwort
human
A.nidulans
ORF C
MKLPESFY-ETI-FLFLESGLI----------------LGSL--GVI-LLT-NIVYSALFLGFVFVCISLLYLLLNADFVAAAQILIYVGAVNVLIIFA
MM:-::::LS:GLVMGFVGFSSKPSPIYGGLV::VS:V:GCV::LN
:NNI--::NDY:SNGL::VL::NDYITNGFXVEFLDIFYII:ITF::FTIISR:P:V:V:::IGL::N:AGILI:AGINYLGLSY::V:::::SI:FL:I
VMLINKK--------------QYSNFFVYWTIGDGI--TLT---LCTSI---FLLLNNF-ISN-TSWSKIFLMTKP-NLVVKDIILIN-T-VRHIGSEL
FGGGYMGLMVFLIYLGGMMVVFGYTTAMAIEEYPEAWGSGVEVLVSVLVGLAMEVGFVLWVKEYDGVVVVVNFNSVGSWMIYEGEGSGFIREDPIGAGA:
L::::IRISELLSETNNDIPLAVLTVLLF:YI::QVLPCN::DKTIIS:LSNR:TGIYNID:::QS:IVG:NQEIGYVSSKGW:NT:VEF:QISG::NIM
LTEFLLPFELMSIILLVALIGAITLARREKKIELEKNDFFNF 191
YDYGRWLVVVTGWP:FVGVYIVIEIARGN 174 6.8%
:TNYSIWLII:S:::LLGM:::IVITIKQ: 228 31.4%

Fig. 4. Comparisons of amino acid sequences of ORFs identified as ndh subunits (ndh1, ndh4, ndh4L, ndh5, and ndh6). (a) Homology of ndh1 to ND1 of human mitochondria. Arrowheads indicate the presence of the introns in the genes. (b) Homology of ndh4 to ND4 of human mitochondria. (c) Homology of ndh4L to ND4L of human mitochondria; (d) homology of ndh5 to ND5 of human mitochondria. (e) Homology of ndh6 to ND6 of human mitochondria (Chomyn et al., 1986) and URFC of As. nidulans (Netzker et al., 1986). Invariant histidine residues required for heme-binding as reported for mitochondrial ND4 and ND5 (de la Cruz et al., 1984) are shown by bold letters.

the size of the ND1 product. The ndh4 (98164-96665) product would be 499 amino acid residues long. Homology with mammalian mitochondrial component of NADH dehydrogenase ND4 (URF4) was seen. The alignment of liverwort ndh4 and human mitochondrial ND4 is shown in Fig. 4b. The ndh4L (99059-98757) product would be a polypeptide of 100 amino acid residues. By a search of the protein data-base, homology with the mammalian mitochondrial component of NADH dehydrogenase ND4L (URF4L) was detected. The alignment of liverwort ndh4L and human mitochondrial ND4L is shown in Fig. 4c. The ndh5 (93179-91101) product would be a polypeptide of 692 amino acid residues. Amino acid sequence homology with the mammalian mitochondrial component of NADH dehydrogenase ND5 (URF5) was significant. The alignment of liverwort ndh5 and human mitochondrial ND5 is shown in Fig. 4d. The ndh6 (99688-99113) product would be a polypeptide of 191 amino acid residues. This gene was named URF6 (Ohyama et al., 1986) based on its homology with human mitochondrial URF6, although greater homology can be detected with the As. nidulans mitochondrial ORF C (Netzker et al., 1982). I renamed this URF6 as ndh6 because the human URF6 gene has been shown to encode a component (ND6) of respiratory chain NADH dehydrogenase complex (Chomyn et al., 1986). This protein is aligned with human mitochondrial ND6 (URF6) (6.8% identity; Chomyn et al., 1986) and As. nidulans mitochondrial ORF C (31.4% homology; Netzker et al., 1982) as shown in Fig. 4e. As. nidulans mitochondrial ORF C is related to ND6 ("URF6") of animal mitochondria.

(iii) Ribosomal protein genes (rps15 and rpl21) encoded in the SSC region

The rps15 (103699-103433) product would be a hydrophilic polypeptide of 88 amino acid residues, with an arginine plus lysine content of 21.6%. This protein can be aligned with E. coli ribosomal protein S15 (Takata et al., 1984) as shown in Fig. 5a. The amino acid sequence of the liverwort rps15 gene product shows 35.2% homology to that of E. coli ribosomal protein S15.

```
(a) rpml5 (S15)

liverwort   MSKNLFMDLSSISEKEK-----GSVEFQIFRLTNRVVKLTYHFKKHGKDYSSQRGLWKILGKRKRLLAYLFKTNFVSYENLIIQLGIRGLKKN   88
E.coli      M:LSTEATAKIV::FGRDANDT::T:V:VAL::AQINH:QG::AE:K::HH:R:::LRMVSQ:RK::D::KRKDVAR:TQ::ER::L:R:      90   35.2%

(b) rpl21 (L21)

liverwort   MSKYAIIETGGGQLRVEPGRFYNIRHFVSLTPNELEQNTKILIYRVLMIRQESTIKMGHPWLKGAIVKGRIL-HSCLEKKITIYKMISKKKTRRKL-GHR
E.coli        M::VFQS::K:H::SE:------QT:R:EKLDIATGETVEFAE::::ANGEEV:I:V:FVD:GVI:AEVVA:GRG::-VK:V:F-RRR:HY::QQ:::

            QKSTRFIVDSIFLNGKEI   116
            :WF:DVKITG:SA        110   28.4%
```

Fig. 5. Amino acid alignments of liverwort chloroplast and *E. coli* ribosomal proteins. (a) Ribosomal protein S15 (Takata *et al.*, 1984). (b) Ribosomal protein L21 (Heiland & Wittmann-Liebold, 1979).

The *rpl21* (93469-93819) product would be a hydrophilic polypeptide of 116 amino acid residues, with an arginine plus lysine content of 22.4%. This protein can be aligned with *E. coli* ribosomal protein L21 (Heiland & Wittmann-Liebold, 1979), as shown in Fig. 5b. The homology between the amino acid sequence of the liverwort *rpl21* gene product and *E. coli* ribosomal protein L21 is 28.4%. This ribosomal protein gene has not been detected in the tobacco chloroplast genome (Shinozaki *et al.*, 1986b).

(iv) Other ORFs in the SSC region

There were seven unidentified open reading frames, ORF1068 (108535-105329), ORF465 (110064-108667), ORF464 (105267-103873), ORF392 (103380-102202), ORF320 (95482-96444), ORF288 (94183-95049), and ORF69 (93886-94059), in the SSC region. Two ORFs gave a significant homology score in computer analysis with functionally unknown ORFs reported in other organisms.

The ORF392 product would be a polypeptide of 392 amino acid residues. Significant homology with LtORF3 (46.2%) and LtORF4 (26.3%) encoded in *L. tarentolae* kinetoplast maxicircle DNA (de la Cruz *et al.*, 1984) are shown in Fig. 6a. These two ORF genes can be converted to one ORF gene by an RNA editing mechanism (Simpson, L., personal communication).

```
(a) ORF392

ORF392           MMILTKNKPMIVSMGPHHPSMHGVLRLIVTLDGEDVLDCEPVLGYLHRGMEKIAENRTIVQYLPYVTRW--DYLATMFTEAITVNAPEKLTNI
LtORF 3      FKFYLADYT:NI:RHLYRFTF::Q::AA::::CCLLY:S::FITYIDVII::::::T::LC:YK:VE:C::MXTGLCKCSL
                                                                LtORF 4      SS:Y::RL::VSVVCN:HLLSLCF:YMLRC

             QVPKRASYIRIIMLELSRIASHLLWLGPFMADIGAQTPFFYIFREREMIYDLFESATGMRMMHNYFRIGGVAVDLPYGWIDKCLDFCDYFLPKINEYERL
             CLAI:CAFM:LL:C:FT:CFNG::CCSCMVM:::SLS:MLWS:E::DKLHTF:DLCC:C::HLAFMCLL:LLD:FVF:F::FL:ML:ISC:FVLDL:DL:

             ITNNPIFLKRVEGIGTVTREEAINWGLSGPMLRASGVQWDLRKVDHYECYDELDWKIQWQKEGDSLARYLVRIGEMKESVKIIQQAL-KAIPGGPFENL-
             FVG:RLLYL:LR:LAFFDIFDLCFNSI::CLS:SL:MV::V:LYSC::L:FM:VFDYCFCYLG::FD:LFLRLFD:RM:ILLCK:CFFVGFFVFG:VC:F

             EARRLNQGKNSEWNLFEYQFISKKPSPTFKLPKQEHYVRVEAPKGELGIFLIGDDSVFPWRLKIRSPGFINLQILPQLVKGHKLADIMTILGSIDIIMGE
             DYMYVDVTIETIIS::-:SL-W---C-CI-::GCS-FAN::H::::YSL::CFLYG-:IS::R::CAD:VHICL:DVMMR:FM:H:LVAVI:NV:VVF:S

             VDR
             :::

(b) mbpY
mbpY   (68)  VLSAYGFTFLTALLATIINAIFGLILAWVLVRYE-FPGKKLLDATV-DLPF-ALPTSV-G--GLTLMT-VFND-KGWIKPI-CSWLNI-KIV-F-NP-IG
hisQ         M:YGFSGVI:QGAIV:LEL:LSSVV::-::IGLV-GA:A::SQNR:TG:I:EGYT:LIR:VPD:V::LLI:YGLQIALNVV-TDS:G:DQ:D-I-D:MVA
malF  (276)  F:AIFVW:VVFS:ITVFLTVAV:MV::-C::QW:ALR::AVYRVLL-I::Y-:V:SFI-S--I:IFKG-L::QSF:E:NMMLSALFGV-:PAW:SD:TTA

             -VLLAMI--FVSLPFVVRTIQPVLQNMEEDLEEAAWCL-GASPWTTFWHILFPPLTPSLLTGTTLGFSRALGEYGSIVLIASNIPMKD (49)   288
             -GIITLG--:IYGAYFTE:FRGAFMAVPKGHI:::TAF-:FTHGQ::RR:M::AMMRYA:P:IGNNWQVI:KATALVS:LGLEDVV:A (50)   228
             RTM:IIVNTWLGY:YMMILCMGL:KAIPD::Y::S-AMD::G:FQN:FK:TL:L:IKP:TPLMIAS:AFNFNNFVL:Q:LTNGG:DRL (59)   514
```

Fig. 6. Amino acid sequence comparison of other ORFs. (a) ORF392 to LtORF3 and LtORF4 of *L. tarentolae* kinetoplast maxicircle DNA (de la Cruz *et al*,, 1984). (b) *mbpY* (ORF288) to *hisQ* in *S. typhimurium* (Higgins *et al*., 1982) and *malF* in *E. coli* (Froshauer *et al*., 1984).

The amino acid sequence of ORF288 showed some homology to the inner membrane permease component encoded by *hisQ* in *S. typhimurium* (Higgins *et al*., 1982) or by *malF* in *E. coli* (Froshauer & Beckwith, 1984). The amino acid sequence homology between *hisQ* and *malF* was not very high, but these components seemed to correspond to each other. I tentatively renamed this ORF288 as the *mbpY* gene. Hydropathy analysis of the *mbpY* gene product showed extensive similarity to that of the *malF* gene product in *E. coli* and of the *hisQ* gene product in *S. typhimurium* (Fig. 7), although the number of amino acid residues was variable. In the liverwort chloroplast genome, there is the *mbpX* gene that encodes a very similar protein to bacterial permease components encoded by *hisP* in *S. typhimurium* or by *malK* in *E. coli* (Ohyama *et al*., 1986; Umesono *et al*., 1988). These observations may indicate an association of the *mbpX* and *mbpY* gene products in a chloroplast membrane complex.

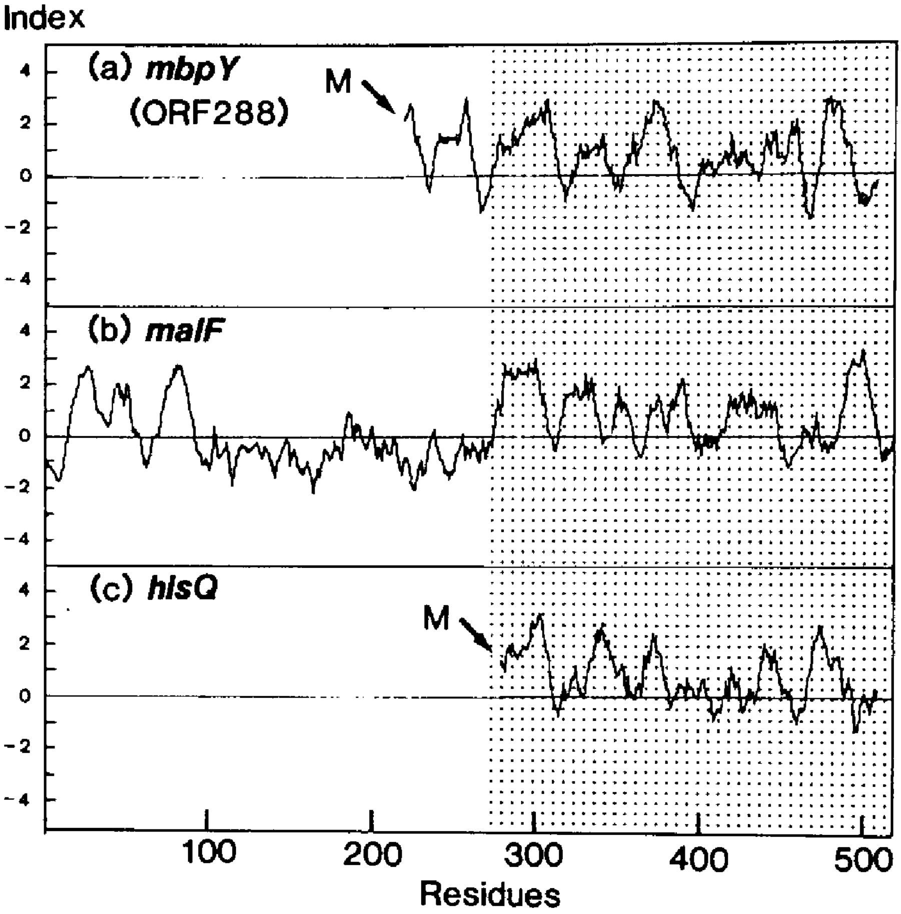


Fig. 7. Hydropathy of the mbpY (ORF288) gene product of the liverwort chloroplast genome. (a) mbpY (ORF288) in liverwort chloroplast. (b) malF in E. coli (Froshauer et al., 1984). (c) hisQ in S. typhimurium (Higgins et al., 1982). The hydropathy is plotted from the N- to C-terminus by the averaging the hydropathy value over a window of 11 residues using the Kyte-Doolittle assignments (1982) and the program DNASIS (HITACHI SK). M denotes the position of the first methionine residue in the reading frames. The region of homology is indicated by the stippled blocks.

The protein of ORF69 was a hydrophilic polypeptide, with an arginine plus lysine content of 26.1%. This ORF could not be identified by computer analysis using the protein database. Chloroplast ribosomal proteins are hydrophilic polypeptides with relatively high levels (more than 20%) of basic amino acids (arginine plus lysine), and more species of ribosomal proteins are reported to be present in chloroplasts than in *E. coli* (Eneas-Filho *et al.*, 1981). ORF69 was located downstream of *rpl21* and can be expected to be co-transcribed with *rpl21*. Further study is required before it can be concluded that ORF69 is the gene for a chloroplast ribosomal protein.

(v) Transfer RNA genes encoded in the SSC region

There were two putative tRNA genes deduced from the DNA sequence in the SSC region. The typical clover-leaf structure could be deduced from the leucine tRNA$_{\mathrm{UAG}}$ gene (95274-95353). This was the only tRNA for CUN-leucine codons and expanded wobble will be required for the codon recognition as seen in mitochondria (Barrell *et al.*, 1980).

A proline tRNA$_{\mathrm{GGG}}$-like sequence (95213-95145) was located on the opposite DNA strand between the *trnL*(UAG) and *mbpY* genes. The secondary structure may be constructed except that both the aminoacyl stem and D-loop are incomplete, although invariant or semi-invariant nucleotides are conserved as shown in Fig. 8a and Table 5 in the preceding chapter. The normal structure of the D-loop can be constructed as shown in Fig. 8b without the formation of the aminoacyl stem. The nucleotide sequence of the reconstructed molecule had homology to liverwort proline tRNA$_{\mathrm{UGG}}$ (Fukuzawa *et al.*, 1988), especially the nucleotide sequences in the loop portion, except for the first nucleotide in the anticodon and a nucleotide in the Tψloop. However, sequence homology of the proline tRNA$_{\mathrm{GGG}}$-like sequence was not detected with proline tRNA$_{\mathrm{GGG}}$ in *Salmonella* (Kuchino *et al.*, 1984) or *Halobacterium* (Gupta, 1984). Another possible structure in Fig. 8c would produce an abnormal tRNA structure with no D-stem as seen in mitochondrial serine tRNA$_{\mathrm{GCU}}$

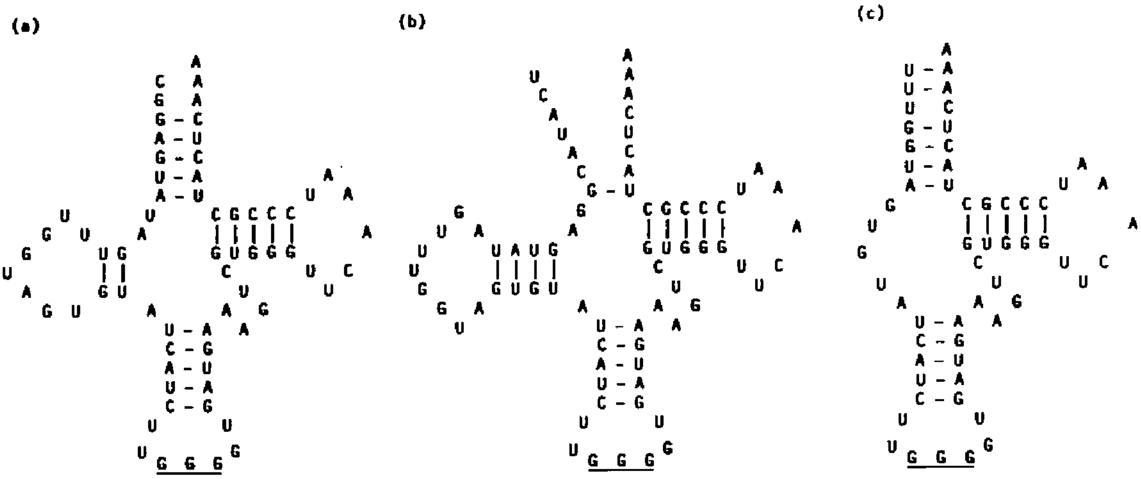


Fig. 8. Secondary structures for proline $tRNA_{GGG}$-like sequence with 2-bp D-stem (a), structure with normal D-loop (b), and structure with no D-loop, like animal mitochondrial serine $tRNA_{AGY}$ (c).

(Arcari & Brownlee, 1982; de Bruijn *et al.*, 1982). None of these three structures is a complete tRNA molecule. No proline $tRNA_{GGG}$ molecules could be detected by preliminary Northern blot hybridization of liverwort chloroplast RNA, indicating the possibility of a pseudogene or very weak expression (data not shown). In the chloroplast genome, the presence of pseudogenes for tRNA has been reported (Howe, 1985; El-Gewely *et al.*, 1984). If the $tRNA_{GGG}$-like sequence is a pseudogene, proline-specific tRNA would be represented by only one species ($tRNA_{UGG}$) in liverwort chloroplasts and expanded wobble might operate to recognize the four proline codons as in mammalian mitochondria (Barrell *et al.*, 1980), and as with the liverwort chloroplast CUN-leucine and GCN-alanine codons described above.

DISCUSSION

Chloroplast IR regions generally contain rrn operons. The liverwort genome has two rrn operons in the IR region that is smallest (10,058 bp) in plant chloroplasts. Chloroplast rrn operons of the fern Osmunda were shown to be located in the IR region (10 kb) of similar size (Stein et al., 1986). Eukaryotic genomes often have multiple rRNA genes (Long & Dawid, 1980) and there are seven rrn operons in E. coli (Lindahl & Zengel, 1982; Fournier & Ozeki, 1985), although some legume chloroplast genomes contain only one rRNA operon. Genes encoded in the higher plant chloroplast IR region such as rpl2, rps'12, rps7, trnL(CAA), and trnI(CAU) were located in the LSC region near the IR region of the liverwort chloroplast genome. No gene located in the higher plant chloroplast IR region had been lost from the liverwort chloroplast genome.

Several interesting features caused by the inverted repeat sequence were observed at the junctions of the inverted repeat and single copy regions. The rps'12 gene and the isoleucine tRNA(CAU) structural gene could be seen in the LSC region immediately next to the IR sequence. The 5' regions of these genes (each in IR_A and IR_B) had common sequences although they are physically separated. The promoter sequence in the IR region must be functional for gene expression both for the rps'12-rps7-ndh2-trnL(CAA) cluster near junction J_{LA} and for the trnI(CAU)-rpl23-rps19-rps22-rps3-rpl16-rpl14-rps8-infA-secX-rps11-rpoA cluster near junction J_{LB}. It may be advantageous to regulate the transcripts of these functionally related genes by the same promoter.

The gene locations at the other ends of the IR region near the junctions (J_{SA} and J_{SB}) with the SSC region were also of interest. The frxC gene may be transcribed from the IR_A region through the junction J_{SA} into the SSC region by a promoter found in the IR_A region. On the other hand, the ndh5 gene may be transcribed from the SSC region through the junction J_{SB}, and transcription may be terminated in the IR_B region by a stem-loop structure. If transcripts for both frxC and ndh5 exist, the nucleotide sequence at the 5'-end of the frxC mRNA molecules would be complementary to that at the 3'-end of the ndh5 mRNA molecules. Translational regulation by complementary RNA is known as mic (mRNA-interfering complementary) RNA (Mizuno et al., 1984). This suggests that frxC mRNA activity may be interfered with the ndh5 mRNA, although the actual gene products and the developmental stage for the expression of each gene are unknown. The promoter sequence for frxC in the IR_A region was also present in the IR_B region. Therefore, anti-mRNA molecules for ndh5 mRNA could be synthesized from the IR_B region. Preliminary Northern hybridization analysis indicated the presence of transcripts of both genes in chloroplasts of liverwort cultured cells (data not shown).

The chloroplast genome encodes all of the ribosomal RNAs and enough kinds of tRNAs to recognize any sense codon. If the chloroplast genome codes for all required RNA species, it may also code for RNA components of certain enzymes such as RNase P for tRNA processing (Stark et al., 1978) and DNA primase for DNA replication (Wong & Clayton, 1986). In fact, RNase P activity has been detected in chloroplasts (Gruissem et al., 1983). The yeast mitochondrial genome has a locus encoding an RNA molecule necessary for tRNA biosynthesis (Miller and Martin, 1983). It is difficult to identify such genes from the nucleotide sequence of the liverwort chloroplast genome because the primary structures of RNA necessary for tRNA synthesis may not be conserved among species. However, a large G + C-rich spacer region was in the IR regions between the trnN(GUU) and the junctions J_{SA} and J_{SB}. The IR region of the liverwort chloroplast genome contains only RNA genes. These results suggest that other specific RNA molecules are encoded in the IR regions.

There were two genes (frxA and frxB) encoding proteins that have features of bacterial 4Fe-4S-type ferredoxin. Ferredoxins generally mediate electron transfer. Chloroplasts contain a 2Fe-2S-type soluble ferredoxin, which is key in photosynthetic electron transport and $NADP^+$-

reduction (Neumann & Drechsler, 1984). This 2Fe-2S-type ferredoxin has also been detected and well-characterized in *M. polymorpha* chloroplasts (Minami *et al.*, 1985a). The complete nucleotide sequence of liverwort chloroplast DNA did not show a coding sequence for a 2Fe-2S-type of ferredoxin protein. There have been a few reports on other iron-sulfur proteins in chloroplasts. An iron-sulfur protein (M_r 8,000) that functions as the primary electron acceptor in photosystem I has been isolated from photosynthetic membranes (Malkin *et al.*, 1974). A spinach photosystem I particle contains an 8-kDa protein that has been tentatively identified as the apoprotein of the iron-sulfur centers of photosystem I (Lagoutte *et al.*, 1984). There are slight differences in the amino acid composition of the liverwort *frxA* product and the 8-kDa cysteine-rich protein in spinach photosystem I. Membrane spanning analysis of the *frxA* product (M_r 8,941.2) showed that it is a peripheral soluble protein (see Table 8 in the preceding chapter). This result is not in agreement with the possibility of the *frxA* protein being an apoprotein of the photosystem I membrane complex. However, the surrounding proteins may provide a special environment so that a soluble protein could be a member of a membrane-bound complex. The chloroplast *frxA* protein was recently found to be a component in the iron-sulfur center of photosystem I, named *psaC* (Oh-oka *et al.*, 1987; Høj *et al.*, 1987). The liverwort chloroplast *frxB* protein probably has properties similar to those of the *frxA* protein except for its size.

Our preliminary Northern hybridization experiments indicate that there was active transcription of the *frxC* genes in liverwort chloroplasts. A second sequence of *nifH* (*nifH**, homologous to *nifH*) has been found in *Az. chroococcum* (Robson *et al.*, 1986). Surprisingly the *nifH** gene was in an operon coding for a bacterial 4Fe-4S-type of ferredoxin protein. Bishop *et al.* (1980) have presented evidence that *Az. vinelandii* contains two N_2-fixing systems, and Premakumar *et al.* (1984) also suggested that a second Fe-protein is involved in the alternative pathway for N_2-fixation. These comprehensive results imply that the liverwort chloroplast genome has retained a gene (frxC) of the alternative pathway for N_2-fixation during chloroplast evolution. However, the R. capsulata F202 ORF homologous to the frxC is located in a photosynthetic gene cluster and therefore frx genes may be involved in electron transfer in photosynthesis.

The ndh1, ndh4, ndh4L, ndh5, and ndh6 genes were encoded in the SSC region of the chloroplast genome. In the ndh4 and ndh5 gene products, the protein domains contained invariant histidine residues required for heme-binding as reported for chloroplast cytochrome b6 (Widger et al., 1984) and mitochondrial ND4 and ND5 (de la Cruz et al., 1984). Although human mitochondrial NADH dehydrogenase (complex I) consists of more than 15 hydrophobic polypeptides, only seven components are encoded in the human mitochondrial genome (Chomyn et al., 1985; 1986). Components of liverwort chloroplast complex are probably encoded in either the chloroplast or the nuclear genome. So far only seven reading frames could be predicted from the nucleotide sequence of the chloroplast genome. A partial nucleotide sequence of a plant mitochondrial NADH dehydrogenase component was reported for watermelon mitochondria (Stern et al., 1986). Local homology of 38.4%, 36.8%, and 35.3% was seen between human mitochondrial ND1 and the liverwort chloroplast ndh1 genes, human mitochondrial ND1 and the watermelon mitochondrial ND1 genes, and liverwort chloroplast ndh1 and the watermelon mitochondrial ND1 genes, respectively (Fig. 9). The divergence between the three species suggests that the chloroplasts ndh genes are not pseudogenes of mitochondrial origin. There have been no reports dealing with these gene products. However, NADPH-plastoquinone oxidoreductase activity has been detected in C. reinhardtii (Neumann & Drechsler, 1984) indicating the possibility that products of these ndh genes are components of the NADH-PQ oxidoreductase in liverwort chloroplasts.

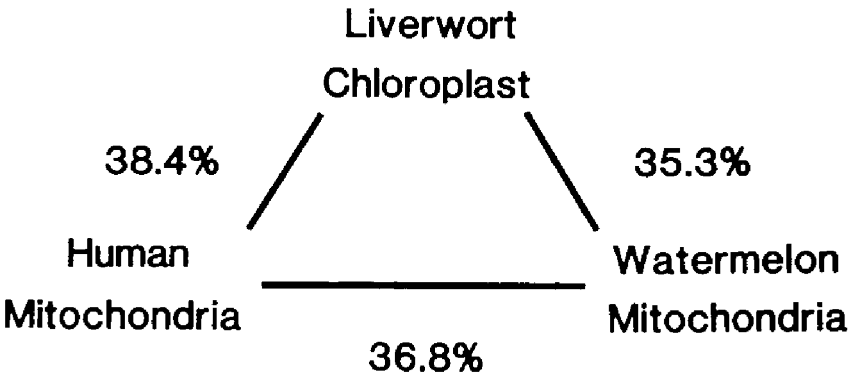


Fig. 9. Scheme of divergence in genes of mammalian mitochondrial ND1, plant mitochondrial ND1, and liverwort chloroplast ndh1. The numbers indicate percentages of amino acid sequence homology between them in the region reported in watermelon mitochondria (Stern *et al.*, 1986).

The chloroplast genome shows some similarity to the human mitochondria genome in the genes encoding components of the transcription-translation system (rRNAs and tRNAs) and the electron-transport system (H^+-ATPase subunits, NADH dehydrogenase components, and cytochrome complex). These results suggest some relationship between chloroplast genomes and mitochondrial genomes. However, mitochondrial genomes contain genes for respiratory polypeptides (COI to III) as the specialized organelles for respiration, and chloroplast genomes contains genes for photosynthetic polypeptides (rbcL, psaA to psaC, psbA to psbG) as the specialized organelles for photosynthesis.

Chapter II Divergent mRNA transcription in the chloroplast *psbB* operon

INTRODUCTION

Chloroplasts have their own genetic system (Whitfeld & Bottomley, 1983; Ohyama *et al.*, 1986; 1988c; Shinozaki *et al.*, 1986b; Umesono & Ozeki, 1987; Ozeki *et al.*, 1987). The entire gene organization of the chloroplast genome has been elucidated for a liverwort, *M. polymorpha* (Ohyama *et al.*, 1986) and for a tobacco, *N. tabacum* (Shinozaki *et al.*, 1986b) by the determination of complete nucleotide sequences. Most functionally related genes are clustered as seen in prokaryotic operons. For instance, the clusters of chloroplast ribosomal protein genes have similar orders to the S10- (Zurawski & Zurawski, 1985), *spc*- (Cerretti *et al.*, 1983), α- (Bedwell *et al.*, 1985), and *str*- (Post & Nomura, 1980) operons in *E. coli.* The H$^+$-ATPase genes in chloroplasts are also clustered as are the *unc* operon in *E. coli* and the gene clusters in cyanobacteria (Cozens *et al.*, 1986). The genes for the components of photosystems and the photoelectron transfer complex, such as *psaA-B*, *psbD-C*, and *psbE-F* in chloroplasts, are also clustered. One typical gene cluster in spinach, *psbB* operon, has been well-characterized (Westhoff *et al.*, 1983), and the gene organization is well conserved in land-plant chloroplasts (Courtice *et al.*, 1985; Ohyama *et al.*, 1986; Shinozaki *et al.*, 1986b; Rock *et al.*, 1987).

Genes on the same DNA strand have been found to overlap in several genomes such as those of coliphages ϕX174 and G4 (Sanger *et al.*, 1977; Godson *et al.*, 1978) and virus SV40 (Fiers *et al.*, 1978). The bovine mitochondrial genes A6L and ATPase-6 also overlap by 40 nucleotides (Fearnley & Walker, 1986). In chloroplasts, except for liverwort (Ohyama *et al.*, 1986) and pea (Zurawski *et al.*, 1986), the stop codon of the gene encoding the β subunit of H$^+$-ATPase, *atpB*, overlaps the initiation codon for the ε subunit, *atpE* (Krebbers *et al.*, 1982; Zurawski *et al.*, 1982). In these genes, the overlapping patterns can be seen either in a single transcript for multiple genes or within two transcripts of the same DNA strand. On the other hand, few genomes have been found to have divergently overlapped genes, which are encoded on both DNA strands in the same region. RNA I and RNA II involved in ColEI plasmid replication have complementary sequences that control the function of the other sequence (Tomizawa *et al.*, 1981; Tomizawa & Itoh, 1981). The O protein gene in the early transcripts from the right promoter in bacteriophage λ overlapped the *oop*RNA coding sequence in the complementary DNA strand (Schwarz *et al.*, 1978). Schwarz *et al.* (1981) have reported the presence of overlapping transcripts in the *in vitro* transcription system of maize chloroplasts. The liverwort ORF43 gene was located on the complementary DNA strand between *psbB* and *psbH* genes in the *psbB* operon. The genes in the *psbB* operon were co-transcribed as a single transcription unit. Recently Tanaka *et al.* (1987) observed a transcription initiation site upstream from *psbB* gene, but did not detect a transcription initiation site in the spacer region between *psbB* and *psbH* genes. However, the ORF43 gene on the complementary DNA strand to the *psbB* operon was actively transcribed in the liverwort chloroplasts *in vivo*. This is the first report to demonstrate divergently overlapping transcription in chloroplasts.

MATERIALS AND METHODS

Isolation of chloroplast RNA

Chloroplast RNA was prepared and purified by repeated Sarkosyl-phenol extraction, LiCl precipitation, and ethanol precipitation from a suspension of cultured cells of a liverwort, *M. polymorpha* (Ohyama *et al.*, 1982; Yamano *et al.*, 1984). Total RNA of the pea *P. sativum* var. Alaska from the seedlings illuminated for a certain time was provided by Dr. Y. Sasaki (Sasaki *et al.*, 1985).

Preparation of DNA fragment

Plasmid pMP710 consisted of *PstI* eleventh fragment of liverwort chloroplast DNA and a cloning vector pUC18 (Ohyama *et al.*, 1988a). Restricted DNA fragments were run on 0.7-1.2% preparative agarose gels or 3.5-8% polyacrylamide gels and eluted from the gels electrophoretically (Maniatis *et al.*, 1982).

Labeling of nucleic acids

For probes of S1 nuclease protection analysis, the 5'-ends of restricted DNA fragments were labelled with [γ-^{32}P] ATP (5000 Ci/mM, Amersham) by polynucleotide kinase (Takara Shuzo). The 3'-ends of DNA fragments were labelled with [α-^{32}P] dCTP (3000 Ci/mM, Amersham) by Klenow fragment of DNA polymerase I (Takara Shuzo). The labelled double-stranded DNA fragments were digested with a restriction endonuclease. Either the 5'- or the 3'-end-labelled DNA fragments were separated electrophoretically. For the hybridization probe, *in vitro* ^{32}P-labelled transcripts from pSP64 vector (Melton *et al.*, 1984) containing appropriate chloroplast DNA fragments were used.

Filter hybridization

Liverwort chloroplast RNA or total pea RNA (10 μg/lane) was dissolved in 20 mM 3-(N-morpholino)propanesulfonic acid (MOPS) buffer (pH 7.0) containing 1 mM EDTA, 2.2 M formaldehyde and 50% formamide, heated at 65°C for 5 min and quenched on ice. Electrophoresis was then done on 1% agarose gel in 20 mM MOPS buffer containing 1 mM EDTA and 2.2 M formaldehyde. Fractionated RNA was blotted to membrane filters (Zeta probe, Bio-Rad Laboratories) by capillary action with 20 X SSC buffer, and fixed by being baked at 80°C for 2 hours *in vacuo*. Prehybridization was done at 45°C for 4 hours in 6 x SSC buffer containing 50% formamide, 0.5% SDS, 0.1% Ficoll, 0.1% polyvinylpyrrolidone, 0.1% BSA, and 200 μg/ml heat-denatured calf thymus DNA. Hybridization was performed at 45°C overnight in the same buffer. The membrane filters were then washed with 2 X SSC buffer containing 0.1% SDS at room temperature, with 1 X SSC buffer containing 0.1% SDS at 65°C, and again at 65°C with 0.1 X SSC buffer containing 0.1% SDS. Autoradiography was done with exposures of from 1 hour to 3 days on X-ray films (RX, Fuji Photo Film Co., Ltd.).

S1 nuclease protection analysis

S1 nuclease protection analysis was involved the procedure of Berk and Sharp (1977) with a slight modification. DNA probes were labelled and purified as described above. For hybridization, 30 μg chloroplast RNA preparations were incubated with labelled probes in 40 mM piperazine-N,N'-bis(2-ethanesulfonic acid) (PIPES) buffer (pH 6.4) containing 80% formamide, 0.4 M NaCl, and 1 mM EDTA for 30 min at 75°C and then for 4 hours at 20°C. The mixture was diluted (1:10) with 30 mM NaOAc buffer (pH 4.6) containing 250 mM NaCl and 4 mM $ZnSO_4$, and incubated with S1 nuclease (1000 units/ml, Takara Shuzo) at 20°C for 30 min. After repetitive phenol extraction and iso-propanol precipitation, the samples were electrophoresed on 6% polyacrylamide gel containing 50% urea together with size markers (^{32}P-labelled *Hpa*II-generated pBR322 DNA fragments) and with chemically cleaved sequence ladders (Maxam & Gilbert, 1977).

In vitro capping analysis of chloroplast RNA *in vivo*

The conditions of the *in vitro* capping reaction were generally as described by Monroy *et al.* (1978) and Strittmatter *et al.* (1985). First, 100 μg of liverwort chloroplast RNA was capped at the triphosphate-bearing 5'-ends in 100 μl of reaction mixture containing 50mM Tris-HCl, pH 7.9, 1.25 mM $MgCl_2$, 6 mM KCl, 2.5 mM DTT, 200 μCi of [α -^{32}P] GTP (400Ci/mM, Amersham), and 8 units guanylyltransferase (Bethesda Research Laboratories) for 120 min at 37°C. The reaction was stopped by the addition of 5 µl of 10% SDS followed by repeated phenol/chloroform extraction. The mixture was precipitated by ethanol with ammonium acetate two times to remove free radioactive GTP.

The hybridization method "Southern Cross" (Potter & Dressler, 1986) was modified. The labelled RNA was separated on agarose gel and blotted on a nylon membrane (Gene Screen, New England Nuclear Corp.) as described above. On the other hand, non-radioactive DNA fragments containing the psbB, ORF43, psbH, and petB/D regions were generated by RsaI digestion of the Pst-eleventh fragments of chloroplast DNA (see Fig. 4), separated on agarose gel and blotted onto Zete-probe membrane in the usual way. The membrane was baked for 2 hours at 80^{o}C and prehybridized with hybridization buffer containing 50% formamide for 16 hours at 45^{o}C. Then the radioactive membrane and non-radioactive membrane sandwich was constructed, heated for 20 min at 80^{o}C, and hybridized for 16 hours at 42^{o}C as described by Potter and Dressler (1986). The recipient membrane was washed for 4 X 5 min at room temperature in 2 X SSC, 0.1% SDS, and then for 2 X 40 min at 48^{o}C in 1 X SSC, 0.1% SDS, after which the membrane was dried and autoradiographed.

RESULTS

Gene organization of the chloroplast psbB gene cluster and ORF43 loci.

The genes psbB, psbH, petB, and petD encode the 51-kDa chlorophyll *a* binding protein (Morris & Herrmann, 1984), the 10 kDa photosystem II phosphoprotein (Westhoff *et al.*, 1986; Hird *et al.*, 1986), cytochrome *b*6, and subunit IV of the cytochrome *b*6/*f* complex (Westhoff *et al.*, 1983), respectively. The order and orientation of these genes are well-conserved among chloroplast genomes in spinach (Westhoff *et al.*, 1983), pea (Courtice *et al.*, 1985), tobacco (Shinozaki *et al.*, 1986b), maize (Rock *et al.*, 1987), and liverwort (Ohyama *et al.*, 1986; Fukuzawa *et al.*, 1988). These genes grouped together have been designated as the psbB operon (Westhoff *et al.*, 1986). No gene has been mapped in the spacer region between psbB and psbH genes; nevertheless, there is a fairly large spacer region. The results of overall DNA sequence analyses of liverwort chloroplasts, such as the G/C content, codon usage pattern (extremely high A/U preference for the third letter in the codons), and secondary structure of nucleotide sequence, suggest that there are two possible reading frames between psbB and psbH (Fig. 1). They are designated as ORF35 and ORF43, and consist of 35 and 43 amino acid

```
                              BstNI.
         .         .         .         .         .         .
CCCATAAGAGGAGTAGTTCCCCAACCAGGAGCCACTTTTCCATACTCTGAATTTAATGGT
 G  M  L  P  T  T  G  W  G  P  A  V  K  G  Y  E  S  N  L  P
         .         .         .         .         .         .
TTTAATATATCACCTATACCACTTTTTTTTCCTTTTGTTTTAGGAGTGTCATCAATTATT
 K  L  I  D  G  I  G  S  K  K  G  K  T  K  P  T  D  D  I  I
                                                    RsaI
         .         .         .         .         .         .
TGTGTAGCCATAAAACTTATCAGATTAGTTGAGATTTATTAACTTTTTGTACTATTATAT
 Q  T  A  M    <psbH        TTGAGA>              TACTAT>
         .         .         .         .         .         .
TAAAAAATATACATTAAAAATATACCATTATTTTGGAACTACCTAACAATGGAAACTGCA
                                  GGA>  ORF43>   M  E  T  A
                               HindIII.
         .         .         .         .         .         .
ACTTTTGTCGCTATCTTCATATCTTGTTTACTTATAAGCTTTACTGGTTATGCTCTTTAT
T  F  V  A  I  F  I  S  C  L  L  I  S  F  T  G  Y  A  L  Y
                               Sau3AI .
         .         .         .         .         .         .
ACCGCATTTGGACAACCTTCTAATGAACTTAGAGATCCATTTGAAGAACATGAAGACTAA
T  A  F  G  Q  P  S  N  E  L  R  D  P  F  E  E  H  E  D  ===
                                                        +---
         .         .         .         .         .         .
TTTTGGACTAATGACTTTTTAAACTAAAAAGTCATTAGTCCAAAATTAGTAATTTAATAT
-------------------> <--------------------------+   +-------
                                                 RsaI.
         .         .         .         .         .         .
TAAATTACTAATTGAATATTAACGTTTTATTTTTTTCCTTTACTTGGTACTTTAGGTGGT
- ----> <---- -------+    ===   K  K  G  K  S  P  V  K  P  P
                              HinfI.
         .         .         .         .         .         .
TCTCTAAAAAAAATAGCAAAAAAAATGATTCCTAAAGTACCTACCAACAAAAATGTATAA
 E  R  F  F  I  A  F  F  I  I  G  L  T  G  V  L  L  F  T  Y
         .         .         .         .         .         .
ACTAATGCTTCCATATATTTGTATATTAGGTAAATAATTTGCAGATAACTTTCGTACTAA
 V  L  A  E  M  <ORF35          +--------     --- ----  --  --
         .         .         .         .         .         .
TTAAATAATATTTTTAGTAAAAAAAAACTGTACTAAAAATTTTATTTATTTAAAAGATAT
- ------ --------->                   <------- -------------- ---
         .         .         .         .         .         .
AAAATATATTTTATATTACTTGTCTTTTTGTTGTTGGATCTCCTAATTTCTGAAACGCTC
 ---- ---+===     I  V  Q  R  K  T  T  P  D  G  L  K  Q  F  A
                                                          <psbB
```

Fig. 1. Nucleotide sequence and information deduced between the psbH and psbB genes. Deduced amino acid sequences are shown under the nucleotide sequence by one-letter symbols. Double underlines indicate termination codons. Possible stem structures are shown by arrows under the sequence. Promoter and ribosome binding sequences are shown under the nucleotide sequence with arrowheads.

residues, respectively. Surprisingly, ORF43 was on the DNA strand opposite to that of the psbB operon. Thus, the nucleotide sequence between the psbB and psbH genes in the liverwort chloroplast genome is efficiently filled with genetic information such as reading frames, stem-loop structures, and promoter elements.

Transcription of the ORF43 gene and psbB operon in liverwort chloroplasts

Systematic Northern hybridization has performed to detect the transcripts of the ORF43 gene and psbB operon. Strand-specific hybridization riboprobes were prepared for psbB (EcoRI-PstI fragment), anti-ORF43 (RsaI-RsaI fragment), and petB/D (RsaI-RsaI fragment) in the psbB operon, and for ORF43 (RsaI-RsaI fragment) encoded on the complementary DNA strand of the psbB operon (Fig. 2). A primary transcript was detected (4.9 kb, lanes 1-3 in Fig. 2 and also see band a in Fig. 4) for the psbB operon. The size of the primary transcript was in good agreement with the results of our S1 protection analyses done to identify the initiation and termination sites of the primary transcript in the psbB operon (unpublished data). The presence of introns in the petB and petD genes (Fukuzawa et al., 1987) and various processing sites gave complicated hybridization patterns in the transcripts of the psbB operon (see also Fig. 4). The hybridization patterns of psbB, petB, and petD in liverwort were similar to those reported for other plant species (Westhoff et al., 1983; Morris & Herrmann, 1984; Rock et al., 1987; Westhoff et al., 1986). On the other hand, Northern hybridization with a strand-specific probe for the ORF43 gene gave a simple hybridization pattern consisting of two bands (230 and 350 nucleotides long; lane 4 in Fig. 2). The DNA region containing ORF43 is unique in the chloroplast genome. It was confirmed by the homology analysis of the entire chloroplast sequence and excluded the possibility that the transcripts for ORF43 do not originate from another loci of the chloroplast genome. These sizes are long enough to encode a polypeptide with 43 amino acid residues.

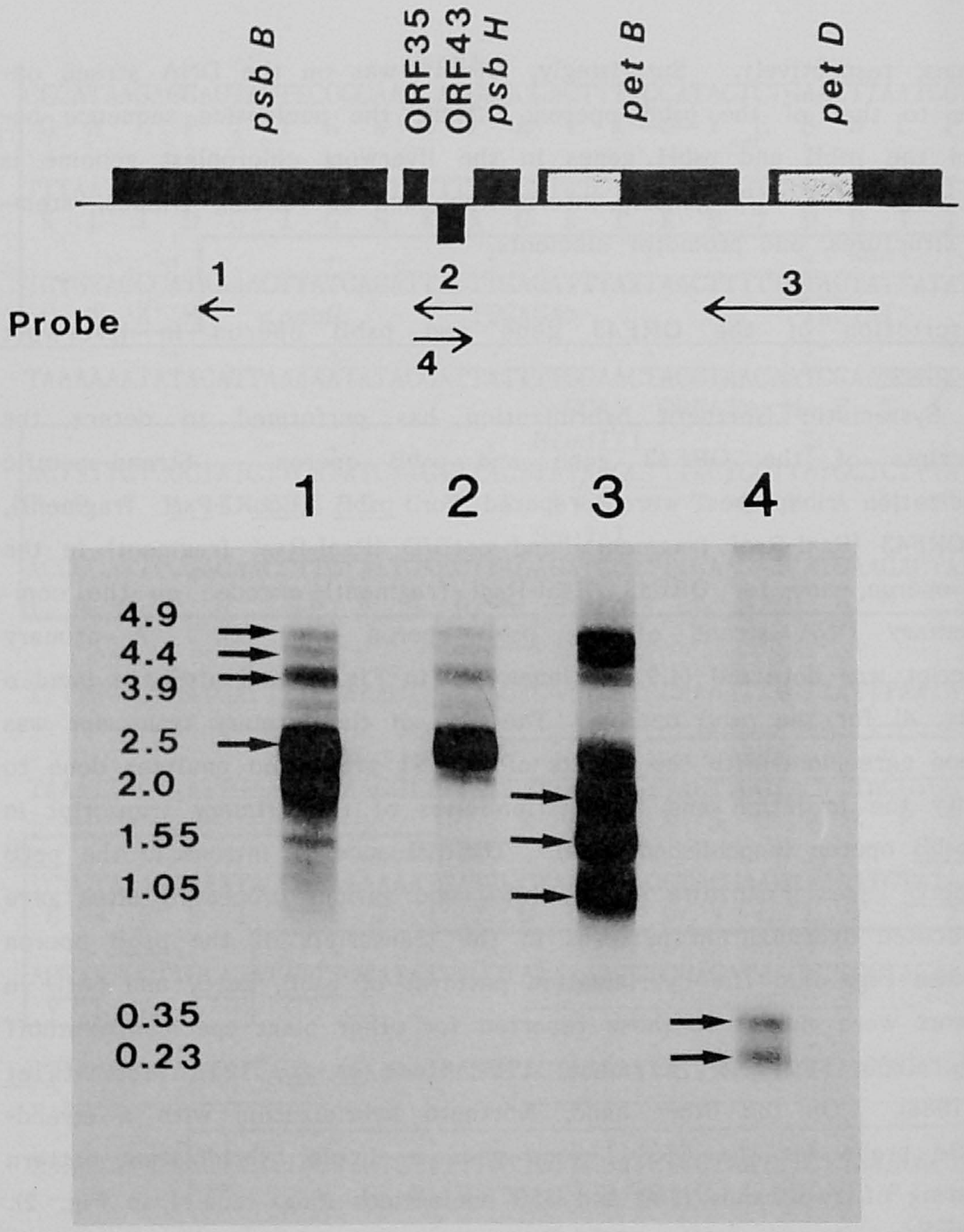


Fig. 2. Northern blot hybridization of the liverwort chloroplast RNA with specific riboprobes. A. Gene organization of liverwort chloroplast psbB operon and ORF43 gene. B. Northern hybridization analysis. Lane 1, psbB probe containing an internal portion in the psbB coding sequence; lane 2, anti-RNA probe of ORF43 containing RsaI-generated fragment (298 bp); lane 3, petB/D probe covering from the petB second exon to the petD second exon; lane 4, ORF43 probe containing RsaI-generated fragment (298 bp) in the opposite direction from the probe in lane 2. Numbers (kb) indicate the size of various transcripts (see also Fig. 4).

S1 protection analysis of ORF43 mRNA showed a single initiation site and two termination sites (Fig. 3). The initiation site was 47 nt upstream from the first AUG codon of ORF43 (panel A in Fig. 3). The termination sites were 50 nt and 170 nt downstream of the termination codon (arrows in panel B, Fig. 3). A signal 50 nt downstream was at the root of a long stem-and-loop structure, and another signal 170 nt downstream was within the ORF35 gene with the opposite orientation. To locate the overlapping region of the ORF43 transcript in processed mRNA for the psbH gene, S1 nuclease protection analysis was done with 5'-end-labelled DNA fragment covering psbH and upstream (panel C in Fig. 3). There were four signals derived from the different lengths of the 5'-leader sequences of the psbH gene. The major nearest signal from the translational initiation codon was about 50 nt upstream (arrow in panel C, Fig. 3) and overlapped divergently by a few nucleotides with the 5'-leader sequences of the ORF43 mRNA.

To locate the primary transcription initiation sites in the psbB operon and ORF43 gene on the complementary DNA strand, *in vitro* chloroplast RNA-capping and hybridization experiments were done (Fig. 4). The primary transcript (4.9 kb: band a) was not observed in the autoradiogram because probably of less efficient transfer of large capped molecules from membrane to membrane. There were two slightly capped molecules corresponding to precursor transcripts; the mRNA in which the first intron in petB gene was spliced (4.4 kb: band b) and the mRNA in which both introns in petB and petD genes were spliced out (3.9 kb: band c). This observation also coincides with that of the ordered splicing of the introns in the maize psbB operon (Rock *et al.*, 1987). Major capped molecules corresponded to the mRNA encoding for psbB to psbH (2.5 kb: band d) and the mRNA encoding for psbB and ORF35 (2.0 kb: band e). These results indicate that rapid processing of the primary transcript occurred at both the 3'-flanking region of ORF35 gene and the 3'- flanking region of psbH gene (Fukuzawa *et al.*, 1987). This fact coincides with presence of monocistronic mRNA for psbH gene (Westhoff *et al.*, 1986).

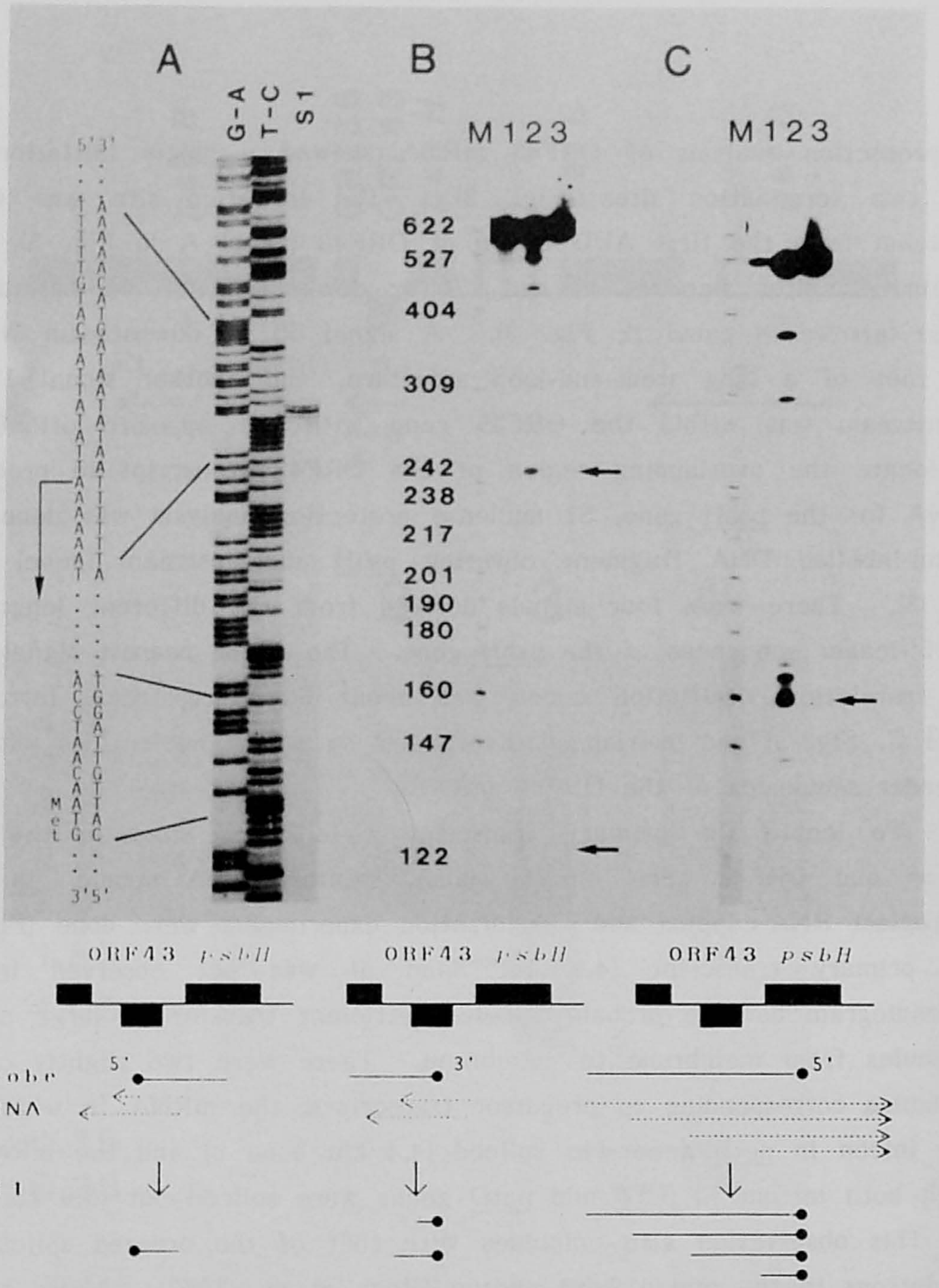


Fig. 3. S1 nuclease protection analysis of ORF43 transcripts and the 5'-terminal of psbH mRNA. (A) 5'-end of ORF43 transcript. The 5'-labeled Sau3AI-BstNI fragment (310 bp) covering ORF43 and psbH was used as a hybridization probe. Lanes G-A and T-C indicate the chemical cleavage ladders specific to the respective bases; lane S1 is for S1 nuclease treatment with chloroplast RNA. The arrow shows the initiation site and the direction of the transcript. (B) 3'-ends of ORF43 transcripts. The 3'-end labeled HindIII-TaqI fragment covering ORF 43 and psbB was used as a probe. Lane M, the size standard of the pBR322-HpaII digest; lane 1, S1 nuclease treatment without chloroplast RNA; lane 2, S1 nuclease treatment with chloroplast RNA; lane 3, probe prepared without S1 nuclease treatment. Arrows indicate major signals. (C) 5'-end of mRNA processed for the psbH. The 5'-end-labeled BstNI-HinfI fragment was used as a probe. Lanes M, 1, 2, and 3 are the same as in panel B. Arrows indicate a signal for major processed psbH mRNA.

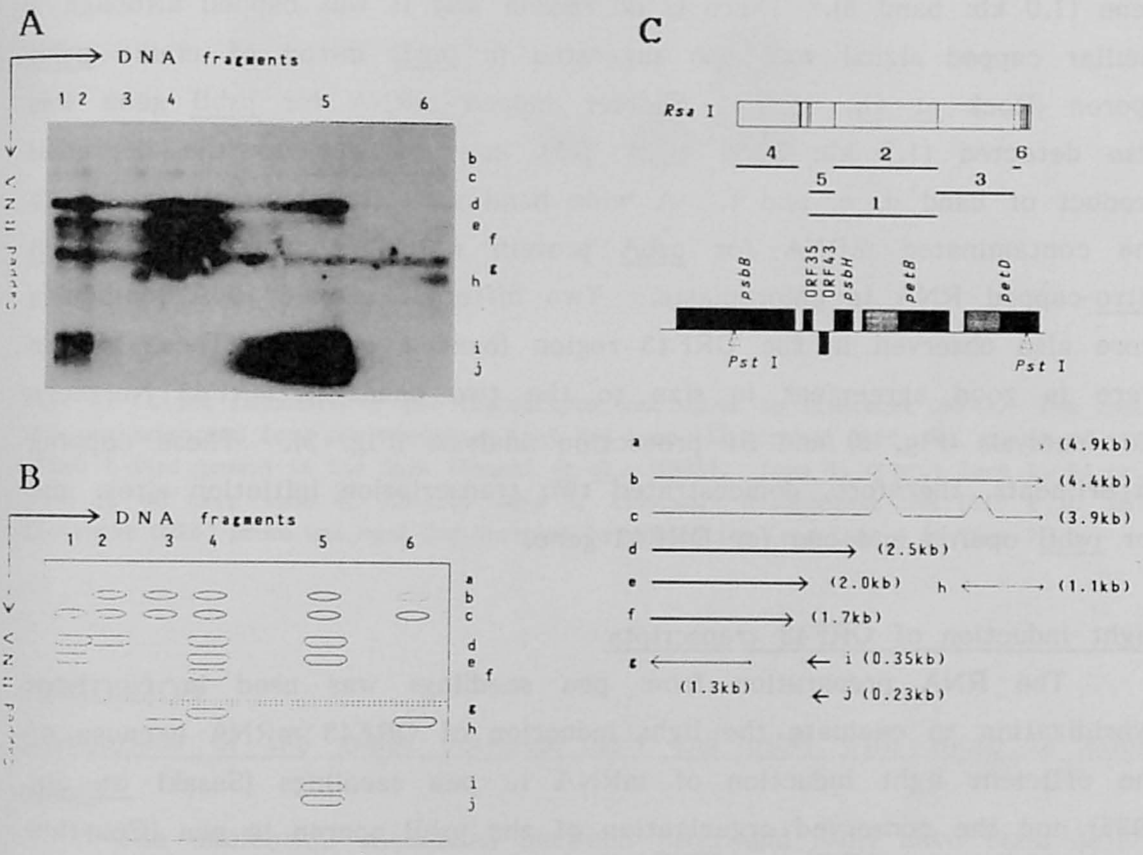


Fig. 4. Southern-cross hybridization of in vitro capped transcripts for psbB operon and ORF43. A. An autoradiogram. DNA fragments were run from left to right in agarose gel and blotted onto membrane. The in vitro capped chloroplast RNA was run from top to bottom and blotted on membrane. B. A schematic presentation of the autoradiogram. C. Physical map, gene organization, and alignment of the detected transcripts. The PstI-eleventh fragment of liverwort chloroplast DNA covering from the psbB gene to the petD gene was digested by RsaI. The resulting fragments except the small fragments are shown. 1, the DNA fragment from the ORF35 gene to psbH gene; 2, from psbH to petB; 3, from petB to petD; 4, psbB; 5, from ORF35 to ORF43; 6, petD. See the text for the explanation of the bands (a to i).

A significant capped molecule (1.7 kb: band f) containing only psbB coding region was also observed. It was not expected that a significant capped molecule appeared in the intron and coding region for the petD gene (1.0 kb: band h). There is no reason why it was capped although a similar capped signal was also suggested in petD intron of maize psbB operon (Rock et al., 1987). Shorter capped mRNA for psbB gene was also detected (1.3 kb: band g). This may be due to the degraded product of band d, e, and f. A wide band with 1.4 kb length is due to the contaminated mRNA for psbA protein which is most abundant in vitro-capped RNA in chloroplasts. Two different capped RNA molecules were also observed in the ORF43 region (band i and j). These signals were in good agreement in size to the two bands in ORF43 Northern blot analysis (Fig. 2) and SI protection analysis (Fig. 3). These capping experiments, therefore, demonstrated two transcription initiation sites: one for psbB operon and one for ORF43 gene.

Light induction of ORF43 transcripts

The RNA preparation from pea seedlings was used in Northern hybridization to evaluate the light induction of ORF43 mRNA because of the efficient light induction of mRNA in pea seedlings (Sasaki et al., 1985) and the conserved organization of the psbB operon in pea (Courtice et al., 1985). A single species of ORF43 mRNA was detected in pea by Northern hybridization with a liverwort ORF43 probe, although liverwort ORF43 mRNAs were composed of two heterogeneous RNAs in terms of their 3'-termini. The amount of ORF43 mRNA that accumulated as the period of illumination increased is shown in Fig. 5. There is no detectable mRNA for ORF43 in pea seedlings grown in the dark. The transcript for ORF43 gene is accumulated under light illumination in chloroplasts.

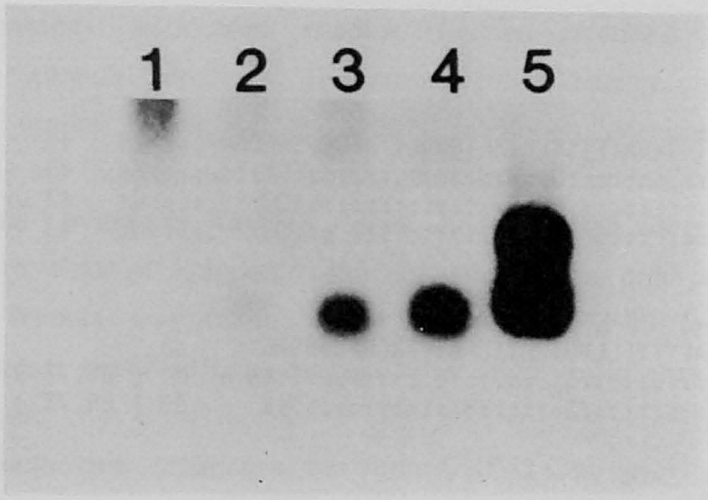


Fig. 5. Light induction of pea transcripts homologous to liverwort ORF43. Pea total RNA was extracted from pea seedlings that had been illuminated different amount of time after 7 days growth in the dark (Sasaki et al., 1985). Lane 1, 0 hr.; lane 2, 24 hr.; lane 3, 48 hr.; lane 4, 72 hr.; lane 5, liverwort chloroplast RNA (0.5×10^{-3}mg). Liverwort ORF43 probe was used for Northern hybridization (see lane 4 in Fig. 3).

DISCUSSION

Comparison of the genes between psbB and psbH with those in other plants

The nucleotide sequences between psbB and psbH have been determined in spinach (Morris & Herrmann, 1984; Westhoff et al., 1986), in tobacco (Shinozaki et al., 1986b), in maize (Rock et al., 1987), and partially in wheat (Hird et al., 1986). The genes homologous to liverwort ORF43 and ORF35 could be deduced from their DNA sequences. The amino acid sequences in liverwort ORF43 and ORF35 are compared to those of the ORFs from other species in Fig. 6. The amino acid sequence of liverwort ORF35 had high homology (85.7%) with that in tobacco and maize, although the reading frame in tobacco is composed of 34 amino acid residues and that in maize is 33 amino acid residues. The amino acid sequence of liverwort ORF43 showed 86.0%, 86.0%, and

```
ORF43

Liverwort        METATFVAIFISCLLISFTGYALYTAFGQPSNELRDPFEEHED   43
Spinach          :::::L::::::G::V::::::::::::::::QQ::::::::G:  43          ( 86.0% )
Tobacco     (MIR)::::::L::::::G::V::::::::::::::::QQ::::::::G:  43 or 46    ( 86.0% )
Maize            :::::L:::S::G::V::::::::::::::::QQ::::::::G:  43          ( 83.7% )

ORF35

Liverwort        MEALVYTFLLVGTLGIIFFAIFFREPPKVPSKGKK   35
Tobacco          :::::::::::S::::::::::::::::::T:KN    34 ( 85.7% )
Maize            :::::::::::S::::::::::::::::::T:K     33 ( 85.7% )
```

Fig. 6. Comparison of amino acid sequences deduced from the nucleotide sequence of liverwort ORF43 and ORF35 with those of the corresponding ORFs from the following species: spinach (Morris and Herrmann, 1984; Westhoff et al., 1986); tobacco (Shinozaki et al., 1986); maize (Rock et al., 1987). The numbers of amino acid residues are shown at the end of the sequences.

```
                                        -35                  -10          ↓
           Met    <psbH
Liverwort  CATAAAA-CT-TATCAG-ATT-AGTTGAGATTTATTAACTTTTTGTACTATTA—TAT-TAAA——A
Maize      ::::::TT::A:T:-:—::-C:-:::-:-:G:-:-G::::G:A::::::::CCG:-:G::GTTGT:
Tobacco    ::::::T-:C-:::T-:T::C:-::::::C:G::G::::G:A:::C::::CCG:-:G::::——T
Spinach    ::::::T-::-G::T-AT::C:-:::::G:C:G::G::::G:A:::C::::CCG:-:G::::——T
Wheat      :::::::T:—:::T-TA::-C:-:::-:-:G:-:-G::::G:A:::C::::CCG:-:G::GTTGT:

                                                                  ORF43>   MetGlu
           AATATAC-AT—TAAAAATATA-CCATTAT—T-TTGGAA—CTACC————————————-TAACAATGGAA
           ::::C::G::-C:TTTC:::G:T:::::G:AA:C:::A::TT:::————————————:::A:::::::
           ::-:—:G::CC::-TC:::G:T:::::GGGG:C:::A::—TTTTATAAATATGA:::G:::::::
           ::-:—TG::TT::-TC:::G:T:::::G:GG:C::AA::TTATC—————————————:::T:::::::
           ::::C::G::-C:TTTCG::G:T::::
```

Fig. 7. Comparison of nucleotide sequences of the 5'-flanking region of ORF43. Liverwort nucleotide sequence upstream ORF43 is compared with that of other plants (Shinozaki et al., 1986; Rock et al., 1987; Westhoff et al., 1986; Hird et al., 1986). Promoter sequences (-10 and -35 regions) are boxed. The arrow indicates the initiation site of ORF43 mRNA in liverwort.

83.7% homology with the reading frames deduced from the DNA sequences in spinach (Morris & Herrmann, 1984), tobacco (Shinozaki et al., 1986b), and maize (Rock et al., 1987), although corresponding frame in tobacco and an additional in-frame methionine codon in the upstream region. The liverwort ORF43 locus had a high degree of sequence homology with that of spinach, tobacco, and maize only when one of the six reading frames was taken. The liverwort ORF35 locus gave similar results. These observations suggest the high reliability of small open reading frames.

The nucleotide sequences between ORF43 and psbH are compared in Fig. 7. The transcription initiation site in liverwort ORF43 was identified by S1 nuclease protection analysis. In vitro capping experiments for the ORF43 gene also showed that there was primary transcription initiation site in the region of the opposite DNA strand of the psbB operon. The promoter elements (-35 and -10 regions) for ORF43 gene were well-conserved among green plants. The stem-and-loop structure could be constructed on the mRNA for ORF43 as shown in Fig. 8A. This structure may corresponding to the termination signal for ORF43 transcription. On the other hand, the precursor mRNA for the psbB operon also form one stem-and-loop structure ($-\Delta G = 23.0$ kcal) between ORF35 and the psbB, two structures ($-\Delta G = 38.3$ kcal; $-\Delta G = 9.24$ kcal) between ORF35 and psbH. These structures may correspond to the processing signals for the primary transcripts.

Possible function of divergent transcripts in an operon

There have been reports describing that gene expression in plastids of higher plants can be efficiently controlled by the RNA stability in post-transcriptional level as a result of environmental changes and/or developmental process (Mullet & Klein, 1987; Deng & Gruissem, 1987). Our experimental results obtained from the liverwort cultured cells grown photomixotrophically under continuous illumination (Katoh, 1983) imply that gene expression in the psbB operon consisting of two deferentially

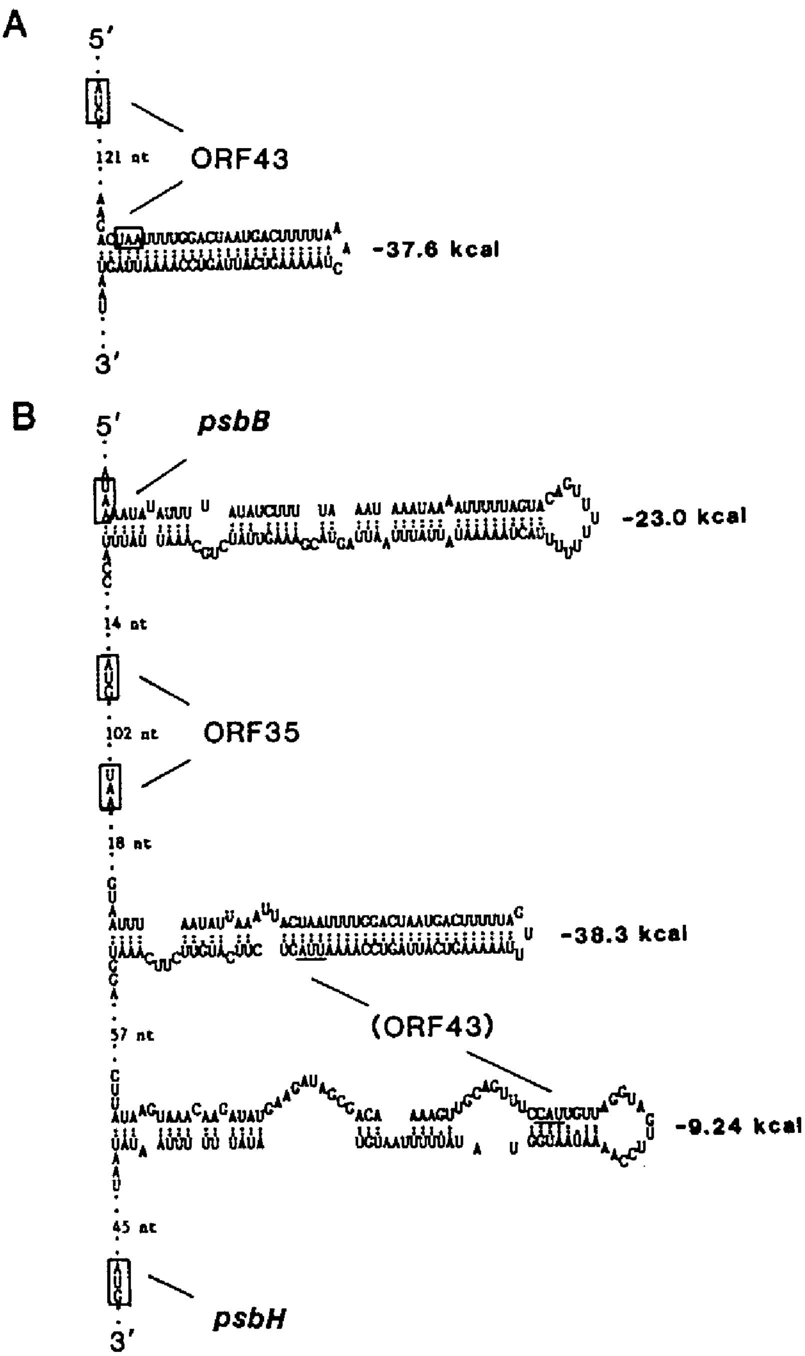


Fig. 8. Possible secondary structures in the spacer regions of ORF43 mRNA (A) and precursor mRNA for psbB operon (B). Base pairings are shown by one dot for G-U pairing and two dots for G-C and A-U pairings. Computer analysis was done by IDEAS-SEQL program (Kanehisa, 1982) with parameters (GMAX, -5.0; LWID, 80; LHMAX, 25; LIMAX, 15; LBMAX, 10; LEN, 100). The stem-and-loop structure with the highest free energy was shown when the overlapped secondary structure can be formed. Numbers indicate free energy (- G).

expressed groups of genes, for components of photosystem II and cytochrome *b6/f* complex, is regulated by the divergently overlapped transcription as well as developmental stages.

Divergently overlapping transcripts within an operon have several effects on gene expression during either transcription or translation. The complementary RNA is a highly specific inhibitor that helps to regulate in plasmid replication (Tomizawa *et al.*, 1981; Tomizawa & Itoh, 1981; Rosen *et al.*, 1981) and bacterial or phage gene expression (Mizuno *et al.*, 1984; Green *et al.*, 1986). In chloroplasts, the expression of the *psbB* gene has been said to be light inducible (Westhoff *et al.*, 1983). In our experiments, the pea chloroplast ORF43 gene was also found to be actively transcribed with the light. On the other hand, mRNA for the *psbH*, *petB*, and *petD* genes has been reported to be present in etioplasts in the dark (Westhoff *et al.*, 1983; Westhoff *et al.*, 1986). *In vitro* capping experiments showed that there was a single transcriptional initiation site upstream from the *psbB* gene in the *psbB* operon (Fig. 4). These observations suggest the presence of controlled mRNA processing or premature transcription termination for gene expression in the *psbB* operon.

There are several possible explanations of the regulation of mRNA processing. RNA is generally transcribed unidirectional from double-stranded DNA by RNA polymeiase. When different messages are encoded on both strands in the same region, RNA polymerase molecules faces each other in simultaneous transcription, resulting in the inhibition of mRNA synthesis. However, this possibility is unlikely because of the high copy number of the genome in chloroplasts. Another possibility is that the ORF43 mRNA acts as a ρ -independent terminator (Yanofsky, 1981) for premature termination in the *psbB* operon because light-inducible transcripts for ORF43 form double-stranded RNA structure with the primary transcripts of the *psbB* operon. The transcription of the following *psbH*, *petB*, and *petD* genes will be then repressed during illumination. Consequently, this explanation coincides with the fact that amounts of transcripts for these genes (*psbH*, *petB* and *petD*) do not change under light illumination. Finally, translational regulation by anti-sense RNA is also one function of messenger RNA interfering complementary (*mic*) RNA (Mizuno *et al.*, 1984). The primary transcripts of *psbB* operon may have the function of micRNA for ORF43 gene expression. The transcripts of the *psbB* operon are processed to make a monocistronic mRNA for the *psbH* gene and two dicistronic mRNA for the *psbB*-ORF35 and *petB*-*petD* genes as translatable mRNA (Westhoff *et al.*, 1986; Rock *et al.*, 1987; Fukuzawa *et al.*, 1987; and this chapter). This fact suggests that the ORF43 mRNA is present as a translatable mRNA, probably because of rapid RNA degradation of antisense RNA. On the other hand, our experimental results that the 5'-end of mature mRNA for *psbH* overlapped the 5'-end of ORF43 mRNA by only a few nucleotides indicate that the ORF43 mRNA may not function as micRNA for monocistronic mRNA of *psbH*. There is no evidence for micRNA regulation in chloroplasts.

CHAPTER III Ordered processing and splicing in a polycistronic transcript in liverwort chloroplasts

INTRODUCTION

In the previous paper, reporting the first case of a trans-split gene in chloroplasts (Fukuzawa et al., 1986), two ORFs, ORF47 and ORF80, were described in the upstream region of rps12'. Later, however, we re-interpreted these ORFs as exons of a split gene that, together with a third ORF, form a gene designated ORF203 (Ohyama et al., 1986). The ORF203 was a unique gene with two cis-introns, 518 and 380 nucleotides long. The second intron is the smallest group II intron in the liverwort chloroplast genome.

In the case of "eukaryotic" (nuclear-encoded) split genes, the processing order of transcripts has been characterized in detail (Padgett et al., 1986). The 5'-end of the precursor RNA is immediately modified by a capping enzyme after the initiation of transcription and the 3'-end of the RNA is generated by poly(A) addition after transcription is terminated. These processing events commonly precede RNA splicing, although poly(A) addition and RNA splicing are not mechanistically coupled (Zeevi et al., 1981). On the other hand, nothing is known about the processing and splicing of multiple introns in the chloroplast genes of land plants. The order of processing of polycistronic precursor mRNA molecules, including the splicing of multiple introns, in the ORF203-rps12'-rpl20 gene cluster transcripts of M. polymorpha chloroplasts.

MATERIALS AND METHODS

Preparation of recombinant plasmid DNA

Plasmid pMP727 consists of a DNA fragment (PstI fragment 6 in Fig. 1A) from liverwort chloroplast DNA and pBR322. Plasmid DNA was prepared as described by Maniatis et al., (1982). DNA restriction fragments were recovered by electrophoresis on an agarose gel with a low melting point (BRL).

Preparation of liverwort chloroplast RNA

Liverwort chloroplasts were isolated either from a two-week-old suspension of cultured cells or from one-day-old exponentially growing cells in fresh medium (Ohyama et al., 1982). Chloroplast RNA was extracted and purified by repetitive phenol extraction and ethanol precipitation as described by Yamano et al., (1984). For primer extension analysis, chloroplast DNA in the RNA preparation was removed by repetitive LiCl precipitation (final concentration of LiCl was 2 M, at 4°C) and ethanol precipitation.

Northern blot hybridization analysis

Oligodeoxyribonucleotides (15-mer) were used as specific probes to each exon and intron. Splicing probe for spliced RNA molecules consisted of 8 nucleotides complementary to the 3' end of an exon and 8 nucleotides complementary to the 5' end of the next exon. The probes were synthesized with a Shimadzu NS-1 DNA synthesis apparatus and labelled with [γ-^{32}P]ATP (Amersham International, Inc., 5000 Ci/mmol) using T4 polynucleotide kinase (Takara Shuzo Co., Ltd.). Each probe was purified by 15% polyacrylamide gel electrophoresis. The locations and nucleotide sequences of the probes are shown in Fig. 2. Liverwort chloroplast RNA (about 10 µg/lane) with chloroplast DNA was heat-denatured, fractionated by electrophoresis in a 1% agarose gel containing 6% formaldehyde, and transferred to a nylon filter (Kohchi et al., 1988b). Prehybridization was done at 65°C for 4 hours in 6 X SSC buffer containing 0.5% SDS, 0.1% Ficoll, 0.1% polyvinylpyrrolidone, 0.1% BSA, and 200 µg/ml heat-denatured calf thymus DNA. Hybridization with individual ^{32}P-labelled probes was performed at 42°C overnight in the same buffer. The membrane filters were washed with 2 X SSC buffer containing 0.1% SDS at room temperature and finally with 1 X SSC buffer containing 0.1% SDS at 42°C for 30 min. Autoradiography was carried out by 1- to 3-day exposures of X-ray film (RX type, Fuji Photo Film Co., Ltd.).

S1 nuclease protection analysis

S1 nuclease protection analysis was done by a slight modification of the procedure of Berk & Sharp (1977). DNA probes were labelled and purified as described above. For hybridization, the chloroplast RNA (30 µg) was incubated with ^{32}P-labelled probes and S1 nuclease was added (1000 units/ml) as described by Kohchi et al., (1988b). The samples were electrophoresed on a 6% polyacrylamide gel containing 50% urea with size markers (^{32}P-labelled HpaII-digested pBR322 DNA fragments) and with chemically cleaved sequence ladders (Maxam & Gilbert 1980).

Primer extension and cDNA sequencing

The primer extension and cDNA sequencing experiments were carried out with 5'-end-labelled oligonucleotide as a primer for 30 µ g of chloroplast RNA by a procedure described before (Williams & Mason 1985, Fukuzawa et al., 1987).

RESULTS AND DISCUSSION

Gene Organization

The gene organization and the nucleotide sequence in the ORF203 region are shown in Figs. 1 and 2, respectively. The coding sequence for

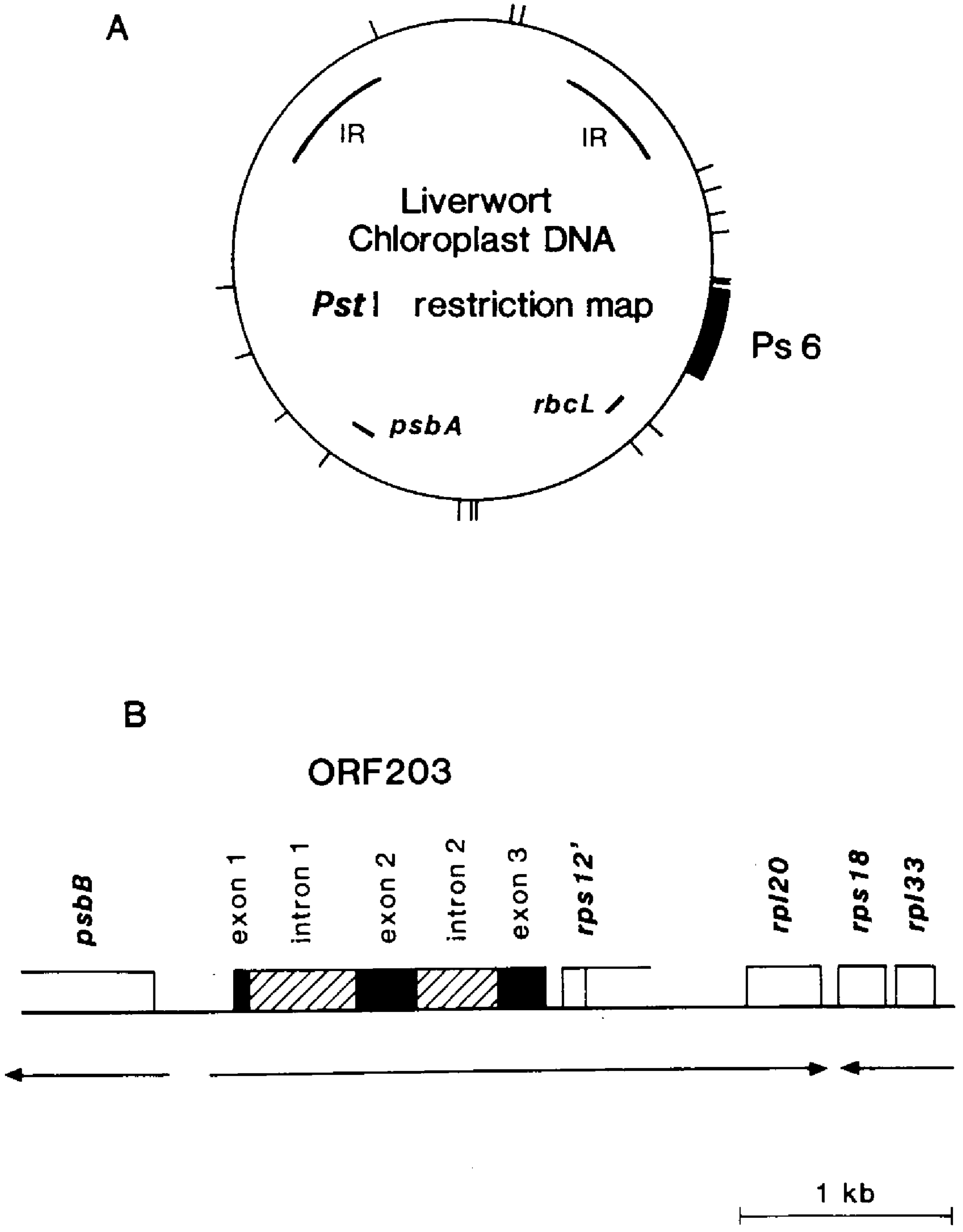


Fig. 1. A, Location of ORF203 gene on the PstI restriction map of a liverwort chloroplast DNA. IR shows an inverted repeat. Genes psbA and rbcL are shown as landmarks. A PstI fragment (Ps6) contains the ORF203 gene. B, The organization of the ORF203 gene cluster in liverwort chloroplast genome. The arrows indicate the direction of transcription. The hatched boxes indicate the regions of introns. ORF203, open reading frame consisting of 203 amino acid residues; rps12', ribosomal protein S12 first exon; rpl20, ribosomal protein L20; rps18, ribosomal protein S18; rpl33, ribosomal protein L33; psbB, 51-kDa chlorophyll a apoprotein in the photosystem II.

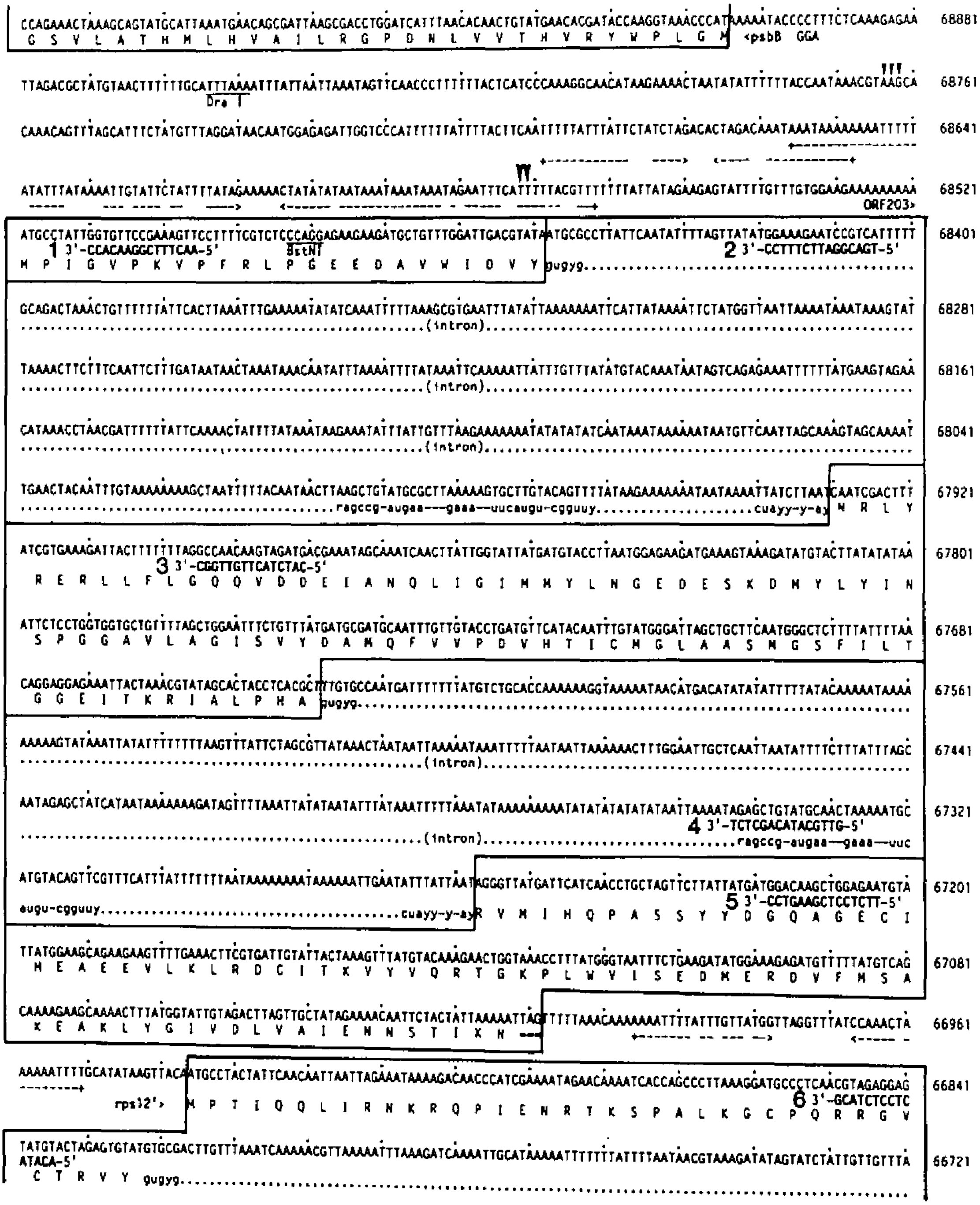


Fig. 2. Nucleotide sequence of ORF203 region. Nucleotide and amino acid sequences deduced for the ORF203 gene are shown. Stem-loop structures are shown by underlining with arrows. The consensus sequence of group II introns in the liverwort chloroplasts is shown by small letters under the DNA sequence. The vertical arrowheads indicate the sites of S1 nuclease protection analysis. The probes and primers are under the recognition sequence for each oligodeoxyribonucleotide sequence.

ORF203 is followed by the first exon of ribosomal protein S12 (*rps12'*) and the ribosomal protein L20 gene (*rpl20*). On the opposite DNA strand, upstream from the gene cluster, there is a gene (*psbB*) for the photosystem II 51-kDa polypeptide. Downstream from the ribosomal protein L20 gene (*rpl20*), a gene cluster of the ribosomal proteins L33 (*rpl33*) and S18 (*rps18*) is present, as on the opposite strand.

RNA processing

To examine the RNA processing and splicing of an ORF203 transcript, Northern blot hybridization was done with radioactively labelled synthetic oligodeoxyribonucleotides as probes. All probes specific to nucleotide sequences for ORF203 and *rps12'* (exon 1) hybridized to the primary precursor RNA molecules (3.00 kb, P1 in Fig. 3), which were long enough to cover the coding sequences from the ORF203 first exon (exon 1) to the *rpl20* genes. These results indicate that the ORF203 gene was co-transcribed with the genes downstream. An RNA band (1.55 kb, a in Fig. 3) was detected by the probes of the ORF203 exons (exons 1, 2, and 3) and by those of the introns (introns 1 and 2). A second RNA band (1.05 kb, b in Fig. 3) was detected by the probes of each exon (exons 1, 2, and 3) and by intron 2 of ORF203, but not intron 1. A third RNA band (0.70 kb, m in Fig. 3) was detected only by the probes of the exons (exons 1, 2, and 3) of ORF203. These results indicate that the ORF203 gene was interrupted by two introns and that the RNA band (0.70 kb) was the mature mRNA for the ORF203 gene. The hybridization pattern of the *rps12'* probe was quite different from the patterns of ORF203 probes, indicating the processing between ORF203 and *rps12'*. The processing point between ORF203 and *rps12'* was detected by S1 nuclease mapping (Kohchi *et al.*, 1988d).

To detect spliced RNA molecules directly, Northern hybridization was performed using two kinds of probes made of oligodeoxyribonucleotides complementary to the spliced RNA junction (Fig. 4). One splicing probe (exon 1 connected to exon 2) hybridized to both

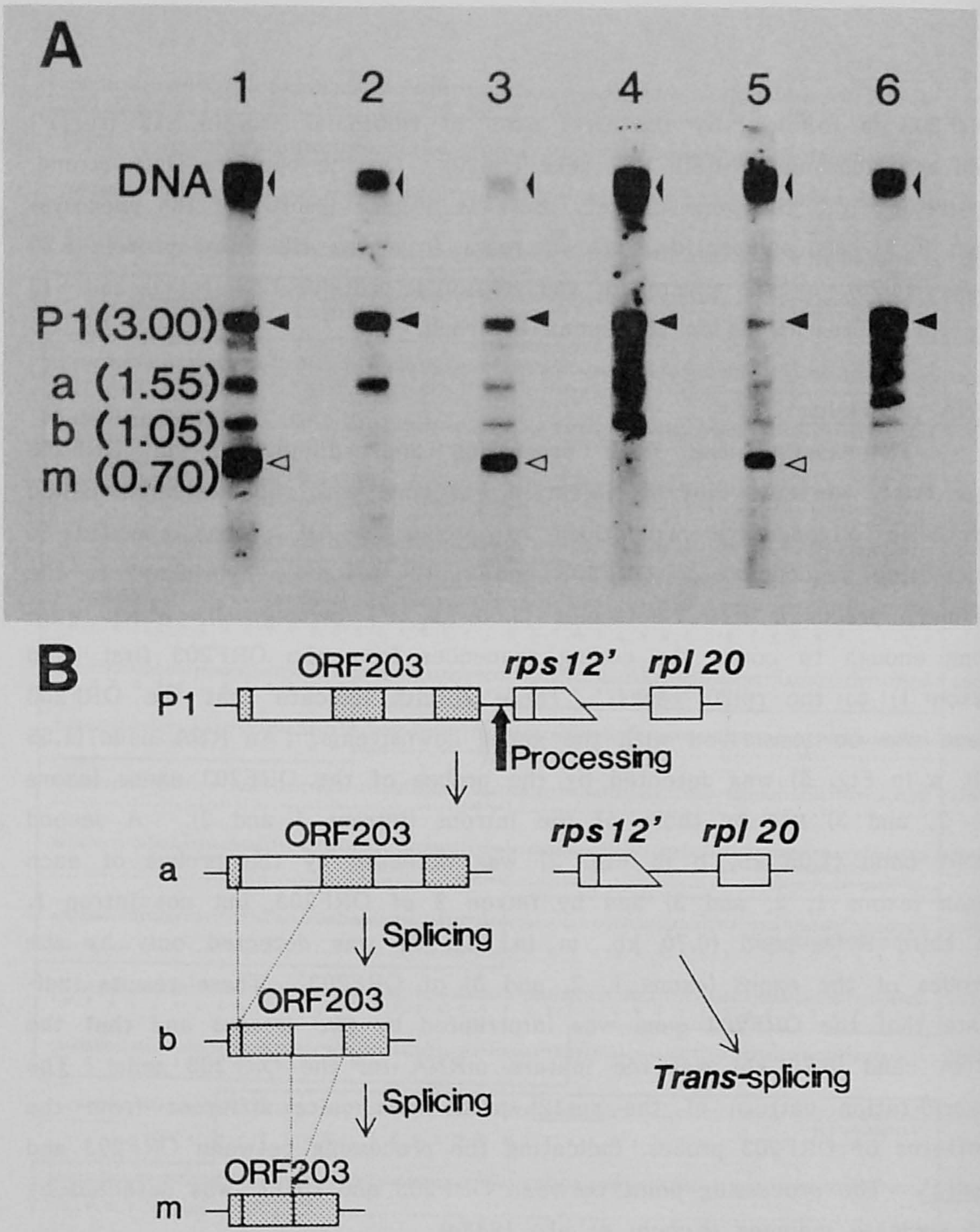


Fig. 3. Northern blot hybridization to chloroplast RNA with probes specific to each region. A. lane 1, ORF203 first exon (exon 1); lane 2, ORF203 first intron (intron 1); lane 3, ORF203 second exon (exon 2); lane 4, ORF203 second intron (intron 2); lane 5, ORF203 third exon (exon 3); lane 6, rps12' (S12 first exon). P1 with an arrowhead indicates precursor mRNA, m with an open arrowhead shows mature mRNA for ORF203. The position and sequence of each probe are shown in Fig. 2. The sizes, in kilobases, are calculated from the size markers (BRL RNA ladders) in gel stained with ethidium bromide. B. Main scheme of ordered processing of ORF203 mRNA. The thick arrow indicates the intercistronic processing site. Dotted boxes indicate the exons of ORF203 transcripts.

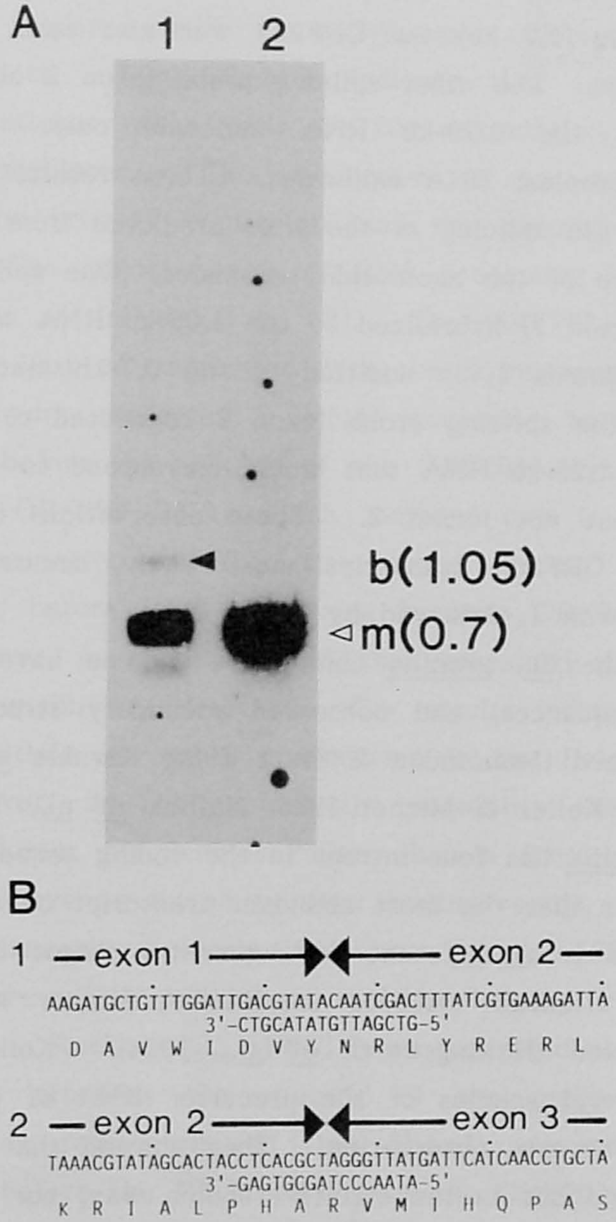

A
1
2
b(1.05)
m(0.7)
B
1 — exon 1 exon 2 —
AAGATGCTGTTTGGATTGACGTATACAATCGACTTTATCGTGAAAGATTA
3'-CTGCATATGTTAGCTG-5'
D A V W I D V Y N R L Y R E R L
2 — exon 2 exon 3 —
TAAACGTATAGCACTACCTCACGCTAGGGTTATGATTCATCAACCTGCTA
3'-GAGTGCGATCCCAATA-5'
K R I A L P H A R V M I H Q P A S

Fig. 4. Northern blot hybridization to chloroplast RNA with specific probes (splicing probes). A. Lane 1, oligodeoxyribonucleotide (16-mer) probe composed of 8 nucleotides complementary to the 5'-end of the ORF203 second exon and 8 nucleotides complementary to the 3'-end of the ORF203 first exon; lane 2, oligodeoxyribonucleotide (16-mer) probe composed of 8 nucleotides complementary to the 5'-end of the ORF203 third exon and 8 nucleotides complementary to the 3'-end of the ORF203 second exon. B. Sequences of each probe are shown.

the 1.05-kb and 0.70-kb RNA molecules. A smaller RNA molecule than the mature mRNA (0.7 kb) for ORF203 was also seen. This molecule may be an artifact. The other splicing probe (exon 2 connected to exon 3) hybridized to the 0.70-kb RNA molecule only. Neither probe hybridized to chloroplast DNA molecules. These results indicate the occurrence of accurate splicing at the sites predicted from the primary and secondary structure of the nucleotide sequences. The splicing probe (exon 1 connected to exon 2) hybridized to the 1.05-kb RNA which contains intron 2, but not intron 1, in addition to the 0.7-kb mature mRNA. On the other hand, the splicing probe (exon 2 connected to exon 3) did not hybridize to any 1.25-kb RNA that would correspond to a precursor RNA with intron 1, but not intron 2. These observations indicate that the splicing of the ORF203 transcript most likely occurred in sequence, beginning with intron 1, followed by intron 2.

Genes in the Eu. gracilis chloroplast genome have multiple introns in the coding sequences, and conserved secondary structures on introns have been described (Montandon & Stutz 1983; Karabin et al., 1984; Koller et al., 1984; Keller & Michel 1985; Hallick et al., 1985). The gene psbA of Eu. gracilis has four introns in the coding sequence. The Northern analysis shows that the most abundant transcript of the psbA gene in RNA preparations from different developmental stages of Eu. gracilis is mature mRNA, although multiple unspliced precursor mRNA transcripts have been detected (Hollingsworth et al., 1984). Koller et al., (1985) detected the several species of the precursor RNA of Eu. gracilis psbA by electron microscopic observation. They showed that the four introns are neither spliced out in the strictly random way, nor in a 5'-3' or 3'-5' direction. However, introns found in Eu. gracilis were often smaller than group II introns of land-plants.

I have observed different ratios of mature RNA to precursor RNA in different stages of cultured cells. For example, chloroplast RNA prepared from the stationary stage of cultured cells accumulated higher proportion of precursor mRNA than of the mature mRNA (data not shown). The major portion of chloroplast RNA was then mature mRNA in chloroplasts prepared from 1 to 2-day-old cultured cells transferred into fresh medium. Such changes indicate that the variety of RNA species from stationary-stage cells are not degradation products but RNA processing intermediates. This facilitates us to investigate both the processing of the primary transcripts and the splicing of introns in the ORF203 gene. The major scheme of RNA processing including the splicing reaction was as follows (Fig. 3B). The ORF203 gene including its two introns was first co-transcribed with the downstream *rps12'* and *rpl20* genes as a primary transcript. RNA processing between ORF203 and *rps12'* occurred before the splicing of the ORF203 introns was completed. The monomeric ORF203 RNA, which included two introns, was spliced, deleting the first intron and then the second intron. A trace of RNA that was spliced before RNA processing between ORF203 and *rps12'* was detected. However, no trace of RNA that contained the first intron but not the second intron was detected. Our results indicate that the splicing order of multiple introns is consecutive in the liverwort chloroplasts. Our data, however, do not exclude the possibility that the introns are spliced with different efficiencies rather than a fixed order. The *petB* and *petD* genes have an intron in their coding region, and the mRNA is present as a dicistronic molecule (Fukuzawa *et al.*, 1987). This is also an example of pre-mRNA splicing of multiple introns in one transcript of chloroplast genes. Rock *et al.*, (1987) have suggested that the splicing of the *petB* intron precedes the splicing of the *petD* intron in maize chloroplasts. Our results showed that 3'-end cleavage of the ORF203 mRNA precedes before the RNA splicing of the ORF203 is completed, although a small fraction of the primary transcripts was also spliced. The 5'-ends of multiple *rps12'* transcripts observed in the tobacco chloroplasts (Koller *et al.*, 1987; Hildebrand *et al.*, 1988) may correspond to intercistronic processing products as describes here.

Transcription Initiation

To identify the initiation site of transcription for an ORF203 gene cluster, S1 nuclease mapping was done in the region between the ORF203 and psbB genes. As a probe, 5'-end-labelled DNA fragments covering the spacer region between the ORF203 and psbB genes were hybridized with chloroplast RNA prepared from exponentially growing cultured cells. The broad multiple signal (P1) was seen at about 52 nt upstream from the first methionine codon of ORF203 (Figs. 5A and 5B). A faint single band (arrowhead, P0) was also seen at around 240 nt upstream from the first codon (Fig. 5A). Primer extension experiments were done with a ^{32}P-labelled synthetic primer (AACTTTCGGAACACC, corresponding to exon 1 of ORF203). There was one major signal (P1) from the 5'-end of a transcript that exactly corresponded to 52 nt upstream from the first methionine codon obtained by S1 nuclease protection analysis (arrowhead, Fig. 5C). Therefore, the ORF203 gene cluster was co-transcribed polycistronically with the genes rps12' (exon 1) and rpl20 by the promoter found at around 240 or 52 nt upstream from the ORF203 coding region. The 52-nt upstream band was probably a major initiation site of transcription for the ORF203 gene cluster, although the 52-nt leader sequences may have been the result of processing from the 240-nt leader sequence.

Secondary structures of the introns

The consensus boundary sequences of liverwort chloroplast introns of group II are 5'-GUGPyG ...... CUAPyPyNPyAPy-3' (Ohyama et al., 1986). However, the 5'-end sequence of the ORF203 first intron (intron 1) is AUGCG, and that of the ORF203 second intron (intron 2) is UUGUG. The 3'-end sequences of ORF203 introns also differ from the 3'-consensus sequences found in other introns. Group II introns generally have the configuration of 6 major stem-and-loop structures (Michel & Dujon 1983). Intron 2, especially the stem I portion, is rich in A + U and is the smallest intron in the liverwort chloroplast genome. Nevertheless, the in-

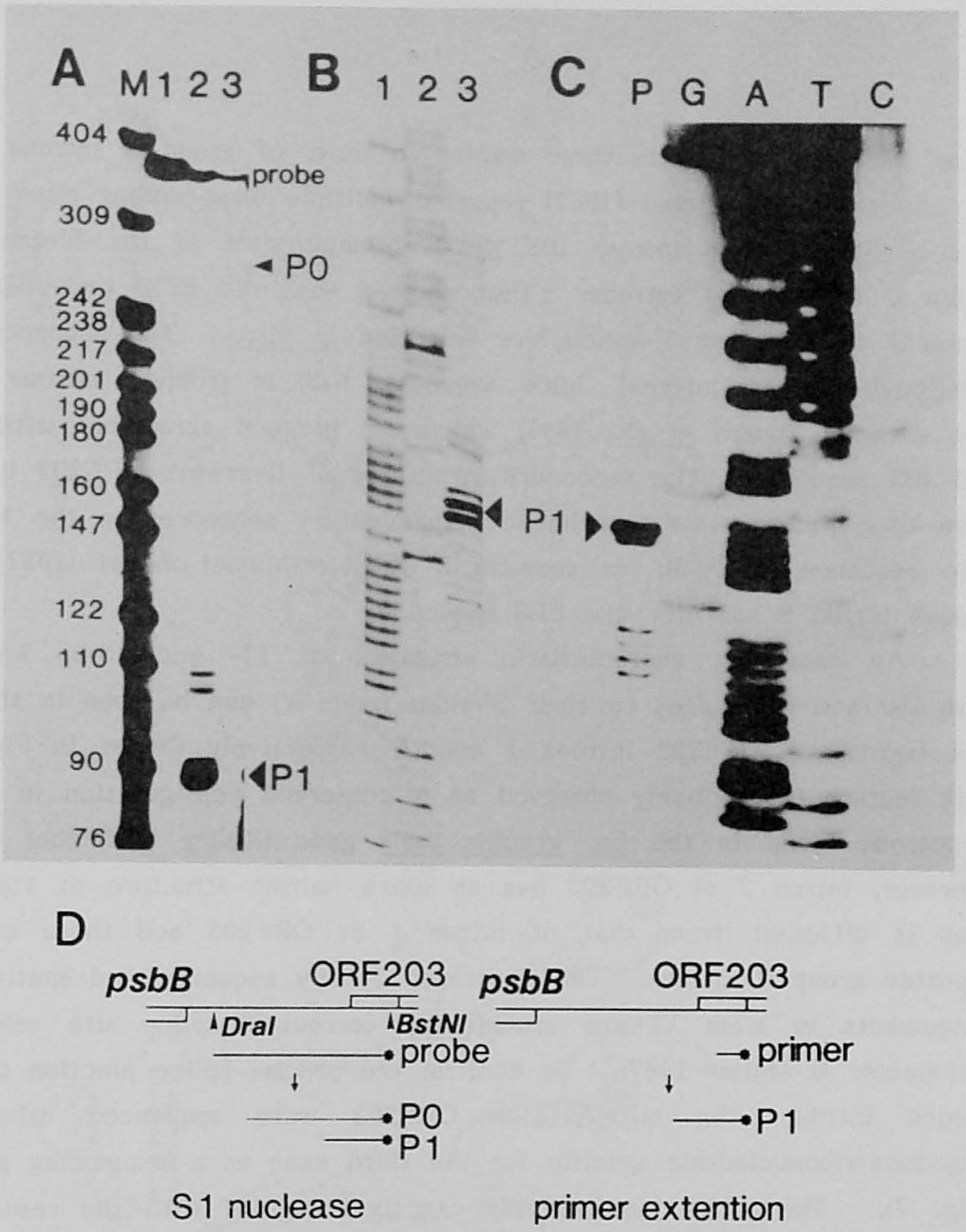


Fig. 5. S1 nuclease protection and primer extension analyses. A DNA restriction fragment (371 bp) was prepared by BstNI and DraI digestion, and labeled at the 5'-end of the BstNI site. Panel A shows patterns of DNA fragments protected from S1 nuclease digestion. Lane M, HpaII-digested pBR322 DNA fragments as size markers; lane 1, DNA probe prepared; lane 2, S1 nuclease treatment with chloroplasts RNA. P0 corresponds to a minor primary transcript. P1 indicates a major protected band of S1 nuclease corresponding to the processed 5' end of ORF203 transcripts; lane 3, S1 nuclease treatment without chloroplast RNA. Panel B is a pattern of S1 nuclease protection analysis of panel A in the high-resolution gel with Maxam-Gilbert sequence ladders; lane 1 shows a base specific ladder (G + A), lane 2 shows another base specific ladder (T + C), and lane 3 shows major protected product of S1 nuclease (P1). Panel C shows primer extension analysis. The 16-mer oligodeoxyribonucleotide complementary to ORF203 exon 1 was used as a primer; lane P is the major product (P1) of primer extension. Lanes G, A, T and C indicate the dideoxy sequencing ladders as a site indicator. Panel D shows schematic diagram of S1 protection and primer extension analyses.

trons form secondary structures similar to those of group II introns (Fig. 6). Jacquier and Michel (1987) reported multiple exon-binding sites (EBS, exon-binding sites in intron; IBS, intron-binding sites in the 5'-exon) in group II self-splicing introns. They showed that the EBSs and IBSs are essential in affecting 5'-splice site selection _in vitro_. The sequences of exon-binding sites (Internal Guide Sequence, IGS) in group II introns were also detected (Ozeki _et al._, 1987) which are in good agreement with EBS and IBS sequence. The secondary structures of liverwort ORF203 introns have exon-binding sites that have complementary sequences to the 5' exonic sequence (Fig. 6), as seen in a Jacquier-Michel model (1987), although intron 2 has only one EBS sequence.

An especially characteristic structure of 12- and 15-bp hairpins with UU and UA bulges on their 3'-sides (stem V) can be seen in the 3'-end regions of ORF203 introns 1 and 2, respectively (boxes in Fig. 6). This feature is commonly observed as a conserved configuration in group II introns found in the _Eu. gracilis psbA_ gene (Keller & Michel 1985). However, intron 2 of ORF203 has an extra hairpin structure at stem VI that is different from that of intron 1 of ORF203 and those of the reported group II introns. The conserved 3'-APy sequence and spatial arrangements in stem VI are crucial for correct 3'-splice site selection (Schmelzer & Muller 1987). To confirm the precise splice junction of the second intron, the mRNAs for ORF203 were sequenced using an oligodeoxyribonucleotide specific for the third exon as a sequencing primer (Fig. 7). The splice junction was exactly coincided with the results of Northern hybridization analysis and the alternative 3' splice site was not seen. The secondary structure was unusual, but our results demonstrated that the 3'-splicing junction was positioned at a distance of 2 to 3 nt from the last stem-loop structure, and a mismatched adenosine residue that is a possible branching point was observed in the last stem-loop structure.

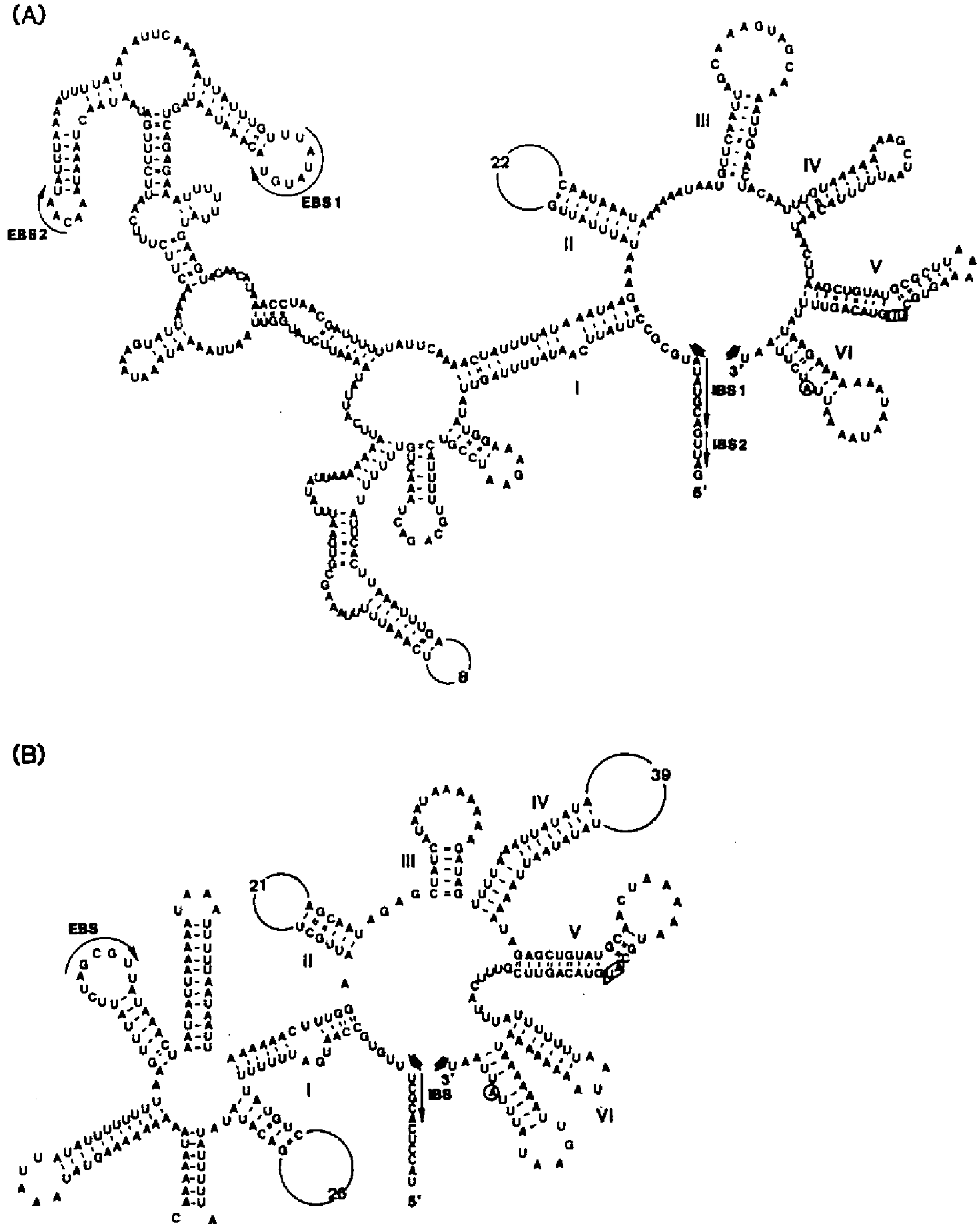


Fig. 6. Secondary structures of introns in the ORF203 gene. (A) First intron of ORF203. (B) Second intron of ORF203. The arrows indicate the 5'- and 3'- splice sites. Possible branching adenine residues are marked by circles. The number in the circles indicates the length of nucleotides. Abbreviations: IBS, indicate intron-binding site, and EBS, exon-binding site (Ozeki et al. 1987; Jacquier and Michel 1987).

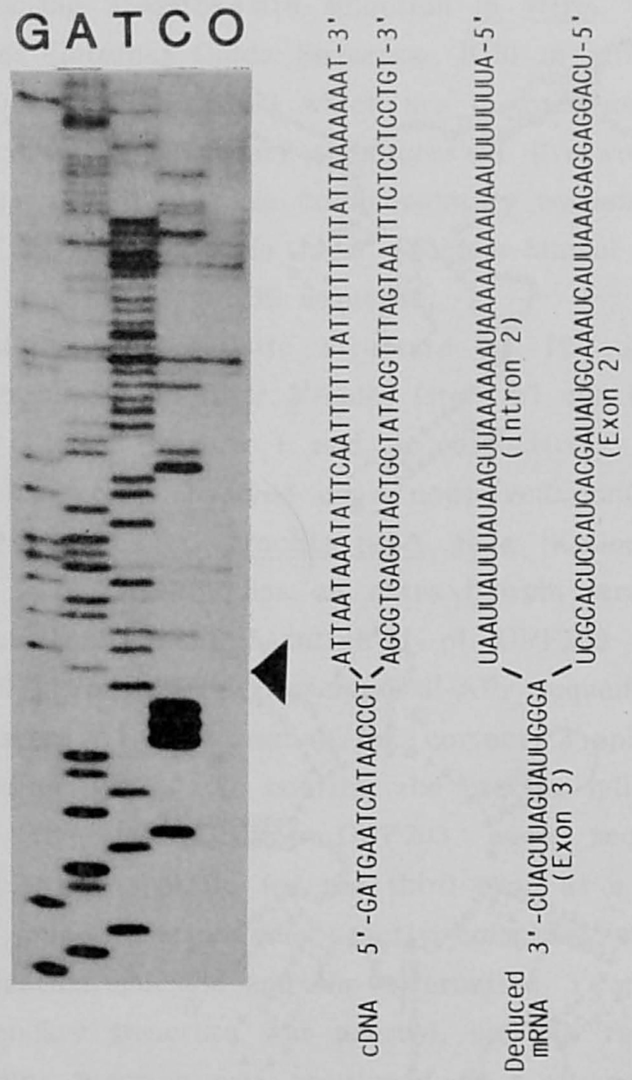

G A T C O
cDNA
5'-GATGAATCATAACCCT
ATTAATAAATATTCAATTTTTTATTTTTTTATTAAAAAAAT-3'
AGCGTGAGGTAGTGCTATACGTTTAGTAATTTCTCTCCTGT-3'
Deduced mRNA
3'-CUACUUAGUAUUGGGA
(Exon 3)
UAAUUAUUUAUAAGUUAAAAAAUAAAAAAUAAUUUUUUUA-5'
(Intron 2)
pre-mRNA (Unspliced)
UCGCACUCCAUCACGAUAUGCAAAUCAUUAAAGAGGAGGACU-5'
(Exon 2)
mature mRNA (Spliced)

Fig. 7. Sequences of ORF203 mRNA. Primer specific for the third exon (Fig. 2) was extended with the use of reverse transcriptase in the chain terminator procedure. In lane 0, no dideoxynucleotide was added. In lanes G, A, T, and C, dideoxynucleotide was added. The cDNA sequence derived from the autoradiogram is given for comparison to the deduced mRNA sequences.

```
                     .         .  ▼      .         .         .        60
liverwort   MPIGVPKVPFRLPGEEDAVWIDVYNRLYRERLLFLGQQVDDEIANQLIGIMMYLNGEDES
tobacco     :::::::::::S::::::S:V::::::::::::::::E::S::S:::::L:V::SI:::T
spinach     :::::::::::S::::::S:V::
                     .         .         .████████████████████  .      120
liverwort   KDMYLYINSPGGAVLAGISVYDAMQFVVPDVHTICMGLAASMGSFILTGGEITKRIALPH
tobacco     ::L::F:::::W:IP:VAI::T::::R:::::::::::::::::::V:::::::L:F::
            ▼        .         .         .         .         .      180
liverwort   ARVMIHQPASSYYDGQAGECIMEAEEVLKLRDCITKVYVQRTGKPLWVISEDMERDVFMS
tobacco     :::::::::::F:EA:T::FVL::::L::::ETL:R:::::::::::V:::::::::::

                     .         .
liverwort   AKEAKLYGIVDLVAIENNSTIKN
tobacco     :T::QA::::::::V:              76.4%
```

Fig. 8. Comparison of the amino acid sequence of liverwort ORF203 to the amino acid sequence deduced from tobacco chloroplasts (Shinozaki et al. 1986, Databank-EMBL release 12.0, complementary strand of position 72,465 - 74,504) and spinach X-gene (Westhoff 1985). Amino acid residues are shown by one-letter symbols and identical amino acid residues are replaced by colons. Sequence homology (%) to liverwort gene product is given at the end of the sequence. Arrows indicate the splicing junctions in their sequences.

Characterization of the ORF203 gene product

The characteristics of the ORF203 gene product were analyzed by a computer for hydropathy of the protein. A putative *trans*-membrane domain (Klein *et al.*, 1985) was found at positions 91-107 of the amino acid sequence (thickbar in Fig. 8) indicating that ORF203 gene product could be associated with a chloroplast membrane. Generally, functionally related genes form a cluster in the chloroplast genome (Ohyama *et al.*, 1986). The ORF203 gene product is not likely to be a chloroplast ribosomal protein, although it is located in a ribosomal protein gene cluster (*rps12'* and *rpl20*). Cluster analysis in a computer by the method of Ward (1963) showed that the ORF203 gene product was related into photosynthetic genes on the basis of amino acid composition (Sano & Ohyama, unpublished results). In chloroplasts, thylakoid polypeptides are translated on thylakoid-bound ribosomes and stromal polypeptides are translated on stromal ribosomes (Minami & Watanabe 1984). This strongly suggests that major RNA processing between ORF203 and *rps12'* would be responsible for the sorting of membrane-bound ORF203 mRNA from the mRNA for stromal ribosomal proteins S12 and L20.

The tobacco chloroplast genome has been sequenced (Shinozaki et al., 1986b). The region containing two ORFs (ORF74B and ORF73) in the tobacco chloroplast genome corresponds to liverwort ORF203 when our consensus sequences for introns are introduced in the tobacco chloroplast sequence. The derived amino acid sequence of the corresponding tobacco frame showed high homology (76.4%) to the liverwort ORF203 gene product (Fig. 8). Twenty out of the 23 amino acids in exon I of ORF203 are identical to a reading frame from the spinach 'X-gene', which gives an unidentified transcript and is also on the opposite DNA strand of the psbB gene (Westhoff 1985). S1 nuclease protection analysis of the X-gene suggested that there is active expression of liverwort ORF203 homologue in spinach chloroplasts, although the presence of introns in the X-gene has not been described (Westhoff 1985). Northern hybridization also shows an abundance of mature mRNA for the ORF203 gene in the liverwort chloroplasts. These results suggest that the product of ORF203 in the liverwort chloroplast genome is functional in chloroplasts, associating with a chloroplast membrane.

CHAPTER IV A nicked group II intron and trans-splicing in liverwort, Marchantia polymorpha, chloroplasts

INTRODUCTION

A number of chloroplast genes are interrupted by introns and thus require post-transcriptional RNA splicing for the gene expression (Ohyama et al., 1986; Ozeki et al., 1987; Umesono & Ozeki, 1987; Shinozaki et al., 1986b). Chloroplast introns belong to either group I or group II depending on their secondary structures, which was first elucidated in mitochondria (Michel & Dujon, 1983). Both of these intron families are accounted as self-splicing types, which suggests that catalytic activity resides in the intron RNA itself even if some protein factors are required for efficient splicing. Twenty different introns have been detected in the chloroplast genome of a liverwort, M. polymorpha; only one intron, in the trnL(UAA) gene, belongs to group I, and the other 19 in group II (Ohyama et al., 1986; Ozeki et al., 1987). Among these split genes, there was an unusual organization for the rps12 gene that encodes the 30S ribosomal protein S12; the gene consists of three exons and two introns, and exon 1 is far (some 60 kb) from the other two exons on the opposite DNA strand (Fukuzawa et al., 1986). Fig. 1 illustrates the genes neighboring the two separated parts, designated as rps12' (A) and rps'12 (B). Essentially the same gene organization as that of rps12 has also been reported for the tobacco chloroplast genome, although the portion of gene containing exons 2-3 (rps'12) is duplicated in the inverted-repeat regions (Shinozaki et al., 1986b; Fromm et al., 1986; Torazawa et al., 1986; Zaita et al., 1987; Hildebrand et al., 1988). This gene organization suggests that the two parts of rps12 are transcribed separately and then spliced trans to assemble exons 1 and 2. Electron microscopic analysis of RNA-DNA hybrids in tobacco showed that there are separate transcripts of exon 1 and 2-3 of the rps12 gene in the chloroplast as well as spliced RNA molecules in which exon 1 is joined to exon 2 (Koller et al., 1987). The gene pslA1 for photosystem I P700 protein in

C. reinhardii chloroplasts has been reported to be a discontinuous gene split into three separated locations, suggesting the operation of trans-splicing (Kück et al., 1987; Choquet et al., 1988).

Two separate RNA molecules can be joined by trans-splicing in vitro in nuclear extracts of HeLa cells (Solnick, 1985; Konarska et al., 1985). In yeast, efficient trans-splicing in vitro of the mitochondrial group II intron suggests that interaction between the 3'-end of the 5'-exon and intron is needed (Jacquier & Rosbash, 1987). Recently several cases of trans-splicing in vivo have been reported (see review: Sharp, 1987). Unlike these cases, however, the chloroplast gene for ribosomal protein S12 involved the trans-splicing of the coding region for a single protein.

There were various RNA molecules derived from the two regions of the liverwort chloroplast genome shown in Fig. 1. Maturation pathways were postulated for mRNAs from the primary RNA transcripts via RNA processing and splicing cis and trans, and a bimolecular interaction model was proposed for trans-splicing according to the folding model of group II introns of Michel and Dujon (1983).

MATERIALS AND METHODS

Preparation of liverwort chloroplast RNA

Liverwort chloroplasts were isolated from the cells of two-week-old suspension cultures grown under continuous illumination (Ohyama et al., 1982). Chloroplast RNA was prepared as described before (Kohchi et al., 1988b).

Northern hybridization analysis

Oligodeoxyribonucleotide probes (16 nucleotides long) were synthesized with a DNA synthesis apparatus (Shimadzu NS-1) and labelled with [γ-^{32}P]ATP (Amersham, 5000 Ci/mmol) and T4 polynucleotide kinase (Takara Shuzo). RNA probes were prepared with in vitro transcription of SP6 vectors containing liverwort chloroplast DNA fragments by the use of [α-^{32}P]UTP (Amersham, 800 Ci/mmol). Northern hybridization was done at 45°C for oligonucleotide probes and in the presence of 50% formamide for riboprobes by a procedure described previously (Kohchi et al., 1988b).

S1-nuclease protection analysis

DNA probes for S1-nuclease protection analysis were obtained from recombinant plasmids containing liverwort chloroplast DNA. The probes were labelled at the 5'-end as described above and at the 3'-end with [α ^{32}P]dCTP (Amersham, 3000 Ci/mmol) and Klenow fragment of DNA

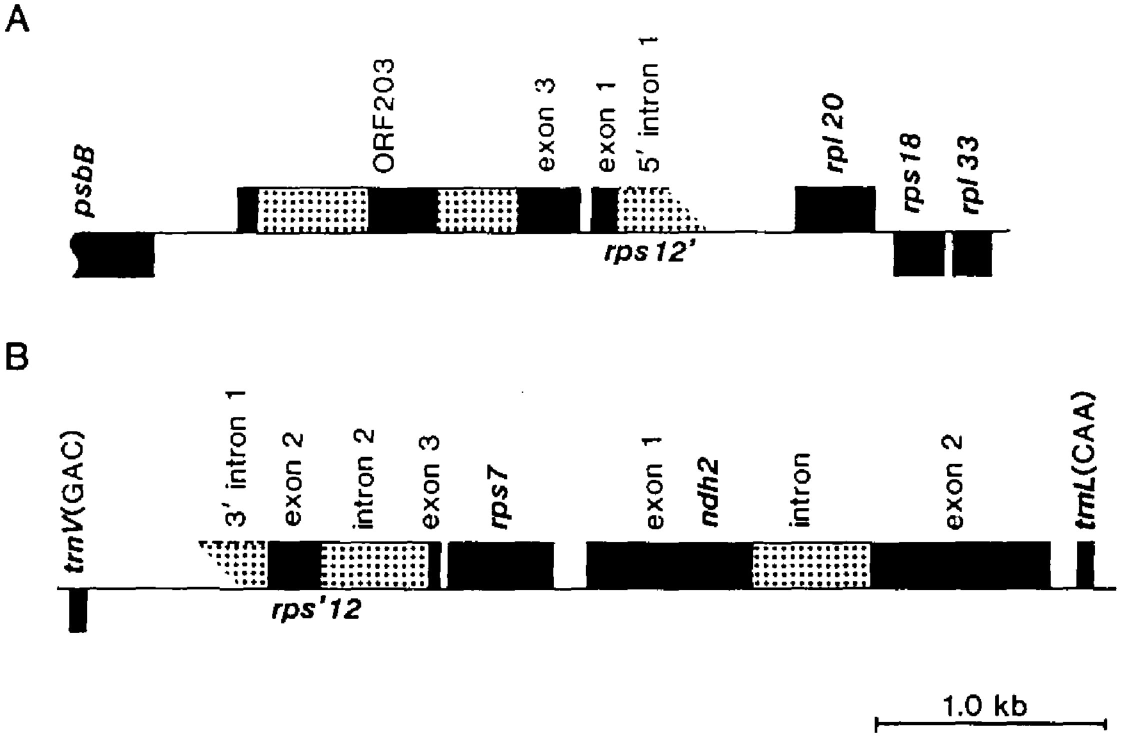


Fig. 1. Gene organization of the rps12 gene. (A) Gene organization in the region coding for rps12'. Symbols are: psbB for the 47-kDa chlorophyll a apoprotein in photosystem II; ORF203 for an open reading frame of 203 amino acids; and rps12, rpl20, rps18, and rpl33 for ribosomal proteins S12, L20, S18, and L33, respectively. (B) Gene organization of the region coding for rps'12. Symbols: rps7 for ribosomal protein S7; ndh2 for NADH dehydrogenase (ND2); and trnV(GAC) and trnL(CAA) for valine tRNA(GAC) and leucine tRNA(CAA), respectively. Genes shown above the lines are transcribed to the right, and those under the lines are transcribed to the left. Hatched boxes are introns.

polymerase I (Takara Shuzo). S1-nuclease protection analysis was done by a published procedure (Berk & Sharp, 1977).

RESULTS

Primary RNA transcripts of rps12' and rps'12 and the processing products

The 5'-portion of rps12 gene (rps12') was proceeded by an unidentified ORF203 that carried two introns, and followed by the rpl20 gene for the 50S ribosomal protein L20 (Fig. 1A). The remaining 3'-portion of rps12 (rps'12) was followed by the rps7 gene for the 30S ribosomal protein S7, ndh2 (a counterpart of human mitochondrial ND2), and

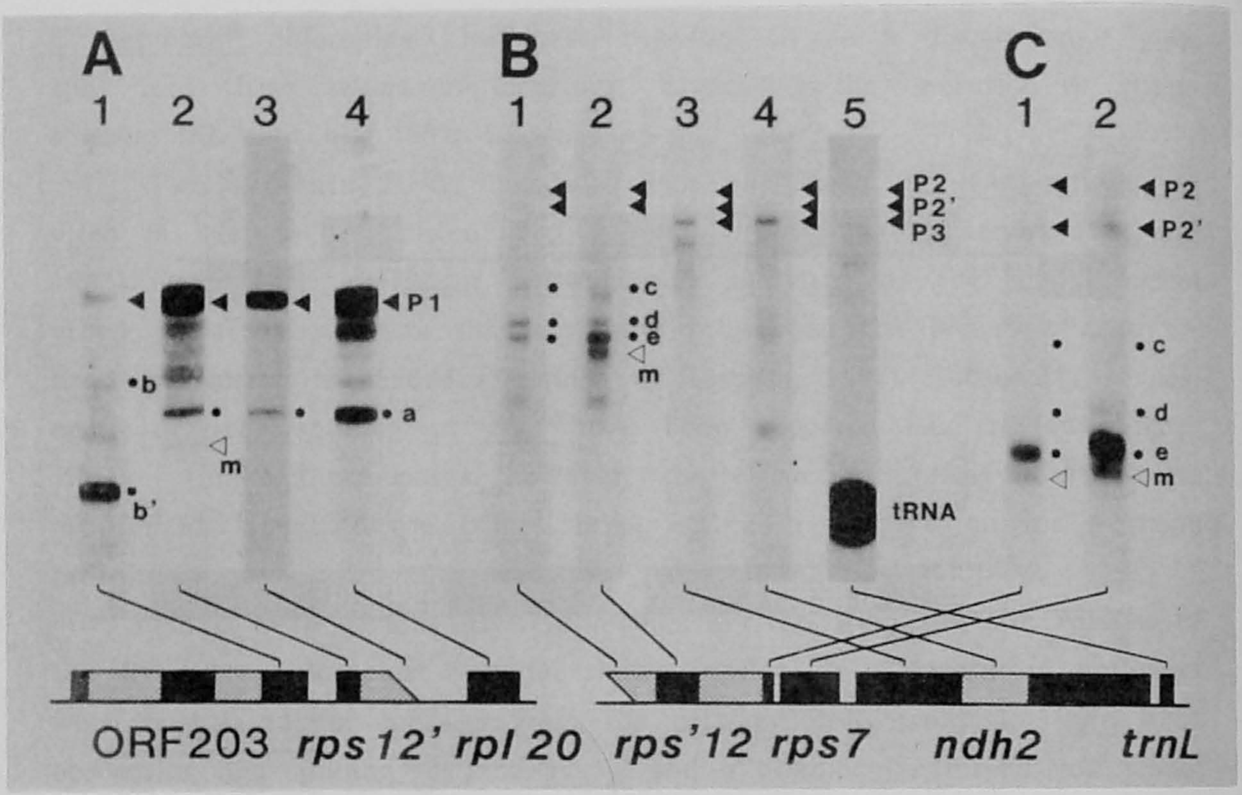


Fig. 2. Northern hybridization with a variety of probes. (A) Probes specific to exon 3 of ORF203 (lane 1), exon 1 of rps12' (lane 2), 5'-intron 1 of rps12' (lane 3), and the coding region of rpl20 (lane 4). P1 with an arrowhead indicates primary transcripts including ORF203 to rpl20; band a indicates processed mRNA containing rps12 (exon 1) and rpl20 genes; band b indicates processed mRNA for ORF203 with introns, and band b' indicates spliced mature mRNA of ORF203. Band m (open triangle) indicates mature mRNA for rps12 gene products. (B) Probes specific to 5'-intron 1 (lane 1), exon 2 (lane 2) of rps'12 gene, exon 1 (lane 3) and intron (lane 4) of ndh2 gene, and for the coding region (lane 5) of the trnL(CAA) gene. P2 and P2' with arrowheads indicate the two kinds of transcripts described in the text. P3 with an arrowhead indicates an additional transcript for ndh2. Band m with an open arrowhead indicates mature mRNA for the rps12 gene. (C) Probes specific to coding regions for exon 3 of the rps'12 and rps7 genes. Lane 1, probe for exon 3 of the rps12 gene; lane 2, for the rps7 gene. Band m with an open arrowhead indicates mature mRNA for the rps12 gene product.

(CAA) for the leucine tRNA(CAA) in this order (Fig. 1B). Chloroplast RNAs were extracted from the cultured liverwort cells liverwort, and Northern hybridization experiments were done with appropriate probes. Two separate transcripts for rps12' and rps'12 were detected (Fig. 2).

The results obtained with the probes for the ORF203 (exon 3), the *rps12* exon 1, the 5'-portion of *rps12* intron 1, and the *rpl20* gene are presented in Fig. 2A. The *rps12'* was transcribed together with the upstream ORF203 and the downstream *rpl20* as a primary RNA of 3.0 kb (band P1, Fig. 2A). I also detected a 1.55-kb RNA (band *b*) that was hybridized only with the ORF203 probe and a 1.45-kb RNA (band *a*) hybridized only with the *rps12'* and *rpl20* probes (Fig. 2A). These two RNA species may be cleaved products of the 3.0-kb primary RNA transcript that was processed at a specific site between ORF203 and *rps12'*. Several bands were observed between bands P1 and *a*, which may be partially spliced molecules of primary transcripts (lane 2, 3 and 4, Fig. 2A). The size of band b' RNA corresponded to that of the mature mRNA of ORF203 after the splicing of its introns (Kohchi *et al.*, 1988c). There was no unique band that was exclusively hybridized with the probe of *rpl20*. However, S1-nuclease protection analysis with a 5'-^{32}P-labelled DNA probe of the *rpl20* coding region gave smeared bands, suggesting that there was no particular fixed end for the 5'-leader of *rpl20* mRNA (data not shown).

Northern hybridization was done with a variety of probes for the 3'-portion of intron I (the 5'-leader sequence of exon 2), exon 2 of the *rps12* gene, exon 1 of *ndh2* gene, an intron of *ndh2* gene, and the *trnL*(CAA) gene (Fig. 2B). The results showed that: (i) the primary transcripts (4.2 kb, band P2; 4.0 kb, band P2') correspond in their entire region from *rps'12* (exons 2 and 3) to the *trnL*(CAA) gene; (ii) the major bands (1.9 kb, 1.4 kb, and 1.2 kb; bands *c*, *d*, and *e*, respectively; Fig. 2B) were mRNA molecules processed upstream from the *ndh2* gene; and (iii) there was an additional transcript for the *ndh2* gene (band P3 in lanes 3 and 4, Fig. 2B). A large amount of *trnL*(CAA) gene product was found in the RNA preparation (lane 5, Fig. 2B). This may not reflect only the stability of tRNA molecules but also the presence of additional initiation sites for *trnL*(CAA) transcription alone. Northern hybridization with probes specific to the coding region for *rps12* exon 3 and *rps7* gave the same pattern, indicating that there was no RNA processing between them (lanes 1 and 2, Fig. 2C). The linkage of rps12 and rps7 was maintained on the mRNA, like that in E. coli (Post & Nomura, 1980).

Processing of two separate rps12 transcripts (exon 1 and exons 2-3)

S1-nuclease protection analysis at the 5'-flanking region of the rps12 exon 1 showed that the processing site was between the ORF203 and the rps12 exon 1 (Fig. 3A). This result coincided with Northern hybridization analysis, which showed that there were major processed mRNA molecules for the ORF203 gene (see band b, Fig. 2A) and for rps12'-rpl20 (see band a, Fig. 2A). The major termination site was 45 nt downstream from the rpl20 gene, although two minor signals (110 nt and 220 nt downstream) were also observed (Fig. 3B). To identify the initiation site of transcripts in the gene cluster of rps'12, S1-nuclease protection analysis was done with a DNA probe containing the HphI-EcoRV fragment, giving four bands (Fig. 3C). One could correspond a spliced RNA (band s, Fig. 3C). A minor band c (Fig. 3C) corresponded to the junction site (83 nt upstream from the 5'-terminal of exon 2) of inverted repeat and large single copy regions, probably because of competitive hybridization with transcripts in the other inverted repeat region. This band, therefore, did not correspond to the initiation site of the transcription. Two additional major bands were detected: one corresponded to the RNA molecule that started 500 nt upstream from the 5'-terminal of exon 2 of rps12 (band a, Fig. 3C), and the other was 310 nt upstream from exon 2 (band b, Fig. 3C). The former could be for the initiation of transcripts of the gene cluster, because of the presence of promoter-like sequences upstream ("TTGACC" and "TAAAAT" as "-35"- and "-10" -sequences, respectively). RNA molecules corresponding to these signals in size were found in the Northern hybridization analysis (see bands P2 and P2', Fig. 2B). S1-nuclease protection analysis to detect the processing site between the rps7 and ndh2 genes gave evidence of two major processed RNA molecules (40 and 90 nt downstream from the termination

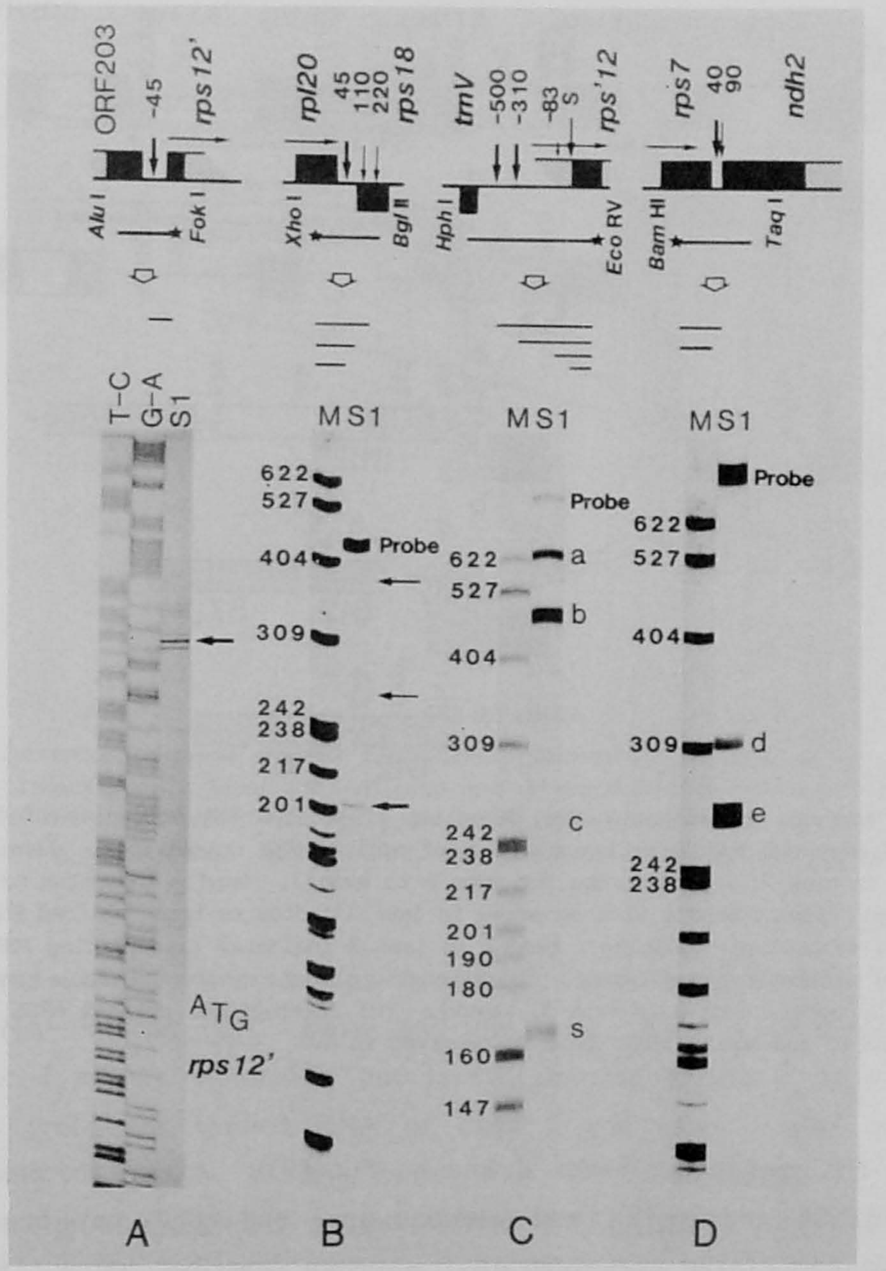


Fig. 3. S1-nuclease protection analysis of transcripts. (A) DNA probe (AluI-FokI fragment) between the ORF203 (exon 3) and rps12' (exon 1) genes. (B) DNA probe (XhoI-BglII fragment) between the rpl20 and rps18 genes. (C) DNA probe (HphI-EcoRV fragment) covering the spacer region between the trnV(GAC) and rps'12 genes. (D) DNA probe (BamHI-TaqI fragment) covering the rps7 and ndh2 genes. Above each panel, signals from S1-nuclease protection analysis are shown with the gene organization. Numbers with arrows indicate the site of signals from the following exons or the coding region of the respective genes. Asterisks indicate the site of ^{32}P-labeled ends. Hatched boxes are introns.

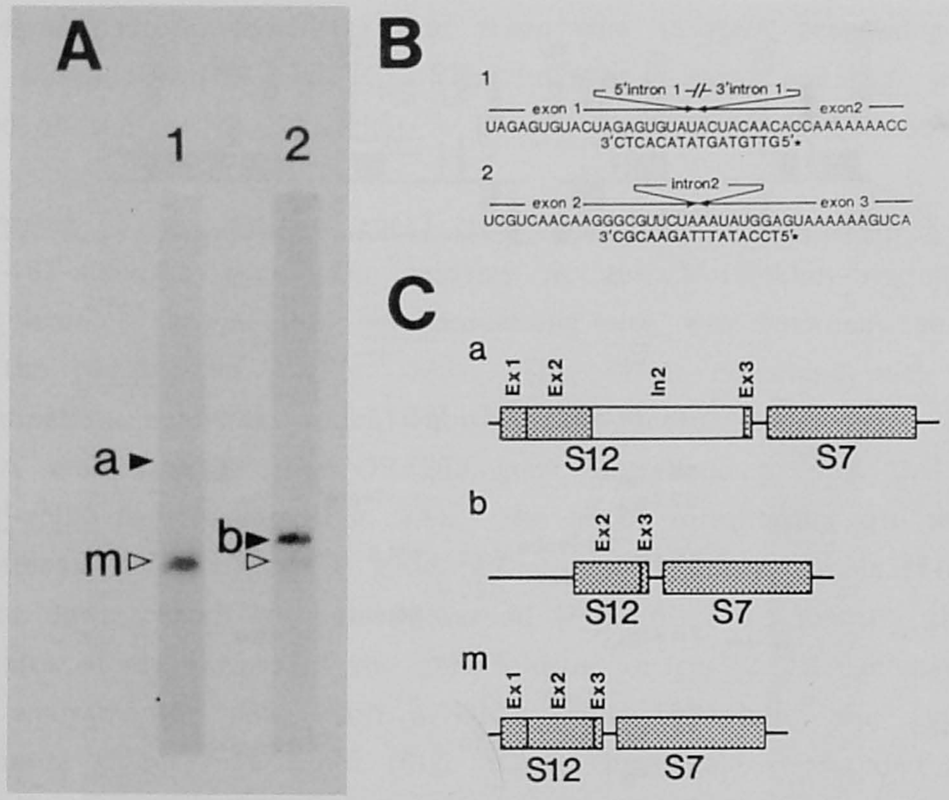


Fig. 4. Northern hybridization for detection of spliced RNA molecules of the rps12 gene. (A) Northern hybridization analysis of spliced RNA transcripts. Lane 1, probe for exon 1 to exon 2; lane 2, probe for exon 2 to exon 3. Band m indicates mature mRNA for the rps12 gene. Band a with an arrow in lane 1 indicates trans-spliced RNA (exon 1 and exon 2) without cis-splicing. Band b in lane 2 indicates cis-splicing RNA (exon 2 and exon 3) without trans-splicing. (B) Synthetic oligodeoxyribonucleotide probes (exon 1 to exon 2, upper; exon 2 to exon 3, lower). (C) Intermediate spliced mRNA molecules (bands a and b) and mature mRNA (band m) observed in A.

codon of the rps7 gene, bands d and e, Fig. 3D). Major processings between ORF203 and rps12', and between rps7 and ndh2 may be required for the folding of the two separate transcripts described below.

Independent trans- and cis-splicing in the rps12 gene.

To confirm the presence of trans- and cis-spliced mRNA molecules, synthetic oligodeoxyribonucleotides 16 nucleotides long consisting of 8 nucleotides complementary to the 3'-end of a 5'-exon and 8 nucleotides complementary to the 5'-end of the following 3'-exon were used as probes for Northern hybridization (Fig. 4B). Spliced RNAs of

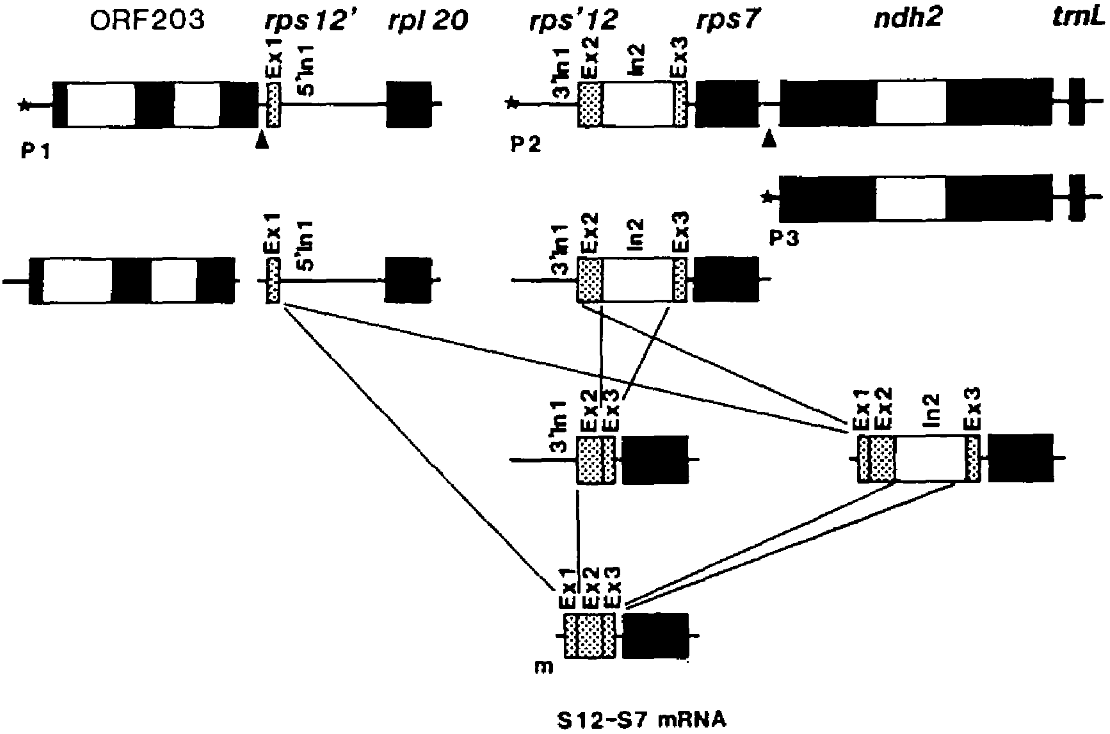


Fig. 5. Schematic pathways of mRNA maturation. Arrowheads indicate major processing sites of transcripts. Asterisks indicate the 5'-terminals of primary transcripts. Dotted boxes indicate exons of the *rps12* gene. In other genes, solid box are exons and white boxes are *cis* introns. Lines indicate the processes of *trans*- and *cis*-splicing.

exon 1 and exon 2 were seen as two bands: the major band *m* was mature mRNA for the *rps12* gene, and the minor band *a* corresponded to *trans*-spliced mRNA molecules, but it still carried intron 2 (lane 1, Fig. 4A). A probe for spliced RNA of exon 2 and exon 3 also gave two bands, with the major band *b* having the 3'-half of intron 1, and the minor band *m* corresponding to the mature RNA for the *rps12* gene (lane 2, Fig. 4A). These results indicate the independent occurrence of *trans*- and *cis*-splicing in the transcripts of the *rps12* gene (Fig. 4C), although *cis*-splicing was apparently more efficient than *trans*-splicing in the liverwort chloroplast.

DISCUSSION

The results obtained are summarized in Fig. 5, which is a probable

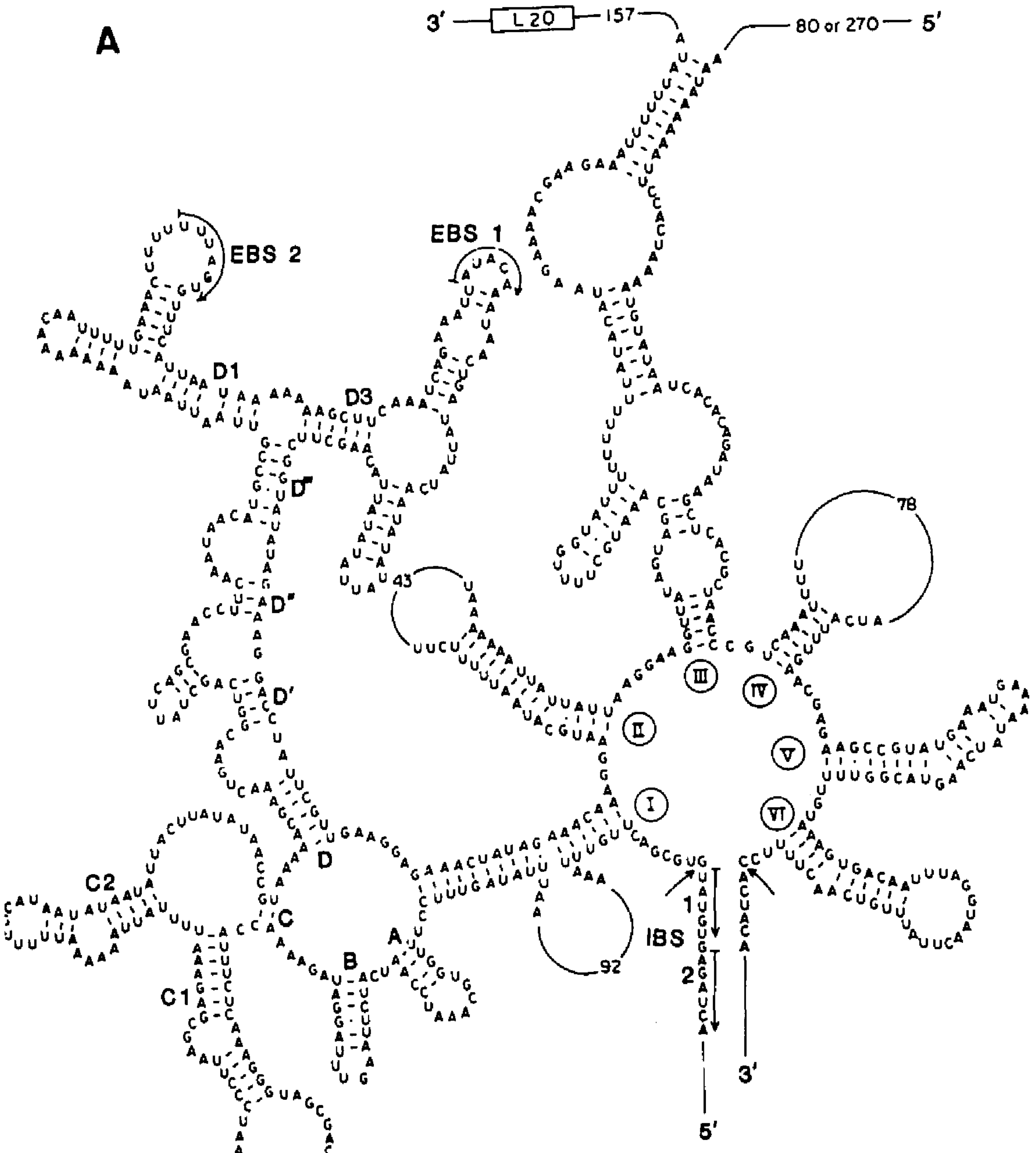
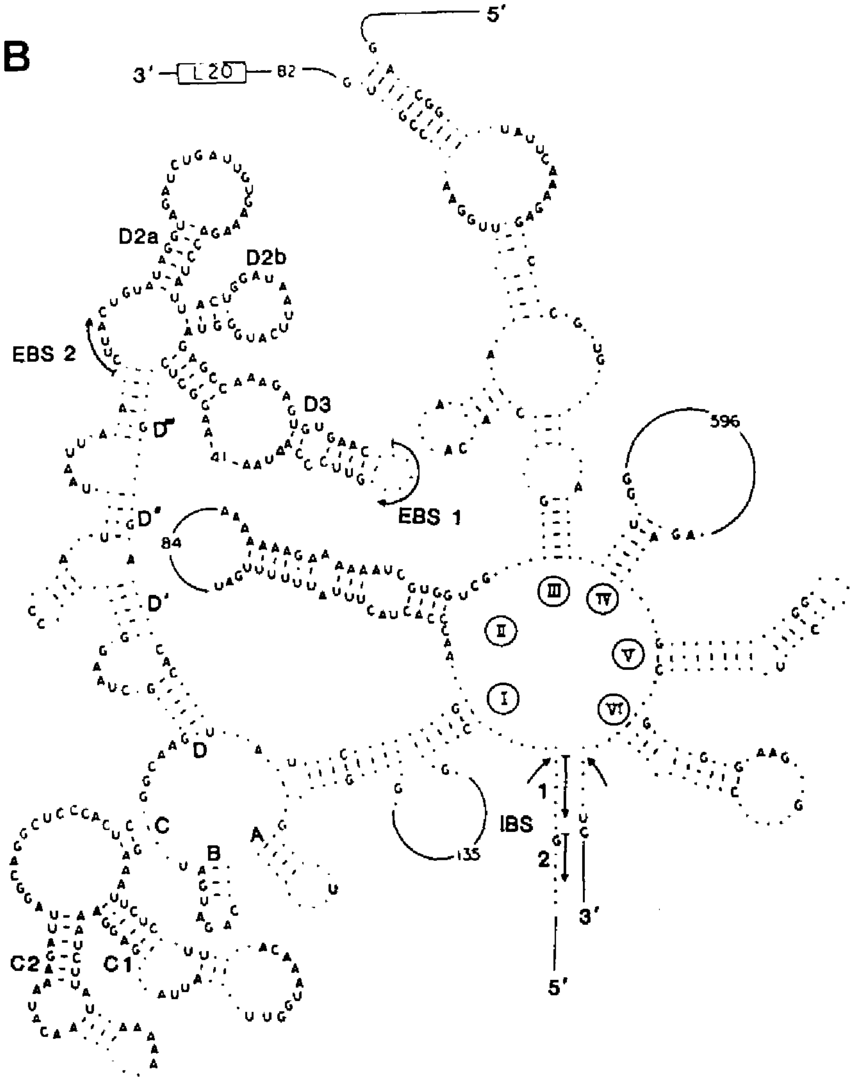


Fig. 6. A model of secondary structure of trans-split introns in the rps12 gene. (A) Liverwort; (B) Tobacco (dots indicate the same nucleotides as liverwort). Arrows indicate splicing sites. Abbreviations IBS and EBS with arrows indicate the intron-binding and exon-binding sites (Ozeki et al., 1987; Jaquier & Michel, 1987; Michel & Jaquier, 1987). The nomenclature of stem-loop structure follows that of the model of Jacquier and Michel (1987).



scheme for the mRNA maturation pathways in the chloroplasts. Note that in this analysis there are a variety of precursor RNAs relatively abundant compared to the amount of matured mRNAs. This is mainly because the liverwort chloroplasts from the cell cultures in the late growth stage were used (2 weeks old, and grown under continuous illumination). If the cells were transferred to fresh medium, mature mRNAs in the chloroplasts became predominant within a day or two, suggesting the conversion of accumulated intermediate RNAs to mRNAs (Kohchi et al., 1988c).

Fig. 6A illustrates a folding model of intron 1 of the rps12 gene for its trans-splicing. On the identified precursor RNAs of rps12' and rps'12, there are stretches of nucleotide sequences complementary with each other as deduced from the DNA sequence, and the model was constructed by assuming base pairings between them. A simplified form of this model has been published by elsewhere (Ozeki et al., 1987). The overall structure has characteristics of group II introns, with its six stem-loop domains (5), but with an interruption in the third loop region. The exon- and intron-binding sites (the EBS and IBS sequence elements) are also indicated in Fig. 6; they are complementary with each other and essential for the splicing (Jacquier & Michel, 1987; Michel & Jacquier, 1987). A structure comparable to the tobacco intron 1 of rps12 can be deduced from the published sequence (Fig. 6B). Zaita et al. (1987) has reported the presence of complementary sequences ("transon 1" and "transon 2"). The sequence transon 1 corresponds to the regions from A to C in stem-loop I, and transon 2 is in the loop portion of stem-loop IV in our model. In the liverwort genome, a complementary structure could not be formed in the regions corresponding to the tobacco transons. Evidence for our model was the compensatory base substitutions observed between liverwort and tobacco to keep base pairings at the stem regions. Two mutations that disrupt the splicing of bI1 intron in yeast mitochondria have been mapped in the stem portion of stem III (Schmelzer et al., 1983), indicating the importance of the stem structure for splicing. The intermolecular base pairings may participate in the folding of group II intron as a "ribozyme" and in the selection of an adequate partner molecule. There was no complementary sequence to the stem portion of stem-loop III of the trans-split intron in any other introns of the liverwort chloroplasts (Ohyama et al., 1986). Thus, the exon shuffling in chloroplast RNA splicing could not be taken into account. Self-splicing of chloroplast group II intron has not yet reported, but catalytic activity would essentially reside in the introns, depending on the tertiary folding structure. The resultant RNA complex may be essential for the removal of trans-split intron.

REFERENCES

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., 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. & Young, I. G. (1981) *Nature* **290**, 457-465.

Arcari, P. & Brownlee, G. G. (1982) *Nucl. Acids Res.* **8**, 5207-5212.

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. & Young, I. G. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3164-3166.

Bedwell, D., Davis, G., Gosink, M., Post, L., Nomura, M., Kestler, H., Zengel, J. M. & Lindahl, L. (1985) *Nucl. Acids Res.* **13**, 3891-3903.

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

Berk, A. J. & Sharp, P. A. (1977) *Cell* **12**, 721-732.

Bishop, P. E., Jarlenski, D. M. L., & Hetherington, D. R. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7342-7346.

Bonitz, S. G., Berlani, R., Coruzzi, G., Li, M., Macino, G., Nobrega, F. G., Nobrega, M. P., Thalenfeld, B. E. & Tzagoloff, A. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3167-3170.

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

Brosius, J., Palmer, M. L., Kennedy, P. J. & Noller, H. F. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4801-4805.

Brosius, J., Dull, T. J. & Noller, H. F. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 201-204.

de Bruijn, M. H. L., Schreier, P. H., Eperon, I. C., Barrell, B. G., Chen, E. Y., Armstrong, P. W., Wong, J. F. H. & Roe, B. A. (1982) *Nucl. Acids Res.* **8**, 5213-5222.

Cantrell, A. & Bryant, D. A. (1988) *Photosynthesis Res.* **16**, 65-81.

Cerretti, D. P., Dean, D., Davis, G. R., Bedwell, D. M. & Nomura, M. (1983) *Nucl. Acids Res.* **11**, 2599-2616.

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

Chomyn, A., Cleeter, M. W. J., Ragan, C. I., Riley, M., Doolittle, R. F. & Attardi, G. (1986) *Science* **234**, 614-618.

Choquet, Y., Goldschmidt-Clermont, M., Girard-Bascou, J., Kück, U., Bennoun, P. & Rochaix, J. (1988) *Cell* **52**, 903-913.

Courtice, G. R. M., Bowman, C. M., Dyer, T. A. & Gray, J. C. (1985) *Curr. Genet.* **10**, 329-333.

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

Crick, F. H. C. (1966a) *Scientific Am.* **245(4)**, 55-62.

Crick, F. H. C. (1966b) *J. Mol. Biol.* **19**, 548-555.

Crouse, E. J., Schmitt, J. M. & Bohnert, H. J. (1985) *Plant Mol. Biol. Rep.* **3**, 43-89.

de la Cruz, V. F., Neckelmann, N. & Simpson, L. (1984) *J. Biol. Chem.*

259, 15136-15147.

Deng, X-W. & Gruissem, W. (1987) *Cell* 49, 379-387.

Edwards, K. & Kössel, H. (1981) *Nucl. Acids Res.* 9, 2853-2869.

El-Gewely, M. R., Helling, R. B. & Dibbits, J. G. T. (1984) *Mol. Gen. Genet.* 194, 432-443.

Ellis, R. J. (1981) *Ann. Rev. Plant Physiol.* 32, 111-137.

Eneas-Filho, J., Hartley, M. R. & Mache, R. (1981) *Mol. Gen. Genet.* 184, 484-488.

Evans, I. J., Holland, I. B., Gray, L., Buckel, S. D., Bell, A. W. & Hermodson, M. A. (1986) *Nature* 323, 448-450.

Fearnley, I. M. & Walker, J. E. (1986) *EMBO J.* 5, 2003-2008.

Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van der Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. & Ysebaert, M. (1978) *Nature* 273, 113-120.

Fournier, M. J. & Ozeki, H. (1985) *Microbiol. Rev.* 49, 379-397.

Fromm, H., Edelman, M., Koller, B., Goloubinoff, P. & Galun, E. (1986) *Nucl. Acids Res.* 14, 883-898.

Froshauer, S. & Beckwith, J. (1984) *J. Biol. Chem.* 259, 10896-10903.

Fukuzawa, H., Kohchi, T., Shirai, H., Ohyama, K., Umesono, K., Inokuchi, H. & Ozeki, H. (1986) *FEBS Lett.* 198, 11-15.

Fukuzawa, H. (1986) PhD. thesis, Kyoto University.

Fukuzawa, H., Yoshida, T., Kohchi, T., Okumura, T., Sawano, Y. & Ohyama, K. (1987) *FEBS Lett.* 220, 61-66.

Fukuzawa, H., Kohchi, T., Sano, T., Shirai, H., Umesono, K., Inokuchi, H., Ozeki, H. & Ohyama, K. (1988) *J. Mol. Biol.* 203, 333-351.

Gamborg, O. L., Miller, R. A. & Ojima, K. (1968) *Exp. Cell Res.* 50, 151-158.

Glotz, C., Zwieb, C., Brimacombe, R., Edwards, K. & Kössel, H. (1981) *Nucl. Acids Res.* 9, 3287-3306.

Godson, G. N., Barrell, B. G., Staden, R. & Fiddes, J. C. (1978) *Nature* 276, 236-247.

Gold, B., Carrillo, N., Tewari, K. K. & Bogorad, L. (1987) *Proc. Natl. Acad. Sci. USA* 84, 194-198.

Green, P. J., Pines, O. & Inouye, M. (1986) *Ann. Rev. Biochem.* 55, 569-597.

Gruissem, W., Greenberg, B. M., Zurawski, G., Prescott, D. M. & Hallick, R. B. (1983) *Cell* 35, 815-828.

Guillemaut, P. & Weil, J. H. (1982) *Nucl. Acids Res.* 10, 1653-1659.

Gupta, R. (1984) *J. Biol. Chem.* 259, 9461-9471.

Hallick, R. B., Gingrich, J. C., Johanningmeier, U., Passavant, C. W. (1985) Introns in *Euglena* and *Nicotiana* chloroplast protein genes. In: van Vloten-Doting, L., Groot, G. S. P., Hall, T. C. (eds) Molecular form and function of the plant genome 1985: pp. 211-220 Plenum Press, New York & London.

Hausinger, R. P. & Howard, J. B. (1982) *J. Biol. Chem.* 257, 2483-2490.

Hearst, J. E., Alberti, M. & Doolittle, R. F. (1985) *Cell* 40, 219-220.

Heiland, I. & Wittmann-Liebold, B. (1979) *Biochemistry* 18, 4605-4612.

Higgins, C. F., Haag, P. D., Nikaido, K., Ardeshir, F., Garcia, G. & Ames, G.F.-L. (1982) *Nature* 298, 723-727.

Higgins, C. F., Hiles, I. D., Salmond, G. P. C., Gill, D. R., Downie, J. A., Evans, I. J., Holland, I. B., Gray, L., Buckel, S. D., Bell, A. W. & Hermodson, M. A. (1986) Nature **323**, 448-450.

Hildebrand, M., Hallick, R. B., Passavant, C. W. & Bourque, D. P. (1988) Proc. Natl. Acad. Sci. USA **85**, 372-376.

Hird, S. M., Dyer, T. A. & Gray, J. C. (1986) FEBS Lett. **209**, 181-186.

Høj, P. B., Svendsen, I., Scheller, H. V. & Møller, B. L. (1987) J. Biol. Chem. **262**, 12676-12684.

Hollingsworth, M. J., Johanningmeier, U., Karabin, G. D., Stiegler, G. L., Hallick, R. B. (1984) Nucl. Acids Res. **12**, 2001-2017.

Howard, J. B., Lorsbach, T. W., Ghosh, D., Melis, K. & Stout, C. D. (1983) J. Biol. Chem. **258**, 508-522.

Howe, C. J. (1985) Curr. Genet. **10**, 139-145.

Jacquier, A. & Michel, F. (1987) Cell **50**, 17-29.

Jacquier, A. & Rosbash, M. (1987) Science **234**, 1099-1104.

Jukes, T. H., Osawa, S., Muto, A. & Lehman, N. (1987) Cold Spring Harbor Symp. Quant. Biol. **52**, 769-776.

Kanehisa, M. I. (1982) Nucl. Acids Res. **10**, 183-196.

Karabin, G. D., Farley, M., Hallick, R. B. (1984) Nucl. Acids Res. **12**, 5801-5812

Kato, A., Takaiwa, F., Shinozaki, K. & Sugiura, M. (1985) Curr. Genet. **9**, 405-409.

Katoh, K. (1983) Physiol. Plant. **57**, 67-74.

Keller, M. & Michel, F. (1985) FEBS Lett. **179**, 69-73

Klein, P., Kanehisa, M. & DeLisi, C. (1985) Biochim. Biophys. Acta **815**, 468-476.

Koch, W., Edwards, K. & Kössel, H. (1981) Cell **25**, 203-213.

Kohchi, T., Shirai, H., Fukuzawa, H., Sano, T., Komano, T., Umesono, K., Inokuchi, H., Ozeki, H. & Ohyama, K. (1988a) J. Mol. Biol. **203**, 353-372.

Kohchi, T., Yoshida, T., Komano, T. & Ohyama, K. (1988b) EMBO J. **7**, 885-891.

Kohchi, T., Ogura, Y., Umesono, K., Yamada, Y., Komano, T., Ozeki, H. & Ohyama, K. (1988c) Curr. Genet. **14**, 147-154.

Kohchi, T., Umesono, K., Ogura, Y., Komine, Y., Nakahigashi, K., Komano, T., Yamada, Y., Ozeki, H. & Ohyama, K. (1988d) Nucl. Acids Res. **16**, 10025-10036.

Koller, B. & Delius, H. (1982) EMBO J. **1**, 995-998.

Koller, B., Gingrich, J. C., Stiegler, G. L., Farley, M. A., Delius, H. & Hallick, R. B. (1984) Cell **36**, 545-553.

Koller, B., Clarke, J. & Delius, H. (1985) EMBO J. **4**, 2445-2450.

Koller, B., Fromm, H., Galun, E. & Edelman, M. (1987) Cell **48**, 111-119.

Kolodner, R. D. & Tewari, K. K. (1975) Nature **256**, 708-711.

Konarska, M. M., Padgett, R. A. & Sharp, P. A. (1985) Cell **42**, 165-171.

Krebbers, E. T., Larrinua I. M., McIntosh, L. & Bogorad, L. (1982) Nucl. Acids Res. **10**, 4985-5002.

Kück, U., Choquet, Y., Schneider, M., Dron, M. & Bennoun, P. (1987) EMBO J. **6**, 2185-2195.

Kuchino, Y., Yabusaki, Y., Mori, F. & Nishimura, S. (1984) Nucl. Acids

Res. **12**, 1559-1562.

Kumano, M., Tomioka, N. & Sugiura, M. (1983) Gene **24**, 219-225.

Kyte, J. & Doolittle, R. F. (1982) J. Mol. Biol. **157**, 105-132.

Lagoutte, B., Setif, P. & Duranton, J. (1984) FEBS Lett. **174**, 24-29.

Lindahl, L. & Zengel, J. M. (1982) Adv. Genetics **21**, 53-121.

Long, E. O. & Dawid, I. B. (1980) Ann. Rev. Biochem. **49**, 727-764.

Malkin, R., Aparicio, P. J. & Arnon, D. I. (1974) Proc. Natl. Acad. Sci. USA **71**, 2362-2366.

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Laboratory Press, New York.

Maxam, A. M. & Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA **74**, 560-564.

Maxam, A. M. & Gilbert, W. (1980) Methods Enzymol. **65**, 499-560.

Melton, D. A., Krieg, P. A., Rebagliati, M. R., Maniatis, T., Zinn, K. & Green, M. R. (1984) Nucl. Acids Res. **12**, 7035-7056.

Messing, J., Crea, R. & Seeburg, P. H. (1981) Nucl. Acids Res. **9**, 309-321.

Michel, F. & Dujon, B. (1983) EMBO J. **2**, 33-38.

Michel, F. & Jacquier, A. (1987) Cold Spring Harbor Symp. Quant. Biol. **52**, 201-212.

Miller, D. L. & Martin, N. C. (1983) Cell **34**, 911-917.

Minami, E. & Watanabe, A. (1984) Arch. Biochem. Biophys. **235**, 562-570.

Minami, Y., Wakabayashi, S., Imoto, S., Ohta, Y. & Matsubara, H. (1985a) J. Biochem. **98**, 649-655.

Minami, Y., Wakabayashi, S., Wada, K., Matsubara, H., Kerscher, L. & Oesterhelt, D. (1985b) J. Biochem. **97**, 745-753.

Mizuno, T., Chou, M.-Y. & Inouye, M. (1984) Proc. Natl. Acad. Sci. USA **81**, 1966-1970.

Monroy, G., Spencer, E. & Hurwitz, J. (1978) J. Biol. Chem. **253**, 4490-4498.

Montandon, P. E. & Stutz, E. (1983) Nucl. Acids Res. **11**, 5877-5892.

Morris, J. & Herrmann, R. G. (1984) Nucl. Acids Res. **12**, 2837-2850.

Mullet, J. E. & Klein, R. R. (1987) EMBO J. **6**, 1571-1579.

Myers, A. M., Grant, D. M., Rabert, D. K., Harris, E. H., Boynton, J. E. & Gillham, N. W. (1982) Plasmid **7**, 133-151.

Netzker, R., Köchel, H. G., Basak, N. & Küntzel, H. (1982) Nucl. Acids Res. **10**, 4783-4794.

Neumann, J. & Drechsler, Z. (1984) Proc. Natl. Acad. Sci. USA **81**, 2070-2074.

Oh-oka, H., Takahashi, Y., Wada, K., Matsubara, H., Ohyama, K. & Ozeki, H. (1987) FEBS Lett. **218**, 52-54.

Ohta, Y., Katoh, K. & Miyake, K. (1977) Planta **136**, 229-232.

Ohtani, T., Uchimiya, H., Kato, A., Harada, H., Sugita, M. & Sugiura, M. (1984) Mol. Gen. Genet. **195**, 1-4.

Ohyama, K., Wetter, L. R., Yamano, Y., Fukuzawa, H. & Komano, T. (1982) Agric. Biol. Chem. **46**, 237-242.

Ohyama, K., Yamano, Y., Fukuzawa, H., Komano, T., Yamagishi, H., Fujimoto, S., & Sugiura, M. (1983) Mol. Gen. Genet. 189, 1-9.

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H., & Ozeki, H. (1986) Nature 322, 572-574.

Ohyama, K., Fukuzawa, H., Kohchi, T., Sano, T., Sano, S., Shirai, H., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H. & Ozeki, H. (1988a) J. Mol. Biol., 203, 281-298.

Ohyama, K., Kohchi, T., Fukuzawa, H., Sano, T., Umesono, K. & Ozeki, H. (1988b) Photosynthetic Res. 16, 7-22.

Ohyama, K., Kohchi, T., Sano, T. & Yamada, Y. (1988c) Trend. Biochem. Sci. 13, 19-22.

Ono, K. (1973) Jpn. J. Genet. 48, 69-70.

Ono, K. (1976) Jpn. J. Genet. 51, 11-18.

Ono, K., Ohyama, K. & Gamborg, O. L. (1979) Plant Sci. Lett. 14, 225-229.

Osawa, S. & Jukes, T. H. (1988) Trends in Genetics 4, 191-198.

Overbeeke, N., Haring, M. A., John, H., Nijkamp, J. & Kool, A. J. (1984) Plant Mol. Biol. 3, 235-241.

Ozeki, H., Ohyama, K., Inokuchi, H., Fukuzawa, H., Kohchi, T., Sano, T., Nakahigashi, K. & Umesono, K. (1987) Cold Spring Harbor Symp. Quant. Biol. 52, 791-804.

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S. & Sharp, P. A. (1986) Ann. Rev. Biochem. 55, 1119-1150.

Palmer, J. D. (1985) Ann. Rev. Genet. 19, 325-354.

Palmer, J. D. & Stein, D. B. (1986) Curr. Genet. 10, 823-833.

Pon, C. L., Wittmann-Liebold, B. & Gualerzi, C. (1979) FEBS Lett. 101, 157-160.

Posno, M., van Vliet, A. & Groot, G. S. P. (1986) Nucl. Acids Res. 14, 3181-3195.

Post, L. E. & Nomura, M. (1980) J. Biol. Chem. 255, 4660-4666.

Potter, H. & Dressler, D. (1986) Gene 48, 229-239.

Premakumar, R., Lemos, E. M. & Bishop, P. E. (1984) Biochim. Biophys. Acta 797, 64-70.

Ravel-Chapuis, P., Heizmann, P. & Nigon, V. (1982) Nature 300, 78-81.

Robson, R., Woodley, P. & Jones, R. (1986) EMBO J. 5, 1159-1163.

Rochaix, J. D. & Malnoe, P. (1978) Cell 15, 661-670.

Rochaix, J. D., van Dillewijn, J. & Rahire, M. (1984) Cell 36, 925-931.

Rock, C. D., Barkan, A. & Taylor, W. C. (1987) Curr. Genet. 12, 69-77.

Rosen, J., Ryder, T., Ohtsubo, H. & Ohtsubo, E. (1981) Nature 290, 794-797.

Sager, R. & Ishida, M. R. (1963) Proc. Natl. Acad. Sci. USA 50, 725-730.

Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 985-1002.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977a) Proc. Natl. Acad. Sci. USA 74, 5463-5467.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M. & Smith, M. (1977b) Nature 265, 687-695.

Sasaki, Y., Tomoda, Y., Tomi, H., Kamikubo, T. & Shinozaki, K. (1985)

Eur. J. Biochem. 152, 179-186.

Schlunegger, B. & Stutz, E. (1984) Curr. Genet. 8, 629-634.

Schmelzer, C. & Muller, M. W. (1987) Cell 51, 753-762.

Schmelzer, C., Schmidt, C., May, K. & Schweyen, R. J. (1983) EMBO J. 2, 2047-2052.

Schwarz, E., Scherer, G., Hobom, G. & Kössel, H. (1978) Nature 272, 410-414.

Schwarz, Z. & Kössel, H. (1980) Nature 283, 739-742.

Schwarz, Z., Jolly, S. O., Steinmetz, A. A. & Bogorad, L. (1981) Proc. Natl. Acad. Sci. USA 78, 3423-3427.

Sharp, P. A. (1987) Cell 50,147-148.

Shinozaki, K., Deno, H., Sugita, M., Kuramitsu, S. & Sugiura, M. (1986a) Mol. Gen. Genet. 202, 1-5.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B. Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N. & Sugiura, M. (1986b) EMBO J. 5, 2043-2049.

Solnick, D. (1985) Cell, 42, 157-164.

Stark, B. C., Kole, R., Bowman, E. J. & Altman, S. (1978) Proc. Natl. Acad. Sci. USA 75, 3717-3721.

Stein, D. B., Palmer, J. D. & Thompson, W. F. (1986a) Curr. Genet. 10, 835-841.

Stern, D. B., Bang, A. G. & Thompson, W. F. (1986b) Curr. Genet. 10, 857-869.

Strittmatter, G., Gozdzicka-Jozefiak, A. & Kössel, H. (1985) EMBO J. 4, 599-604.

Sugita, M., Kato, A., Shimada, H. & Sugiura, M. (1984) Mol. Gen. Genet. 194, 200-205.

Takaiwa, F. & Sugiura, M. (1982a) Eur. J. Biochem. 124, 13-19.

Takaiwa, F. & Sugiura, M. (1982b) Nucl. Acids Res. 10, 2665-2676.

Takata, R., Mukai, T., Aoyagi, M. & Hori, K. (1984) Mol. Gen. Genet. 197, 225-229.

Tanaka, A. (1984) PhD. thesis, Kyoto University.

Tanaka, M., Nakashima, T., Benson, A., Mower, H. & Yasunobu, K. T. (1966) Biochemistry 5, 1666-1681.

Tanaka, M., Haniu, M., Yasunobu, K. T., Evans, M. C. W. & Rao, K. K. (1974) Biochemistry 13, 2953-2959.

Tanaka, M., Obokata, J., Chunwongse, J., Shinozaki, K. & Sugiura, M. (1987) Mol. Gen. Genet. 209, 427-431.

Tinoco, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. & Gralla, J. (1973) Nature New Biology 246, 40-41.

Tohdoh, N., Shinozaki, K. & Sugiura, M. (1981) Nucl. Acids Res. 9, 5399-5406.

Tohdoh, N. & Sugiura, M. (1982) Gene 17, 213-218.

Tomioka, N. & Sugiura, M. (1983) Mol. Gen. Genet. 191, 46-50.

Tomizawa, J. & Itoh, T. (1981) Proc. Natl. Acad. Sci. USA 78, 6096-6100.

Tomizawa, J., Itoh, T., Selzer, G. & Som, T. (1981) Proc. Natl. Acad.

Sci. USA 78, 1421-1425.

Torazawa, K., Hayashida, N., Obokata, J., Shinozaki, K. & Sugiura, M. (1986) Nucl. Acids Res. 14, 3143.

Umesono, K., Inokuchi, H., Ohyama, K. & Ozeki, H. (1984) Nucl. Acids Res. 12, 9551-9565.

Umesono, K. & Ozeki, H. (1987) Trends in Genetics 3, 281-287.

Umesono, K., Inokuchi, H., Shiki, Y., Takeuchi, M., Chang, Z., Fukuzawa, H., Kohchi, T., Shirai, H., Ohyama, K. & Ozeki, H. (1988) J. Mol. Biol. 203, 299-332.

Vallet, J. M., Rahire, M. & Rochaix, J. D. (1984) EMBO J. 3, 415-421.

Wada, A. & Sako, T. (1987) J. Biochem. 101, 817-820.

Waddell, J., Wang, X. M. & Wu, M. (1984) Nucl. Acids Res. 12, 3843-3856.

Wang, X. M., Chang, C. H., Waddell, J. & Wu, M. (1984) Nucl. Acids Res. 12, 3857-3872.

Ward, J. H. Jr. (1963) J. Amer. Stat. Ass. 58, 236-244.

Westhoff, P., Alt, J. & Herrmann, R. G. (1983) EMBO J. 2, 2229-2237.

Westhoff, P. (1985) Mol. Gen. Genet. 201, 115-123.

Westhoff, P., Farchaus, J. W. & Herrmann, R. G. (1986) Curr. Genet. 11, 165-169.

Whitfeld, P. R. & Bottomley, W. (1983) Ann. Rev. Plant Physiol. 34, 279-310.

Widger, W. R., Cramer, W. A., Herrmann, R. G. & Trebst, A. (1984) Proc. Natl. Acad. Sci. USA 81, 674-678.

Wilbur, W. J. & Lipman, D. J. (1983) Proc. Natl. Acad. Sci. USA 80, 726-730.

Williams, J. G. & Mason, P. J. (1985) In: Hames, B. D. & Higgins, S. J. (eds) Nucleic acid hybridization pp 139-160 IRL Press, Oxford.

Wong, T. W. & Clayton, D. A. (1986) Cell 45, 817-825.

Wu, M., Lou, J. K., Chang, D. Y., Chang, C. H. & Nie, Z. Q. (1986) Proc. Natl. Acad. Sci. USA 83, 6761-6765.

Wu, M., Kong, X. F., & Kung, S. D. (1987) Curr. Genet. 10, 819-822.

Yamano, Y., Ohyama, K. & Komano, T. (1984) Nucl. Acids Res. 12, 4621-4624.

Yamano, Y., Kohchi, T., Fukuzawa, H., Ohyama, K. & Komano, T. (1985) FEBS Lett. 185, 203-207.

Yamano, Y. (1985b) PhD. thesis, Kyoto University.

Yanisch-Perron, C., Vieira, J. & Messing, J. (1985) Gene 33, 103-119.

Yanofsky, C. (1981) Nature 289, 751-758.

Yasunobu, K. T. & Tanaka, M. (1980) Methods in Enzymol. 69, 228-238.

Youvan, D. C., Bylina, E. J., Alberti, M., Begusch, H. & Hearst, J. E. (1984) Cell 37, 949-957.

Zaita, N., Torazawa, K., Shinozaki, K. & Sugiura, M. (1987) FEBS Lett. 210, 153-156.

Zeevi, M., Nevins, J. R. & Darnell, J. E. Jr. (1981) Cell 26, 39-46.

Zurawski, G., Bottomley, W. & Whitfeld, P. R. (1982) Proc. Natl. Acad. Sci. USA 79, 6260-6264.

Zurawski, G., Bottomley, W. & Whitfeld, P. R. (1984) Nucl. Acids Res. 12, 6547-6558.

Zurawski, G. & Zurawski, S. M. (1985) Nucl. Acids Res. 13, 4521-4526.

Zurawski, G., Bottomley, W. & Whitfeld, P. R. (1986) Nucl. Acids Res. 14, 3974.

Zyskind, J. W. & Smith, D. W. (1980) Proc. Natl. Acad. Sci. USA 77, 2460-2464.

SUMMARY

CHAPTER I Gene organization of the liverwort chloroplast genome

I-1 Structure and organization of *Marchantia polymorpha* chloroplast genome --- Cloning and gene identification

The complete nucleotide sequence of chloroplast DNA from a liverwort, *M. polymorpha*, was determined using a clone bank of chloroplast DNA fragments. The circular genome consisted of 121,025 base pairs (bp) and includes two large inverted repeats (IR_A and IR_B, each 10,058 bp), a large single-copy region (LSC, 81,096 bp), and a small single-copy region (SSC, 19,813 bp). The nucleotide sequence was analyzed with a computer to deduce the entire gene organization, assuming the universal genetic code and the presence of introns in the coding sequences. 136 possible genes were detected, 103 gene products of which are related to known stable RNA or protein molecules, as reported in brief earlier (Ohyama *et al.*, 1986). Stable RNA genes for four species of ribosomal RNA and 32 species of transfer RNA (tRNA) were located, although one of the tRNA genes may be defective. Twenty genes encoding polypeptides involved in photosynthesis and electron transport were identified by comparison with known chloroplast genes. Twenty-five open reading frames (ORFs) showed structural similarities to *E. coli* RNA polymerase subunits, 19 ribosomal proteins and two related ones. Seven ORFs were comparable with human mitochondrial NADH dehydrogenase genes. A computer-aided homology search predicted possible chloroplast homologues of bacterial proteins; two ORFs for bacterial 4Fe-4S-type ferredoxin, two for distinct subunits of a protein-dependent transport system, one ORF for a component of nitrogenase, and one for an antenna protein of a light harvesting complex. The other 33 ORFs consisting of 29 to 2,136 codons remain to be identified, but some of them seem to be conserved in evolution. There were 22 introns in 20 genes (8 tRNA genes and 12 ORFs), which may be classified into the groups I and II found in fungal mitochondrial genes. The structural gene for ribosomal protein S12 is trans-split on the opposite DNA strand (Fukuzawa et al., 1986). The universal genetic code was confirmed by the substitution pattern of simultaneous codons and also by possible codon recognition of the chloroplast-encoded tRNA molecules, assuming no importation of tRNA molecules from the cytoplasm. The nucleotide residue A or T was preferred at the third position of the codons (G + C, 11.9%) and in intergenic spacers (G + C, 19.5%), resulting in an overall G + C content that was low (28.8%) throughout the liverwort chloroplast genome. Possible gene expression signals such as promoters and terminators for transcription, and predicted locations of gene products were discussed.

I-2 Structure and organization of Marchantia polymorpha chloroplast genome --- Inverted repeat and small single copy regions

The genes in the regions of large inverted repeats (IR_A and IR_B, 10,058 base pairs each) and a small single copy (SSC, 81,095 bp) of chloroplast DNA from M. polymorpha were characterized. The IR regions contained genes for four ribosomal RNAs (16S, 23S, 4.5S, and 5S rRNAs) and five transfer RNAs (valine $tRNA_{GAC}$, isoleucine $tRNA_{GAU}$, alanine $tRNA_{UGC}$, arginine $tRNA_{ACG}$, and asparagine $tRNA_{GUU}$). The gene organization of the IR regions in the liverwort chloroplast genome was conserved, although the IR regions were smaller (10,058 bp) than any reported in higher plant chloroplasts. The SSC region (19,813 bp) encoded genes for 17 open reading frames (ORFs), a leucine $tRNA_{UAG}$, and a proline $tRNA_{GGG}$-like sequence. Twelve ORFs were identified by homology of their coding sequences with a 4Fe-4S-type ferredoxin protein, a bacterial nitrogenase reductase component (Fe-protein), five human mitochondrial components of NADH dehydrogenase (ND1, ND4, ND4L, ND5, and ND6), two E. coli ribosomal proteins (S15 and L21), two putative proteins encoded in the kinetoplast maxicircle DNA of L. tarentolae (LtORF 3 and LtORF 4), and a bacterial permease inner membrane component (encoded by malF in E. coli or hisQ in S. typhimurium).

CHAPTER II Divergent mRNA transcription in the chloroplast psbB operon

The genes psbB, psbH, petB, and petD for the components in photosystem II and the cytochrome b6/f complex are clustered and co-transcribed in liverwort M. polymorpha chloroplasts. On the opposite DNA strand in the spacer region between the genes psbB and psbH, an open reading frame consisting of 43 sense codons was deduced and designated as the ORF43 gene. The ORF43 gene was actively transcribed in liverwort chloroplasts. The ORF43 transcripts were entirely complementary to a part of the primary transcripts of the psbB operon. Heterogeneous Northern hybridization showed that the mRNA transcripts for the ORF43 gene increased with the greening in pea seedlings. This is the first demonstration of a divergent overlapping transcription in chloroplasts.

CHAPTER III Ordered processing and splicing in a polycistronic transcript in liverwort chloroplast

From the complete sequence of the chloroplast DNA in a liverwort, M. polymorpha, an unidentified open reading frame, ORF203, was found between the psbB and rps12' (trans-split) genes. ORF203 was a split gene consisting of three exons and two group II introns. Multiple transcripts for ORF203 were detected on Northern blots of the chloroplast RNA preparation. The ORF203 locus was primarily co-transcribed with the downstream genes rps12' and rpl20, and then processed into a monomeric precursor. S1 nuclease mapping gave the transcription initiation site 52 nucleotides upstream from the coding sequence of ORF203. The spliced RNA molecules were identified, as predicted, by the use of synthetic oligodeoxyribonucleotide probes specific to ligated exon sequences. The splicing reaction proceeded successively from the 5' to 3' direction. These results indicate that ordered RNA processing occurs in the chloroplasts of land plants. Trans-membrane analysis by a computer indicated that ORF203 gene product could be associated with a chloroplast membrane.

CHAPTER IV Nicked group II intron and its trans-splicing in the liverwort chloroplasts

The chloroplast gene *rps12* for ribosomal protein S12 in a liverwort, *M. polymorpha*, is split into three exons by two introns, one of which (intron 1) is discontinuous. Exon 1 of *rps12* for the N-terminal portion of the S12 protein is far from exons 2 and 3 for the C-terminal portion on the opposite DNA strand. S1-nuclease protection analysis and Northern hybridization with RNA isolated from the liverwort chloroplasts showed that: (i) the exons 1 and 2-3 of the *rps12* gene with the neighboring genes were transcribed separately, (ii) the *trans*-splicing of intron 1 occurred after the processing of two primary transcripts to two pre-mRNAs, and (iii) there was no particular order for the splicing of intron 1 (*trans*) and intron 2 (*cis*) in the *rps12* gene. A bimolecular interaction model was proposed for *trans*-splicing by assuming that intermolecular base pairings between two pre-mRNAs result in the formation of the structure typical of group II introns except for disruption in the loop III region. This structure could be constructed in intron 1 of tobacco *rps12* gene.

LIST OF PUBLICATIONS

(A) Yamano, Y., Kohchi, T., Fukuzawa, H., Ohyama, K. & Komano, T. (1985). Nucleotide sequences of chloroplast 4.5S ribosomal RNA from a leafy liverwort, *Jungermannia subulata*, and a thalloid liverwort, *Marchantia polymorpha*. *FEBS Letters*, **185**:203-207.

(B) Fukuzawa, H., Kohchi, T., Shirai, H., Ohyama, K., Umesono, K., Inokuchi, H. & Ozeki, H. (1986). Coding sequences for chloroplast ribosomal protein S12 from the liverwort, *Marchantia polymorpha*, are separated far apart on the different DNA strands. *FEBS Letters*, **198**:11-15.

(C) Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Inokuchi, H. & Ozeki, H. (1986). Chloroplast gene organization from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*, **322**:572-574.

(D) Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Inokuchi, H. & Ozeki, H. (1986). Complete nucleotide sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Plant Mol. Biol. Reporter*, **4**:148-175.

(E) Fukuzawa, H., Yoshida, T., Kohchi, T., Okumura, T., Sawano, Y. & Ohyama, K. (1987). Splicing of group II introns in mRNAs coding for cytochrome b6 and subunit IV in liverwort *Marchantia polymorpha* chloroplast genome: Exon specifying a region coding for two genes with the spacer. *FEBS Letters*, **220**:61-66.

(F) Ozeki, H., Ohyama, K., Inokuchi, H., Fukuzawa, H., Kohchi, T., Sano, T., Nakahigashi, K. & Umesono, K. (1987). Genetic system of chloroplasts. *Cold Spring Harbor Symp. Quant. Biol.,* 52:791-804.

(G) Kohchi, T., Yoshida, T., Komano, T. & Ohyama, K. (1988). Divergent mRNA transcription in the chloroplast *psbB* operon. *EMBO J.,* 7:885-891.

(H) Ohyama, K., Kohchi, T., Sano, S. & Yamada, Y. (1988). Newly identified groups of genes in chloroplasts. *Trend. Biochem. Sci.,* **13**:19-22.

(I) Ohyama, K., Kohchi, T., Fukuzawa, H., Sano, T., Umesono, K. & Ozeki, H. (1988). Gene organization and newly identified groups of genes of the chloroplast genome from a liverwort, *Marchantia polymorpha*. *Photosynthesis Res.,* **16**: 7-22.

(J) Kohchi, T., Ogura, Y., Umesono, K., Yamada, Y., Komano, T., Ozeki, H. & Ohyama, K. (1988). Ordered processing and splicing in a polycistronic transcript in liverwort chloroplast. *Curr. Genet.,* **14**:147-154.

(K) Ohyama, K., Fukuzawa, H., Kohchi, T., Sano, T., Sano, S., Shirai, H., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H. & Ozeki, H. (1988). Structure and organization of *Marchantia polymorpha* chloroplast genome. I. Cloning and gene identification. *J. Mol. Biol.,* **203**:281-298.

(L) Umesono, K., Inokuchi, H., Shiki, Y., Takeuchi, M., Chang, Z., Fukuzawa, H., Kohchi, T., Shirai, H., Ohyama, K. & Ozeki, H. (1988). Structure and organization of *Marchantia polymorpha* chloroplast genome. II. Gene organization of the large single-copy region from *rps'12* and *atpB*. *J. Mol. Biol.,* 203:299-332.

(M) Fukuzawa, H., Kohchi, T., Sano, T., Shirai, H., Umesono, K., Inokuchi, H., Ozeki, H. & Ohyama, K. (1988). Structure and organization of *Marchantia polymorpha* chloroplast genome. III. Gene organization of the large single copy region from *rbcL* to *trnI*(CAU). *J. Mol. Biol.,* 203:333-351.

(N) Kohchi, T., Shirai, H., Fukuzawa, H., Sano, T., Komano, T., Umesono, K., Inokuchi, H., Ozeki, H. & Ohyama, K. (1988). Structure and organization of *Marchantia polymorpha* chloroplast genome. IV. Inverted repeat and small single copy regions. *J. Mol. Biol.,* 203:353-372.

(O) Kohchi, T., Umesono, K., Ogura, Y., Komine, Y., Nakahigashi, K., Komano, T., Yamada, Y., Ozeki, H. & Ohyama, K. (1988). A nicked group II intron and *trans*-splicing in liverwort, *Marchantia polymorpha,* chloroplasts. *Nucl. Acids Res.,* 16:10025-10036.

Chapter I-1 is described in reference (K).
Chapter I-2 is described in reference (N).
Chapter II is described in reference (G).
Chapter III is described in reference (J).
Chapter IV is described in reference (O).

ACKNOWLEDGMENT

The author wishes to express his sincere thanks to Professor Tohru Komano, Department of Agricultural Chemistry, Faculty of Agriculture, Kyoto University, for his kind guidance and continuous encouragement throughout this work.

The author greatly appreciates to Associate Professor Kanji Ohyama, Research Center for Cell and Tissue Culture, Faculty of Agriculture, Kyoto University, for his valuable suggestion, discussion, critical reading of the manuscript, and warm encouragement in carrying out this study.

Special thanks must be given to Professor Haruo Ozeki, Associate Professor Hachiro Inokuchi, and Dr. Kazuhiko Umesono, Department of Biophysics, Faculty of Science, Kyoto University, for their thoughtful suggestion and collaboration during this work.

The author is grateful to Professor Yasuyuki Yamada, Research Center for Cell and Tissue Culture, Faculty of Agriculture, Kyoto University, for his strong encouragement. The author thanks to Dr. Yukiko Sasaki, Department of Food Science and Technology, Faculty of Agriculture, Kyoto University, for her kind permission to use her light-induced RNA and for her valuable discussion.

The author is indebted to Dr. Hideya Fukuzawa, Institute for Applied Microbiology, University of Tokyo; Dr. Yosiaki Yamano, and all the other members of Laboratory of Biochemistry, Department of Agricultural Chemistry; Mr. Hiromasa Shirai, Mr. Satoshi Sano, Mr. Tohru Sano, Mr. Yutaka Ogura, Mr. Tadashi Yoshida, and all the other members of Laboratories of Plant Molecular Biology and Fundamental Cell Technology, Research Center for Cell and Tissue Culture, Faculty of Agriculture, Kyoto University.

March, 1989

Takayuki Kohchi

Takayuki Kohchi