© 1991 Oxford University Press

Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes

Hiroaki Shimada and Masahiro Sugiura[1]*
Mitsui Plant Biotechnology Research Institute, 2-1-6 Sengen, Tsukuba 305 and [1]Center for Gene Research, Nagoya University, Nagoya 464-01, Japan

Received December 21, 1990; Accepted January 31, 1991 EMBL accession nos X15901, X04465, Z00044

ABSTRACT

The entire nucleotide sequences of the rice, tobacco and liverwort chloroplast genomes have been determined. We compared all the chloroplast genes, open reading frames and spacer regions in the plastid genomes of these three species in order to elucidate general structural features of the chloroplast genome. Analyses of homology, GC content and codon usage of the genes enabled us to classify them into two groups: photosynthesis genes and genetic system genes. Based on comparisons of homology, GC content and codon usage, unidentified ORFs can also be assigned to each of these groups such that it is possible to speculate about the functions of products which may be produced by these ORFs. The spacer regions and intron sequences were compared and found to have no obvious homology between rice and liverwort or between tobacco and liverwort.

INTRODUCTION

Chloroplasts are intracellular organelles which contain their own genetic systems and a number of chloroplast components are encoded in their genomes (1,2). Most chloroplast genomes in land plants consist of homogeneous circular DNA molecules which range in size from 120 to 160 kbp. Determination and comparison of corresponding DNA sequences from a variety of chloroplast genomes reveal more accurate information on chloroplast gene structure and on the evolution of these genes. Since the entire nucleotide sequences of the chloroplast DNAs from tobacco, liverwort and rice have been determined (3–5), the structure of all chloroplast genes, open reading frames (ORFs) and spacer regions in these three species can be analyzed at the nucleotide sequence level. We made a complete sequence comparison of these three genomes. We have grouped the chloroplast genes and ORFs into two groups based on homologies, GC contents and codon usages. This classification scheme enables us to predict the function of unidentified gene products from ORFs present in the chloroplast genome.

MATERIALS AND METHODS

The entire nucleotide sequences of the rice (*Oryza sativa* L., 134,525 bp) and tobacco (*Nicotiana tabacum*, 155,844 bp) chloroplast genomes (accession numbers x15901 and z00044, respectively, in the EMBL data library) were determined in our laboratory (3,5). The nucleotide sequence of the liverwort chloroplast genome (*Marchantia polymorpha*; 121,024 bp) was obtained from the GenBank database (release 58.0). Computer-aided analysis of nucleotide sequences was carried out using the GENETYX program (Software Development Co., Japan) on an NEC PC9801 computer.

RESULTS AND DISCUSSION

RNA genes

As with all higher plant chloroplast genomes examined to date, four kinds of rRNA genes, 23S, 16S, 5S and 4.5S rDNA are present in the rice, tobacco and liverwort chloroplast genomes. The deduced rRNA sequences of these four genes strongly resemble the corresponding counterparts among these three plant species (Table 1). Among them, the rice 23S rDNA is larger than the other two because an extra 68 bp sequence is present in the middle of the gene. This extra sequence is flanked by short direct GTA repeats and has a significant level of homology to the region immediately upstream of the 23S rDNA sequence (data not shown). This suggests that the 68 bp sequence originated by a duplication of the upstream region. A similar sequence has also been reported in the maize chloroplast 23S rDNA (6), so it seems that this extra 68 bp sequence could be specific in monocots.

Chloroplast tRNA genes (*trn*s) are highly conserved (more than 80% homologous) between rice, tobacco and liverwort (Table 1). A unique tRNA gene, *trnR*-CCG, is present only in liverwort (4). Four pseudogenes corresponding to tRNA genes, Ψ*trnfM/G*, Ψ*trnG*, Ψ*trnT* and Ψ*trnE*, have been identified in rice. These are proposed to have resulted from multiple genome rearrangements, which must have occurred more than four times if the initial rice chloroplast genome was similar in structure to that of tobacco (7). Similar genome rearrangements and the presence of pseudogenes have also been reported in wheat (8,9).

* To whom correspondence should be addressed

Table 1. RNA genes in rice, tobacco and liverwort chloroplasts

Gene	Gene product	Number of bp			Homology(%)	
		Rice	Tobacco	Liverwort	T/R	L/R
Ribosomal RNA genes						
23S rDNA	23S rRNA	2884	2810	2811	94	91
16S rDNA	16S rRNA	1491	1489	1496	97	94
5S rDNA	5S rRNA	121	121	119	97	89
4.5S rDNA	4.5S rRNA	95	103	103	83	82
transfer RNA genes						
trnA-UGC*	Ala-tRNA(UGC)	73	73	73	100	97
trnR-ACG	Arg-tRNA(ACG)	74	74	74	100	96
trnR-CCG	Arg-tRNA(CCG)	–	–	74	–	–
trnR-UCU	Arg-tRNA(UCU)	72	72	72	100	89
trnN-GUU	Asn-tRNA(GUU)	72	72	72	97	92
trnD-GUC	Asp-tRNA(GUC)	74	74	74	100	93
trnC-GCA	Cys-tRNA(GCA)	71	72	71	89	93
trnQ-UUG	Gln-tRNA(UUG)	72	72	72	96	93
trnE-UUC	Glu-tRNA(UUC)	73	73	73	96	89
trnG-GCC	Gly-tRNA(GCC)	71	71	71	82	83
trnG-UCC*	Gly-tRNA(UCC)	72	71	70	97	94
trnH-GUG	His-tRNA(GUG)	75	75	75	100	92
trnI-CAU	Ile-tRNA(CAU)	74	74	74	92	91
trnI-GAU*	Ile-tRNA(GAU)	72	72	72	100	99
trnL-CAA	Leu-tRNA(CAA)	81	81	80	98	94
trnL-UAA*	Leu-tRNA(UAA)	85	85	85	97	87
trnL-UAG	Leu-tRNA(UAG)	80	80	80	96	93
trnK-UUU*	Lys-tRNA(UUU)	72	72	72	97	97
trnfM-CAU	fMet-tRNA(CAU)	73	73	74	92	88
trnM-CAU	Met-tRNA(CAU)	73	73	74	97	96
trnF-GAA	Phe-tRNA(GAA)	73	73	73	97	96
trnP-UGG	Pro-tRNA(UGG)	74	74	74	99	96
trnS-GCU	Ser-tRNA(GCU)	88	88	88	94	94
trnS-GGA	Ser-tRNA(GGA)	87	87	88	97	89
trnS-UGA	Ser-tRNA(UGA)	88	92	88	91	85
trnT-GGU	Thr-tRNA(GGU)	72	72	72	96	94
trnT-UGU	Thr-tRNA(UGU)	73	73	73	93	85
trnW-CCA	Trp-tRNA(CCA)	74	74	74	99	93
trnY-GUA	Tyr-tRNA(GUA)	84	84	82	99	59
trnV-GAC	Val-tRNA(GAC)	72	72	72	100	97
trnV-UAC*	Val-tRNA(UAC)	74	73	72	97	92

*:genes containing introns

No similar pseudogenes have been found in the corresponding regions of the tobacco or liverwort chloroplast genome. However, unique pseudogenes corresponding to *trnR*-UCG and *trnP*-GGG have been reported in tobacco and liverwort, respectively (4,10).

Polypeptide genes and conserved ORFs

The chloroplast genome contains over 62 genes which encode polypeptides (including putative genes). The polypeptides encoded by these genes include ribosomal proteins, RNA polymerase subunits, photosystem components and polypeptides homologous to the mitochondrial NADH dehydrogenase subunits, among others. Among ORFs found in the three chloroplast genomes studied, those with similar sizes and which are located in the corresponding regions of the three genomes are designated as conserved ORFs. The sizes and homologies of the predicted translation products are listed and compared between rice and tobacco and between rice and liverwort in Table 2.

The gene encoding a 30S ribosomal protein 16, *rps16*, is present in the same region of the rice and tobacco chloroplast genomes. (It is also present in all other angiosperms analyzed to date). The corresponding region in the liverwort genome contains ORF513 which has no homology with *rps16*, suggesting that ORF513 and *rps16* are of different origins. The gene for a unique 50S ribosomal protein, *rpl21*, is present in the liverwort genome and no corresponding sequence occurs in the rice or tobacco genomes.

A gene encoding an initiation factor, *infA*, appears to be incomplete in the tobacco genome and lacks a known translation initiation codon. The chloroplast genomes of other higher plants such as spinach have been reported to contain an *infA* reading frame starting from ATG (11). Therefore it seems that the tobacco gene may have lost a portion of its sequence and may thus be a pseudogene.

The following gene have listed in Table 2 as NADH dehydrogenase (*ndh*) sequences because they have some relation or homology to the mitochondrial NADH dehydrogenase genes (12–15): *psbG*, ORF393/393/392 , ORF178/167/*frxB* and ORF159/158/169 (ORFs are listed as rice/tobacco/liverwort, respectively).

A conserved ORF, ORF63/55/69, in the small-single copy region has recently been identified as the gene encoding the 50S ribosomal protein CS32, (gene name: *rpl32*), which was purified from tobacco chloroplasts and subjected to amino acid sequence determination (16). The corresponding ORFs in rice and liverwort good homology to tobacco *rpl32* gene, especially in the N-terminal region.

ORF2280 and ORF2136 are related and are present in the tobacco and liverwort chloroplast genomes, respectively. The

Table 2. Polypeptide genes and conserved ORFs in rice, tobacco and liverwort chloroplasts

Genes	Gene products	Number of codons			Homology(%)		Remarks
		Rice	Tobacco	Liverwort	T/R	L/R	
			Genetic System Genes				
30S ribosomal proteins							
rps2	CS2	236	236	235	79	65	
rps3	CS3	239	218	217	79	50	
rps4	CS4	201	201	202	80	70	
rps7	CS7	156	155	155	85	71	
rps8	CS8	136	134	132	75	59	
rps11	CS11	143	138	130	68	63	
*rps12**	CS12	124	123	123	89	86	1)
rps14	CS14	103	100	100	85	73	
rps15	CS15	78	87	88	80	53	
*rps16**	CS16	86	86	–	90	–	
rps18	CS18	168	101	75	70	66	
rps19	CS19	93	92	92	68	65	
50S ribosomal proteins							
*rpl2**	CS2	273	274	277	90	69	
rpl14	CS14	123	123	122	83	77	
*rpl16**	CS16	136	134	143	88	79	
rpl20	CS20	119	128	116	69	52	
rpl21	CS21	–	–	116	–	–	
rpl22	CS22	103	155	119	61	55	
rpl23	CS23	93	93	91	85	59	
rpl32	CS32	63	55	69	70	65	
rpl33	CS33	66	66	65	73	59	
rpl36	CS36	37	37	37	92	87	
RNA polymerase subunits							
rpoA	alpha	337	337	340	69	52	
rpoB	beta	1075	1070	1065	81	64	
*rpoC1**	beta′	682	687	684	78	62	
rpoC2	beta″	1514	1392	1386	64	47	
Initiation factor							
infA	IF1	107	(96)	78	66	57	2)
			NADH Dehydrogenase Genes				
*ndhA**	ND1	362	364	368	76	70	4)
*ndhB**	ND2	510	387	501	96	68	4)
ndhC	ND3	120	120	120	87	72	4)
ndhD	ND4	500	509	499	82	72	4)
ndhE	ND4L	101	101	100	83	67	4)
ndhF	ND5	734	710	692	67	53	4)
ndhG	ND6	176	176	191	76	55	4)
psbG(ndhI)		246	284	243	82	64	5)
ORF393/393/392(*ndhH*)		393	393	392	89	83	5)
ORF178/167/*frxB*		178	167	183	81	78	5)
ORF159/158/169		159	158	169	85	72	
			Photosynthesis Genes				
Ribulose 1,5-diphosphate carboxylase/oxygenase							
rbcL	Large subunit	477	477	475	93	92	
Photosystem I							
psaA	P700 (A1)	750	750	750	96	91	
psaB	P700 (A2)	734	734	734	97	92	
psaC	9 kDa protein	81	81	81	95	93	
psaI	I protein	36	36	36	89	71	
psaJ	J protein	44	44	42	89	76	
Photosystem II							
psbA	D1	353	353	353	99	97	
psbB	47kDa protein	508	508	508	97	91	
psbC	43kDa protein	473	473	473	97	95	
psbD	D2	353	353	353	98	97	
psbE	b559(9 kDa)	83	83	83	98	88	
psbF	b559(4 kDa)	39	39	39	100	97	
psbH	10kDa protein	73	73	74	90	67	
psbI	I protein	36	52	36	97	94	
psb(J)	J protein(?)	40	40	40	90	85	3)
psbK	K protein	61	98	55	72	65	

Table 2. *(continued)*

Genes	Gene products	Number of codons			Homology(%)		Remarks
		Rice	Tobacco	Liverwort	T/R	L/R	
psbL	L protein	38	38	38	100	92	
psbM	M protein	34	34	34	100	88	
psbN	N protein	43	43	43	98	84	
Cytochromen b/f complex							
petA	cytochrome f	320	320	320	91	78	
*petB**	cytochrome b6	215	215	215	99	96	
*petD**	IV	160	160	160	99	94	
petG	V	37	37	36	100	87	
H⁺-ATPase subunit							
atpA	alpha	507	507	507	88	83	
atpB	beta	498	498	492	92	88	
atpE	epsilon	137	133	135	73	53	
*atpF**	I	180	184	184	79	53	
atpH	III	81	81	81	99	98	
atpI	IV	247	247	248	93	83	
			Conserved ORFs				
ORF29/29/29		29	29	29	100	86	
ORF31/31/31		31	31	31	90	77	
ORF35/34/35		35	34	35	100	89	
ORF62/62/62		62	62	62	87	81	
IRF170/168/167*		170	168	167	95	84	
ORF185/184/184		185	184	184	80	60	
ORF216/IRF196/203*		216	196	203	69	61	
ORF230/229/434		230	229	434	62	46	
ORF321/313/320		321	313	320	70	54	
ORF106/512/316		106	512	316	50	45	6)
ORF2280/2136		–	2280	2136	–	–	
ORF542/509/370		542	509	370	59	32	

Gene names are listed according to Hallick *et al.* (18,19). ORFs are shown listed as rice/tobacco/liverwort, respectively. Hyphens indicate genes or ORFs which are absent in a particular genome. IRF indicated an intron-containing reading frame. Asterisks indicate genes containing introns.

1) divided gene (20).
2) tobacco *infA* is assumed to be the largest reading frame.
3) see reference 19.
4) Liverwort *ndh* are called as *ndh1-6*. The *ndh* products were not identified.
5) see references 12–15.
6) There is local homology between 74 amino acids in the protein.

Table 3. Codon usages in rice, tobacco and liverwort chloroplasts.

(A) Rice genetic system genes

UUU-Phe	168	UCU-Ser	104	UAU-Tyr	170	UGU-Cys	58	UUG-Leu	134	UCG-Ser	46	UAG-***	9	UGG-Trp	155
UUC-Phe	84	UCC-Ser	93	UAC-Tyr	61	UGC-Cys	26	CUU-Leu	145	CCU-Pro	137	CAU-His	127	CGU-Arg	90
UUA-Leu	200	UCA-Ser	90	UAA-***	31	UGA-***	12	CUC-Leu	44	CCC-Pro	70	CAC-His	46	CGC-Arg	44
UUG-Leu	132	UCG-Ser	45	UAG-***	8	UGG-Trp	72	CUA-Leu	94	CCA-Pro	73	CAA-Gln	164	CGA-Arg	60
CUU-Leu	119	CCU-Pro	93	CAU-His	123	CGU-Arg	111	CUG-Leu	39	CCG-Pro	43	CAG-Gln	60	CGG-Arg	19
CUC-Leu	46	CCC-Pro	60	CAC-His	43	CGC-Arg	41	AUU-Ile	260	ACU-Thr	176	AAU-Asn	175	AGU-Ser	97
CUA-Leu	81	CCA-Pro	74	CAA-Gln	216	CGA-Arg	120	AUC-Ile	101	ACC-Thr	84	AAC-Asn	67	AGC-Ser	28
CUG-Leu	49	CCG-Pro	36	CAG-Gln	61	CGG-Arg	55	AUA-Ile	106	ACA-Thr	92	AAA-Lys	153	AGA-Arg	71
AUU-Ile	262	ACU-Thr	118	AAU-Asn	224	AGU-Ser	91	AUG-Met	156	ACG-Thr	30	AAG-Lys	52	AGG-Arg	24
AUC-Ile	98	ACC-Thr	62	AAC-Asn	81	AGC-Ser	37	GUU-Val	161	GCU-Ala	265	GAU-Asp	194	GGU-Gly	219
AUA-Ile	178	ACA-Thr	98	AAA-Lys	379	AGA-Arg	193	GUC-Val	51	GCC-Ala	76	GAC-Asp	56	GGC-Gly	79
AUG-Met	129	ACG-Thr	49	AAG-Lys	150	AGG-Arg	65	GUA-Val	178	GCA-Ala	152	GAA-Glu	242	GGA-Gly	190
GUU-Val	156	GCU-Ala	115	GAU-Asp	182	GGU-Gly	115	GUG-Val	70	GCG-Ala	64	GAG-Glu	86	GGG-Gly	118
GUC-Val	45	GCC-Ala	53	GAC-Asp	65	GGC-Gly	33								
GUA-Val	138	GCA-Ala	108	GAA-Glu	296	GGA-Gly	174								
GUG-Val	51	GCG-Ala	51	GAG-Glu	120	GGG-Gly	98								

AU: 4599
GC: 2099 (GC:31.3%)

AU: 4517
GC: 2149 (GC:32.2%)

(B) Rice photosynthesis genes

UUU-Phe	262	UCU-Ser	115	UAU-Tyr	155	UGU-Cys	35
UUC-Phe	155	UCC-Ser	86	UAC-Tyr	55	UGC-Cys	14
UUA-Leu	261	UCA-Ser	48	UAA-***	11	UGA-***	9

(C) Rice NADH dehydrogenase genes

UUU-Phe	159	UCU-Ser	70	UAU-Tyr	107	UGU-Cys	38
UUC-Phe	71	UCC-Ser	52	UAC-Tyr	30	UGC-Cys	5
UUA-Leu	157	UCA-Ser	56	UAA-***	4	UGA-***	1
UUG-Leu	57	UCG-Ser	18	UAG-***	4	UGG-Trp	63
CUU-Leu	106	CCU-Pro	50	CAU-His	30	CGU-Arg	19
CUC-Leu	30	CCC-Pro	24	CAC-His	13	CGC-Arg	12

Table 3. *(continued)*

CUA-Leu	67	CCA-Pro	37	CAA-Gln	64	CGA-Arg	26
CUG-Leu	23	CCG-Pro	11	CAG-Gln	17	CGG-Arg	9
AUU-Ile	145	ACU-Thr	71	AAU-Asn	84	AGU-Ser	46
AUC-Ile	61	ACC-Thr	22	AAC-Asn	27	AGC-Ser	18
AUA-Ile	112	ACA-Thr	45	AAA-Lys	72	AGA-Arg	34
AUG-Met	101	ACG-Thr	15	AAG-Lys	18	AGG-Arg	15
GUU-Val	74	GCU-Ala	77	GAU-Asp	83	GGU-Gly	66
GUC-Val	19	GCC-Ala	16	GAC-Asp	10	GGC-Gly	18
GUA-Val	51	GCA-Ala	49	GAA-Glu	103	GGA-Gly	104
GUG-Val	14	GCG-Ala	21	GAG-Glu	27	GGG-Gly	35

AU: 2207
GC: 876 (GC: 28.4%)

(D) Tobacco genetic system genes

UUU-Phe	139	UCU-Ser	126	UAU-Tyr	178	UGU-Cys	69
UUC-Phe	53	UCC-Ser	66	UAC-Tyr	29	UGC-Cys	15
UUA-Leu	188	UCA-Ser	79	UAA-***	20	UGA-***	1
UUG-Leu	125	UCG-Ser	40	UAG-***	3	UGG-Trp	58
CUU-Leu	112	CCU-Pro	94	CAU-His	132	CGU-Arg	109
CUC-Leu	35	CCC-Pro	57	CAC-His	34	CGC-Arg	31
CUA-Leu	62	CCA-Pro	70	CAA-Gln	180	CGA-Arg	144
CUG-Leu	30	CCG-Pro	40	CAG-Gln	61	CGG-Arg	45
AUU-Ile	278	ACU-Thr	107	AAU-Asn	224	AGU-Ser	81
AUC-Ile	97	ACC-Thr	55	AAC-Asn	65	AGC-Ser	24
AUA-Ile	176	ACA-Thr	116	AAA-Lys	324	AGA-Arg	152
AUG-Met	132	ACG-Thr	35	AAG-Lys	97	AGG-Arg	61
GUU-Val	129	GCU-Ala	120	GAU-Asp	185	GGU-Gly	127
GUC-Val	49	GCC-Ala	55	GAC-Asp	50	GGC-Gly	40
GUA-Val	137	GCA-Ala	110	GAA-Glu	262	GGA-Gly	199
GUG-Val	50	GCG-Ala	30	GAG-Glu	76	GGG-Gly	76

AU: 4430
GC: 1714 (GC:27.9%)

(E) Tobacco photosynthesis genes

UUU-Phe	272	UCU-Ser	125	UAU-Tyr	157	UGU-Cys	36
UUC-Phe	153	UCC-Ser	71	UAC-Tyr	50	UGC-Cys	12
UUA-Leu	259	UCA-Ser	61	UAA-***	13	UGA-***	7
UUG-Leu	146	UCG-Ser	29	UAG-***	9	UGG-Trp	153
CUU-Leu	146	CCU-Pro	145	CAU-His	127	CGU-Arg	105
CUC-Leu	42	CCC-Pro	46	CAC-His	46	CGC-Arg	26
CUA-Leu	98	CCA-Pro	83	CAA-Gln	176	CGA-Arg	65
CUG-Leu	46	CCG-Pro	46	CAG-Gln	61	CGG-Arg	14
AUU-Ile	239	ACU-Thr	170	AAU-Asn	182	AGU-Ser	97
AUC-Ile	104	ACC-Thr	100	AAC-Asn	69	AGC-Ser	30
AUA-Ile	112	ACA-Thr	87	AAA-Lys	159	AGA-Arg	67
AUG-Met	153	ACG-Thr	35	AAG-Lys	45	AGG-Arg	25
GUU-Val	166	GCU-Ala	272	GAU-Asp	190	GGU-Gly	250
GUC-Val	48	GCC-Ala	95	GAC-Asp	55	GGC-Gly	80
GUA-Val	196	GCA-Ala	153	GAA-Glu	244	GGA-Gly	193
GUG-Val	58	GCG-Ala	55	GAG-Glu	78	GGG-Gly	97

AU: 4642
GC: 2078 (GC: 30.9%)

(F) Tobacco NADH dehydrogenase genes

UUU-Phe	157	UCU-Ser	69	UAU-Tyr	122	UGU-Cys	34
UUC-Phe	60	UCC-Ser	35	UAC-Tyr	33	UGC-Cys	10
UUA-Leu	151	UCA-Ser	51	UAA-***	5	UGA-***	2
UUG-Leu	63	UCG-Ser	25	UAG-***	3	UGG-Trp	59
CUU-Leu	92	CCU-Pro	53	CAU-His	33	CGU-Arg	28
CUC-Leu	21	CCC-Pro	19	CAC-His	11	CGC-Arg	7
CUA-Leu	54	CCA-Pro	49	CAA-Gln	60	CGA-Arg	38
CUG-Leu	16	CCG-Pro	7	CAG-Gln	17	CGG-Arg	9
AUU-Ile	156	ACU-Thr	68	AAU-Asn	89	AGU-Ser	37
AUC-Ile	65	ACC-Thr	19	AAC-Asn	22	AGC-Ser	17
AUA-Ile	122	ACA-Thr	58	AAA-Lys	78	AGA-Arg	36
AUG-Met	100	ACG-Thr	12	AAG-Lys	18	AGG-Arg	8
GUU-Val	85	GCU-Ala	83	GAU-Asp	89	GGU-Gly	67
GUC-Val	15	GCC-Ala	32	GAC-Asp	20	GGC-Gly	21
GUA-Val	55	GCA-Ala	42	GAA-Glu	99	GGA-Gly	103
GUG-Val	15	GCG-Ala	20	GAG-Glu	23	GGG-Gly	37

AU: 2265
GC: 839 (GC: 27.0%)

(G) Liverwort genetic system genes

UUU-Phe	234	UCU-Ser	152	UAU-Tyr	195	UGU-Cys	60
UUC-Phe	13	UCC-Ser	20	UAC-Tyr	20	UGC-Cys	15
UUA-Leu	412	UCA-Ser	92	UAA-***	25	UGA-***	0
UUG-Leu	34	UCG-Ser	10	UAG-***	1	UGG-Trp	59
CUU-Leu	125	CCU-Pro	111	CAU-His	106	CGU-Arg	112
CUC-Leu	5	CCC-Pro	13	CAC-His	26	CGC-Arg	15
CUA-Leu	30	CCA-Pro	105	CAA-Gln	276	CGA-Arg	116
CUG-Leu	6	CCG-Pro	12	CAG-Gln	12	CGG-Arg	11
AUU-Ile	385	ACU-Thr	135	AAU-Asn	358	AGU-Ser	95
AUC-Ile	21	ACC-Thr	18	AAC-Asn	38	AGC-Ser	7
AUA-Ile	229	ACA-Thr	150	AAA-Lys	623	AGA-Arg	138
AUG-Met	116	ACG-Thr	17	AAG-Lys	15	AGG-Arg	7
GUU-Val	152	GCU-Ala	140	GAU-Asp	172	GGU-Gly	128
GUC-Val	16	GCC-Ala	27	GAC-Asp	18	GGC-Gly	24
GUA-Val	108	GCA-Ala	124	GAA-Glu	321	GGA-Gly	203
GUG-Val	15	GCG-Ala	13	GAG-Glu	22	GGG-Gly	20

AU: 5617
GC: 666 (GC: 10.6%)

(H) Liverwort photosynthesis genes

UUU-Phe	374	UCU-Ser	167	UAU-Tyr	165	UGU-Cys	33
UUC-Phe	47	UCC-Ser	17	UAC-Tyr	40	UGC-Cys	8
UUA-Leu	461	UCA-Ser	69	UAA-***	27	UGA-***	0
UUG-Leu	42	UCG-Ser	14	UAG-***	1	UGG-Trp	155
CUU-Leu	156	CCU-Pro	171	CAU-His	147	CGU-Arg	124
CUC-Leu	3	CCC-Pro	11	CAC-His	23	CGC-Arg	14
CUA-Leu	32	CCA-Pro	111	CAA-Gln	218	CGA-Arg	39
CUG-Leu	8	CCG-Pro	14	CAG-Gln	16	CGG-Arg	3
AUU-Ile	339	ACU-Thr	223	AAU-Asn	208	AGU-Ser	99
AUC-Ile	46	ACC-Thr	24	AAC-Asn	51	AGC-Ser	21
AUA-Ile	88	ACA-Thr	121	AAA-Lys	209	AGA-Arg	80
AUG-Met	161	ACG-Thr	11	AAG-Lys	22	AGG-Arg	6
GUU-Val	236	GCU-Ala	364	GAU-Asp	216	GGU-Gly	309
GUC-Val	14	GCC-Ala	19	GAC-Asp	30	GGC-Gly	28
GUA-Val	182	GCA-Ala	182	GAA-Glu	296	GGA-Gly	223
GUG-Val	19	GCG-Ala	20	GAG-Glu	29	GGG-Gly	31

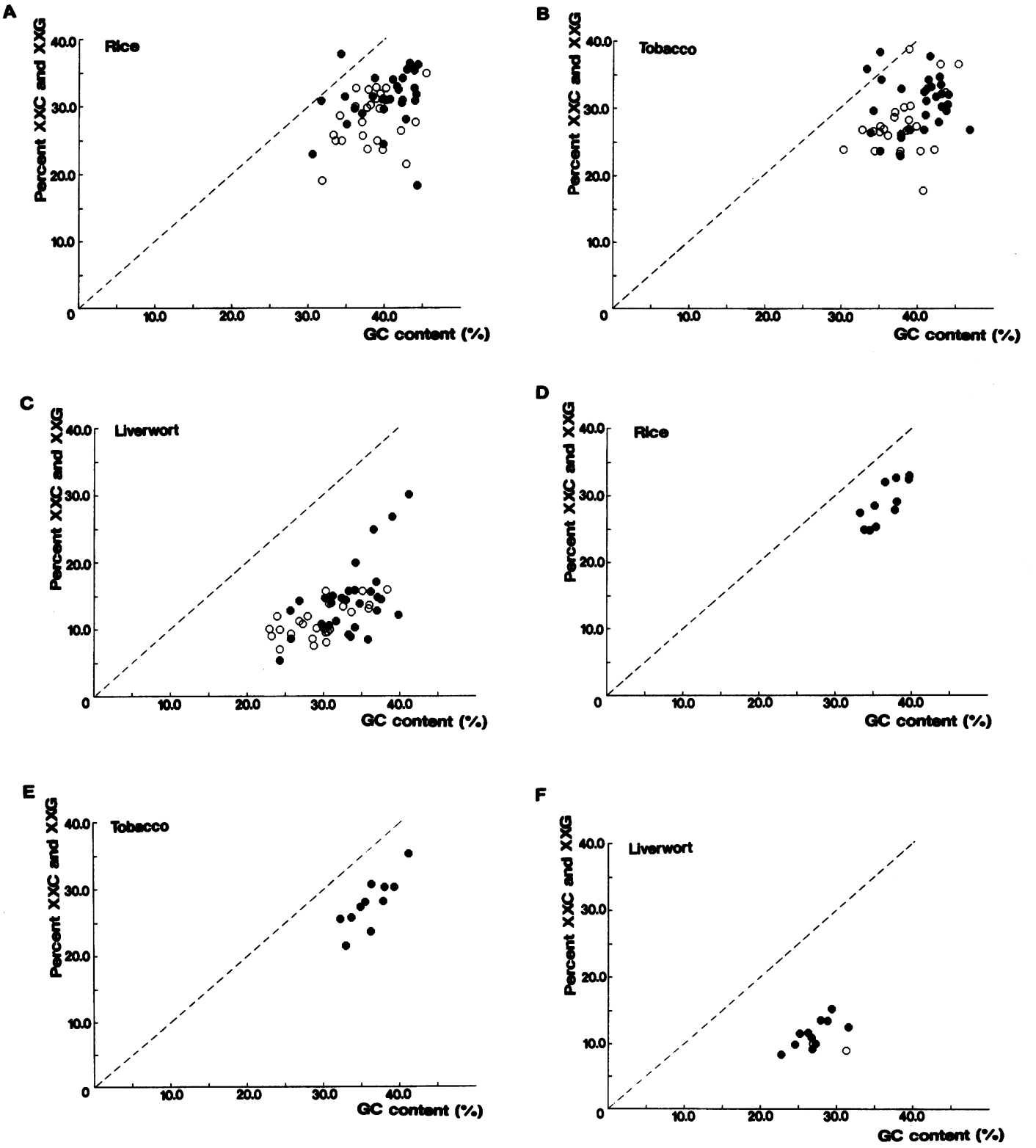
AU: 5668
GC: 949 (GC:14.2%)

(I) Liverwort NADH dehydrogenase genes

UUU-Phe	267	UCU-Ser	81	UAU-Tyr	130	UGU-Cys	38
UUC-Phe	13	UCC-Ser	12	UAC-Tyr	9	UGC-Cys	6
UUA-Leu	297	UCA-Ser	59	UAA-***	7	UGA-***	1
UUG-Leu	26	UCG-Ser	5	UAG-***	1	UGG-Trp	58
CUU-Leu	74	CCU-Pro	67	CAU-His	35	CGU-Arg	22
CUC-Leu	5	CCC-Pro	4	CAC-His	2	CGC-Arg	3
CUA-Leu	23	CCA-Pro	39	CAA-Gln	72	CGA-Arg	26
CUG-Leu	1	CCG-Pro	4	CAG-Gln	2	CGG-Arg	3
AUU-Ile	237	ACU-Thr	67	AAU-Asn	121	AGU-Ser	67
AUC-Ile	12	ACC-Thr	1	AAC-Asn	9	AGC-Ser	5
AUA-Ile	118	ACA-Thr	63	AAA-Lys	126	AGA-Arg	29
AUG-Met	92	ACG-Thr	4	AAG-Lys	4	AGG-Arg	3
GUU-Val	79	GCU-Ala	104	GAU-Asp	76	GGU-Gly	73
GUC-Val	7	GCC-Ala	7	GAC-Asp	6	GGC-Gly	9
GUA-Val	51	GCA-Ala	45	GAA-Glu	113	GGA-Gly	99
GUG-Val	1	GCG-Ala	7	GAG-Glu	9	GGG-Gly	19

AU: 2706
GC: 347 (GC: 11.4%)

The total number of codons which have AU or GC in the third position are listed below each table. The genes used to determine the data in Table 3 are listed in Table 2.


A
Rice
Percent XXC and XXG
GC content (%)
B
Tobacco
Percent XXC and XXG
GC content (%)
C
Liverwort
Percent XXC and XXG
GC content (%)
D
Rice
Percent XXC and XXG
GC content (%)
E
Tobacco
Percent XXC and XXG
GC content (%)
F
Liverwort
Percent XXC and XXG
GC content (%)
0
10.0
20.0
30.0
40.0

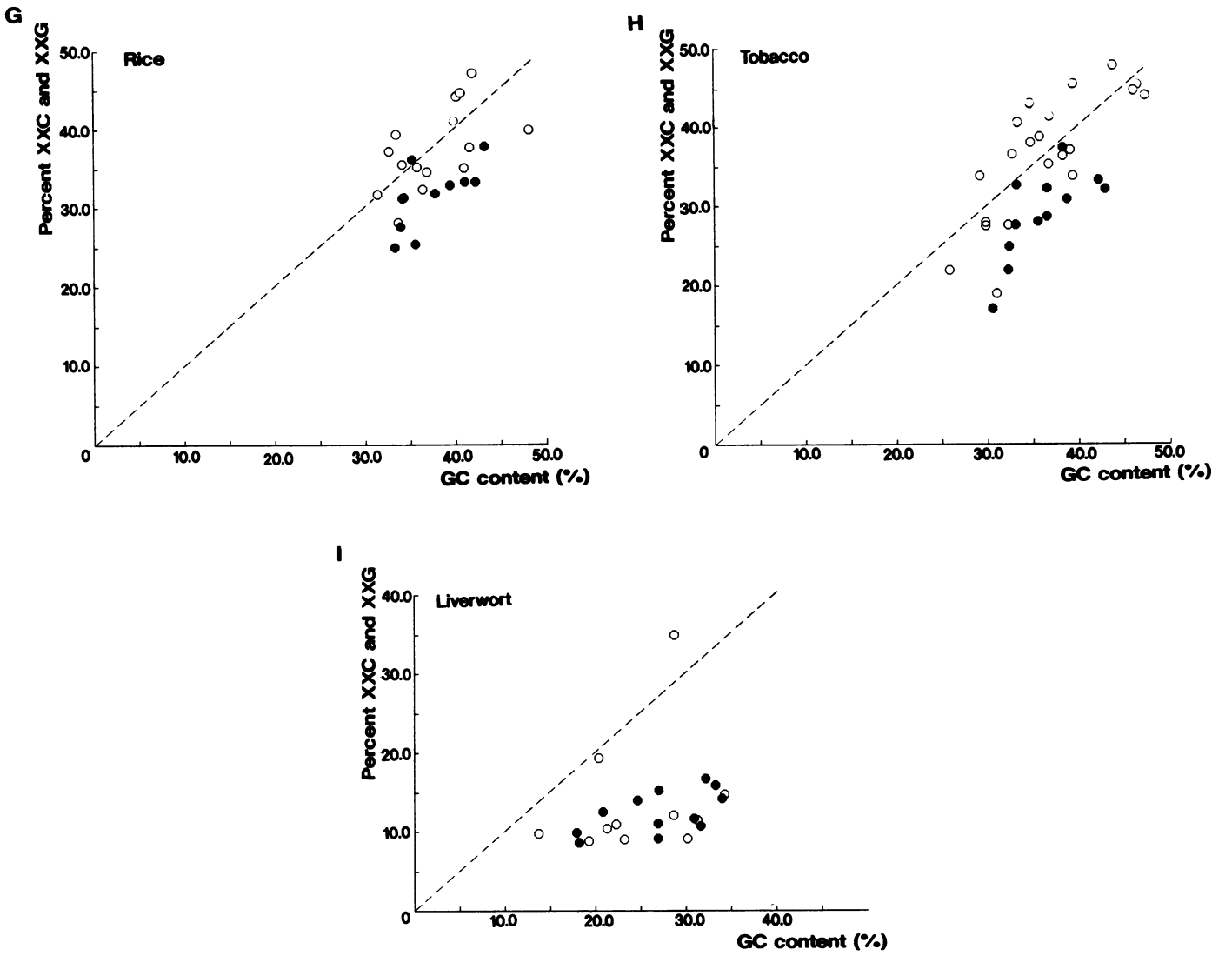


Figure 1. Plots of GC content of individual genes against the GC content of the third codon position in the same genes. A–C represent genes for the genetic system (open circles) and for photosynthesis (filled circles). D–F represent *ndh* sequences (filled circles) and unique ORFs in liverwort (*mbpX* and *frxC*; open circleıs). G–I represent conserved ORFs (filled circles) and unique ORFs in each genome (open circles). Vertical axes indicate GC content (%) in the third codon position and horizontal axes indicate GC content (%) in individual genes.

Table 4. GC contents (%) of whole genes (including ORFs) and of the third codon positions.

Genes	Rice		Tobacco		Liverwort		Remarks
	whole	3rd	whole	3rd	whole	3rd	
Genetic system Genes							
30S ribosomal proteins							
rps2	37.8	32.5	39.0	30.4	29.1	10.2	
rps3	33.3	25.8	36.1	26.0	29.8	10.6	
rps4	37.1	25.7	38.5	26.7	27.3	10.8	
rps7	39.7	23.6	40.4	23.7	32.5	13.5	
rps8	37.0	27.7	35.1	27.4	26.8	11.3	
rps11	42.8	21.5	45.3	36.6	38.4	16.0	
*rps12**	42.1	26.4	42.2	23.9	36.0	13.7	
rps14	40.1	32.7	42.9	36.6	30.7	13.9	
rps15	34.1	28.6	30.3	23.9	22.9	10.1	
*rps16**	31.8	19.1	37.6	23.3	–	–	
rps18	33.5	25.0	35.0	26.5	30.3	15.8	
rps19	39.4	31.9	34.4	23.7	30.5	9.7	
50S ribosomal proteins							
*rpl2**	45.4	35.0	43.6	32.4	33.6	12.6	
rpl14	39.0	25.0	39.8	27.4	30.4	8.1	

Table 4. *(continued)*

Genes	Rice		Tobacco		Liverwort		Remarks
	whole	3rd	whole	3rd	whole	3rd	
*rpl16**	44.0	27.7	40.7	17.8	35.9	13.2	
rpl20	36.1	30.0	37.0	29.5	28.5	8.6	
rpl21	–	–	–	–	27.9	12.0	
rpl22	36.2	32.7	35.5	26.9	30.8	10.0	
rpl23	37.6	29.8	36.9	28.7	23.9	12.0	
rpl32	34.4	25.0	32.7	26.8	24.3	7.1	
rpl33	38.8	32.8	38.8	38.8	23.2	9.1	
rpl36	37.7	23.7	37.7	23.7	35.1	15.8	
RNA polymerase subunits							
rpoA	38.1	30.2	34.2	26.6	25.7	9.4	
rpoB	39.7	31.2	38.8	26.9	30.1	10.1	
*rpoC1**	39.3	29.7	38.8	28.3	30.3	9.6	
rpoC2	38.8	31.2	38.2	30.2	24.3	10.0	
initiation factor							
infA	41.4	43.5	–	–	28.7	7.6	
NADH dehydrogenase							
*ndhA**	34.5	24.8	36.3	23.6	27.3	10.0	
*ndhB**	38.0	32.7	38.1	30.2	25.2	11.6	
ndhC	39.7	33.1	36.4	30.6	24.5	9.9	
ndhD	36.5	32.1	34.9	27.3	27.9	13.6	
ndhE	33.3	27.5	32.3	25.5	26.7	10.9	
ndhF	33.8	25.0	33.7	25.7	26.3	11.7	
ndhG	35.4	25.4	33.0	21.5	22.7	8.3	
psbG	38.1	29.2	37.9	28.1	28.8	13.5	
ORF393/393/392	37.8	27.9	39.4	30.2	31.5	12.5	
ORF178/167/*frxB*	35.2	28.5	35.5	28.0	26.8	9.2	
ORF159/158/169	39.6	32.5	41.3	35.2	29.4	15.3	
Photosynthesis Genes							
Ribulose 1,5-diphosphate carboxylase/oxygenase							
rbcL	44.1	31.8	43.7	29.7	37.5	14.5	
Photosystem I							
psaA	43.8	35.4	43.1	32.2	36.2	15.7	
psaB	41.5	32.9	41.1	31.2	34.7	13.9	
psaC	42.7	28.1	42.7	28.0	35.8	8.5	
psaI	34.2	37.8	34.2	29.7	24.3	5.4	
psaJ	37.0	28.9	41.5	37.8	31.0	14.0	
Photosystem II							
psbA	42.3	34.2	42.8	34.7	41.2	30.2	
psbB	43.9	30.8	43.9	30.6	37.0	12.8	
psbC	43.8	32.7	44.0	32.1	37.1	14.8	
psbD	44.3	36.2	43.0	33.6	36.9	17.2	
psbE	40.1	31.0	41.7	33.3	36.5	25.0	
psbF	41.7	32.5	40.8	32.5	34.2	20.0	
psbH	39.6	31.1	37.8	25.7	32.4	14.7	
psbI	36.0	29.7	33.8	26.4	29.7	10.8	
psb(J)	39.8	24.4	39.0	26.8	39.0	26.8	
psbK	35.0	27.4	35.0	38.4	26.8	14.3	
psbL	31.6	30.8	33.3	35.9	25.6	12.8	
psbM	30.5	22.9	35.2	34.3	25.7	8.6	
psbN	43.2	36.4	42.4	31.8	34.1	15.9	
Components of cytochrome b/f complex							
petA	41.0	34.0	41.4	34.3	30.7	10.6	
*petB**	39.8	29.6	41.2	32.9	32.9	14.4	
*petD**	40.6	31.1	37.7	23.0	33.3	9.3	
petG	38.6	34.2	35.1	23.7	33.3	15.8	
H^+–ATPase subunits							
atpA	42.3	30.9	40.8	26.8	33.5	8.9	
atpB	42.2	30.5	43.1	30.3	34.1	10.3	
atpE	42.8	35.5	41.0	29.1	30.2	14.7	
*atpF**	34.7	31.5	37.8	33.0	31.2	15.1	
atpH	44.3	18.3	46.8	26.8	39.8	12.2	
atpI	38.4	31.5	37.8	26.2	31.6	11.3	

			Conserved ORFs				
ORF29/29/29	42.2	33.3	42.2	33.3	32.2	16.7	
ORF31/31/31	33.3	25.0	32.3	21.9	20.8	12.5	
ORF35/34/35	35.2	36.1	30.5	17.1	26.9	11.1	
ORF62/62/62	35.5	25.4	36.5	28.6	33.3	15.9	
IRF170/168/167*	39.4	32.8	39.3	31.4	31.6	10.7	
ORF216/196/203	43.2	37.8	42.9	32.1	34.0	14.2	
ORF185/184/184	41.0	33.3	38.7	30.8	26.9	9.2	
ORF230/229/434	34.1	31.2	33.2	32.6	24.6	14.0	
ORF2280/2136	–	–	38.3	37.4	17.9	9.9	
ORF321/313/320	33.9	27.6	32.4	24.8	27.0	15.3	
ORF106/512/316	37.7	31.8	36.5	32.2	30.9	11.7	
			Unique ORFs				
ORF100	31.4	31.7	–	–	–	–	
ORF91	34.1	35.5	–	–	–	–	
ORF70	35.7	35.2	–	–	–	–	
ORF42	32.6	37.2	–	–	–	–	
ORF133	40.3	41.8	–	–	–	–	
ORF85	33.7	40.7	–	–	–	–	
ORF82	41.0	43.4	–	–	–	–	
ORF137	48.1	39.9	–	–	–	–	1)
ORF28	36.8	34.5	–	–	–	–	2)
ORF64	36.4	32.3	–	–	–	–	2)
ORF249	43.5	47.2	–	–	–	–	
ORF72	41.2	43.1	–	–	–	–	
ORF85	41.5	37.7	–	–	–	–	
ORF23	44.4	62.5	–	–	–	–	
ORF63	33.6	28.1	–	–	–	–	
ORF56	40.9	35.1	–	–	–	–	3)
ORF64	–	–	29.2	33.8	–	–	
ORF51	–	–	32.7	36.5	–	–	
ORF41	–	–	33.3	40.5	–	–	
ORF154	–	–	35.7	38.7	–	–	
ORF105	–	–	36.5	40.6	–	–	
ORF70A	–	–	34.7	38.0	–	–	
ORF99A	–	–	34.7	42.0	–	–	
ORF99B	–	–	31.0	19.0	–	–	
ORF103	–	–	29.8	27.9	–	–	
ORF87	–	–	46.2	44.3	–	–	
ORF92	–	–	47.0	43.0	–	–	
ORF115	–	–	46.0	45.7	–	–	
ORF79	–	–	38.3	36.3	–	–	
ORF70B	–	–	39.4	33.8	–	–	
ORF131	–	–	39.1	37.1	–	–	
ORF38	–	–	39.3	46.2	–	–	
ORF75	–	–	45.6	48.7	–	–	
ORF350	–	–	36.7	35.3	–	–	
ORF228	–	–	29.8	27.5	–	–	
ORF273(*ssb*)	–	–	25.8	21.9	–	–	
ORF1244	–	–	32.3	27.6	–	–	
ORF135*	–	–	–	–	22.3	11.0	
ORF33	–	–	–	–	34.3	14.7	
ORF30	–	–	–	–	20.4	19.4	
ORF32	–	–	–	–	23.2	9.1	
ORF513	–	–	–	–	31.3	11.5	
ORF50	–	–	–	–	13.7	9.8	
ORF42a	–	–	–	–	28.7	34.9	
ORF288	–	–	–	–	28.6	12.1	
ORF289(*frxC*)	–	–	–	–	31.3	9.0	
ORF370(*mbpX*)	–	–	–	–	26.9	10.0	
ORF464	–	–	–	–	19.3	8.8	
ORF1068	–	–	–	–	21.3	10.4	
ORF465	–	–	–	–	30.2	9.2	
			ORFs within the trn's introns				
ORF542/509/370	34.3	31.3	33.1	27.6	18.2	8.6	4)
ORF133	50.3	50.0	–	–	–	–	5)
ORF109	50.0	46.4	–	–	–	–	6)

1) This ORF lies in the strand opposite the *rpl2* gene.
2) This ORF has local homology to tobacco ORF2280.
3) This ORF corresponds to ORF393 in the other inverted repeat.
4) Within *trnK*.
5) Within *trnI*.
6) Within *trnA*.
* Genes containing introns.

tobacco sequence for ORF2280 has recently been revised, and defined as a single ORF which was previously thought contain two ORFs, ORF581 and ORF1708 (T. Wakasugi, personal communication). Though ORF2280 in tobacco corresponds in location to ORF2136 in liverwort, these ORFs have only local homology in common. Rice has no ORF which corresponds in size to ORF2280 and ORF2136 in tobacco and liverwort, however, it contains two short ORFs, ORF28 and ORF64, which bear homology to parts of tobacco ORF2280. Thus, these rice ORFs could be derived from an ORF similar to tobacco ORF2280 as a result of large deletions in the sequence.

Based on homology of translation products, the chloroplast genes from rice, tobacco and liverwort have been classified into two groups. One group includes the genes encoding ribosomal protein, RNA polymerase and NADH dehydrogenase. This group of genes include homologies of 70−80%, between rice, tobacco and liverwort. The second group of genes all have homologies of more than 80% and mainly encode components of the photosynthetic apparatus. The conserved ORFs are distributed into these two groups based on homology. Among them, ORF29/29/29, ORF31/31/31, ORF35/34/35 and IRF170/168/167 (IRF: intron-containing reading frame) have homologies of 80% or more and are grouped with photosynthesis genes. ORF542/509/370 lies within an intron in the *trnK*-UUU gene and has less homology than all the other genes and ORFs, suggesting that ORF542/509/370 may not be functional.

The average GC content of photosynthesis genes (excluding introns) in rice, tobacco and liverwort are calculated to be 41.9%, 41.6% and 34.8%, respectively. These values are substantially higher than those found in the ribosomal protein genes, 38.8%, 38.5% and 28.7% (in rice, tobacco and liverwort, respectively). The GC content of the ribosomal protein genes in comparable to GC content of the entire genomes which are 39.0%, 37.9% and 28.8% in rice, tobacco and liverwort, respectively.

The codon usage pattern (Table 3) suggests that the codons containing A or U on the third position are given preference in the chloroplast genome. Figure 1 shows the relationship between the GC contents of an individual genes (including ORFs) and

Table 5. Introns in rice, tobacco and liverwort chloroplasts.

Gene	Rice		Tobacco		Liverwort		Remarks
	bp	%GC	bp	%GC	bp	%GC	
trnA-UGC	812	50.9 (56.2)	709	51.1 (56.2)	768	41.2 (58.9)	
trnI-GAU	947	49.8 (59.7)	707	50.8 (59.7)	886	37.5 (37.5)	
trnG-UCC	678	32.5 (54.2)	691	32.1 (53.5)	593	19.2 (51.4)	
trnK-UUU	2504	33.7 (54.2)	2526	33.0 (54.2)	2111	18.9 (56.9)	
trnL-UAA	540	35.0 (50.6)	503	33.8 (48.2)	315	22.9 (52.9)	
trnV-UAC	597	38.0 (50.0)	571	36.4 (50.7)	530	24.5 (51.4)	
atpF	828	32.7 (34.7)	695	32.2 (37.8)	587	24.9 (31.2)	
ndhA	987	32.0 (34.5)	1107	32.1 (36.3)	712	22.8 (27.3)	
ndhB	712	39.8 (38.0)	517	39.2 (38.1)	536	17.2 (25.2)	
petB	811	33.4 (39.8)	753	32.9 (41.2)	495	18.6 (32.9)	
petD	744	37.0 (40.6)	741	36.0 (37.7)	493	19.1 (33.3)	
3'-rps12	540	40.4 (42.1)	536	39.4 (42.2)	500	21.2 (36.0)	
rps16	878	34.6 (31.8)	860	35.9 (37.6)			
rpl2	663	40.4 (45.4)	666	39.0 (43.6)	544	23.5 (33.6)	
rpl16	1059	29.9 (44.0)	1020	32.2 (40.7)	536	21.1 (35.9)	
rpoC1	–	– (39.3)	738	36.9 (38.8)	596	19.8 (30.3)	
IRF170/168/167(1)	745	40.5	738	38.5	608	25.2	1)
(2)	729	37.3 (39.4)	783	32.2 (39.3)	–	– (31.6)	2)
ORF216/IRF196/203(1)	–	–	806	32.8	381	17.8	1)
(2)	–	– (43.2)	643	32.8 (42.9)	518	14.4 (34.0)	2)
IRF135*						18.8 (22.3)	

The GC content of exons are shown in parentheses. Hyphens indicate that no intron in present in the corresponding gene or ORF.
1) First intron in the IRF.
2) Second intron in the IRF.
* only in liverwort.

Table 6. Spacer regions in rice, tobacco and liverwort chloroplasts.

Spacer regions	Number of bp			Remarks
	Rice	Tobacco	Liverwort	
rps19(rpl2)–*trnH*-GUG	52	5	239	1)
trnH-GUG–*psbA*	81	453	165	
psbA–*trnK*-UUU	229	214	145	
trnK-UUU–*trnQ*-UUG	2683	3008	2164	2)
trnQ-UUG–*psbK*	345	347	198	
psbK–*psbI*	389	329	330	
psbI–*trnS*-GCU	110	123	17	
psbD–*psbC*	−53	−53	−53	
psbC–*trnS*-UGA	176	239	121	
trnS-UGA–ORF62/62/62	346	362	152	
ORF62/62/62–*trnG*-GCC	205	275	199	
trnG-GCC–*trnfM*-CAU	437	227	50	3)
trnT-GGU–*trnE*-UUC	518	848	1579	4)
trnE-UUC–*trnY*-GUA	61	59	71	
trnY-GUA–*trnD*-GUC	363	108	77	
psbM–ORF29/29/29	766	1132	1063	
ORF29/29/29–*trnC*-GCA	413	670	393	
trnC-GCA–*rpoB*	1084	1281	138	
rpoB–*rpoC1*	37	5	30	
rpoC1–*rpoC2*	199	153	73	
rpoC2–*rps2*	271	226	83	
rps2–*atpI*	250	226	127	
atpI–*atpH*	794	1158	377	
atpH–*atpF*	456	401	208	
atpF–*atpA*	98	54	45	
atpA–*trnR*-UCU	132	123	72	
rps14–*psaB*	147	122	88	
psbB–*psaA*	25	25	26	
psbA–ORF170/168/167	600	752	280	
ORF170/168/167–*trnS*-GGA	600	852	245	
trnS-GGA–*rps4*	285	330	492	
rps4–*trnT*-UGU	299	371	227	
trnT-UGU–*trnL*-UAA	770	710	188	
trnL-UAA–*trnF*-GAA	242	356	76	
trnF-GAA–ORF159/158/169	490	676	162	
ORF159/158/169–*psbG*	97	105	50	
psbG–*ndhC(3)*	−10	−121	−10	
ndhC(3)–*trnV*-UAC	704	1087	173	
trnV-UAC–*trnM*-CAU	181	190	148	
trnM-CAU–*atpE*	112	221	80	
atpE–*atpB*	−4	−4	5	
atpB–*rbcL*	784	817	508	
rbcL–*psaI*	1693	3054	1410	5)
psaI–ORF185/184/184	369	444	221	
ORF185/184/184–ORF230/229/434	417	222	71	
ORF230/229/434–*petA*	231	230	185	
petA–*psb(J)*	1001	1065	190	
psb(J)–*psbL*	126	124	119	
psbL–*psbF*	22	22	21	
psbF–*psbE*	10	9	9	
psbE–ORF31/31/31	1196	1163	597	6)
ORF31/31/31–*petG*	172	181	122	
petG–*trnW*-CCA	116	131	69	
trnW-CCA–*trnP*-UGG	126	164	88	
trnP-UGG–*psaJ*	318	438	238	
psaJ–*rpl33*	441	432	117	
rpl33–*rps18*	242	186	27	
rps18–*rpl20*	222	199	81	
rpl20–*5'-rps12*	675	811	786	
5'-rps12–ORF216/196/203	134	149	72	
ORF216/196/203–*psbB*	510	445	385	
psbB–*psbN*	320	378	310	
psbN–*psbH*	103	111	97	
psbH–*petB*	129	129	107	
petB–*petD*	192	189	148	
petD–*rpoA*	219	187	111	
rpoA–*rps11*	63	65	32	
rps11–*rpl36*	174	101	50	
rpl36–*infA*	111	12	36	
infA–*rps8*	136	134	86	
rps8–*rpl14*	139	168	81	

Table 6. *(continued)*

Spacer regions	Number of bp			Remarks
	Rice	Tobacco	Liverwort	
rpl14–rpl16	109	124	97	
rpl16–rps3	145	146	57	
rps3–rpl22	55	−16	48	
rpl22–rps19	68	53	17	
rps19–rpl2	261	60	36	7)
(*rps19–trnH*-GUG	131	–	–)	
(*trnH*-GUG–*rpl2*	55	60	–)	
rpl2–rpl23	18	18	36	
rpl23–trnI-CAU	174	165	158	
trnL-CAA–*ndhB(2)*	603	539	123	
ndhB(2)–*rps7*	304	647	154	
rps7–3'-rps12	58	53	49	
3'-rps12–trnV-GAC	1722	1607	801	8)
trnV-GAC–16SrRNA	231	227	223	
16SrRNA–*trnI*-GAU	310	300	273	
trnI-GAU–*trnA*-UGC	64	64	76	
trnA-UGC–23SrRNA	145	153	159	
23SrRNA–4.5SrRNA	95	101	107	
4.5SrRNA–5SrRNA	227	256	226	
5SrRNA–*trnR*-ACG	256	257	204	
trnR-ACG–*trnN*-GUU	251	581	705	
trnN-GUU–*ndhF(5)*	2145	1366	768	9)
ndhF(5)–rpl32	714	768	706	10)
rpl32–trnL-UAG	530	932	1178	11)
trnL-UAG–ORF321/313/320	82	103	128	
ORF321/313/320–*ndhD(4)*	196	237	220	
ndhD(4)–psaC	119	90	124	
psaC–ndhE(4L)	446	260	222	
ndhE(4L)–ndhG(6)	209	223	53	
ndhG(6)–ORF178/167/*frxB*	242	396	90	
ORF178/167/*frxB–ndhA(1)*	94	84	51	
ndhA(1)–ORF393/393/392	1	1	1	
ORF393/393/392–*rps15*	138	111	52	
rps15–trnN-GUU	1530	6431	8088	12)
Spacer that contains an inversion breakpoint				
trnS-GCU–*psbD*	983	–	–	13)
trnfM–trnG-UCC	97	–	–	
trnG-UCC–*trnT*-GGU	1307	–	–	14)
trnR-UCU–*rps14*	371	–	–	15)
trnS-GCU–*trnG*-UCC	–	779	845	
trnG-UCC–*trnR*-UCU	–	169	64	
psbM–trnD-GUC	–	380	1074	
trnT-GGU–*psbD*	–	1218	416	
trnL-CAA–*psbM*	–	–	242	
trnfM–rps14	–	149	103	
trnI-CAU–*trnV*-GAC	–	–	756	
trnI-CAU–*trnL*-CAA	1498	7656	–	16)
trnH-GUG–ORF2136	–	–	239	
ORF2136–*trnD*-GUC	–	–	91	
rps15–ndhF	614	–	–	

1) *rps19* in rice and *rpl2* in tobacco and liverwort.
2) *rps16* in rice and tobacco, ORF513 in liverwort are present within the spacer.
3) Ψ*trnG* in rice.
4) Ψ*trnT*, Ψ*trnE* in rice, ORF370(*mbpX*) in liverwort.
5) Ψ*rpl23*, ORF133 and ORF106 in rice, ORF512 in tobacco, *trnR*–CCG and ORF316 in liverwort.
6) Ψ*rps12* in rice.
7) Junction between IR and LSC in rice and tobacco, *trnH* in rice.
8) Junction between IR and LSC in liverwort.
9) Junction between IR and SSC in rice and tobacco, *rps15* in rice and ORF350 in tobacco.
10) *rpl21* in liverwort.
11) ORF288 and *trnP*-GGG in liverwort.
12) Junction between IR and SSC in three genomes, ORF228, ORF273(*ssb*) and ORF1244 in tobacco, ORF464, ORF1068 and *frxC* in liverwort.
13) ORF100 in rice.
14) Ψ*rps12* in rice.
15) Ψ*trnfM/G* in rice.
16) ORF249 in rice, ORF581 and ORF1708 in tobacco.

the numbers of codons in the same gene with use G or C in the third position. Plots of this relationship for most genes are localized in the lower right quadrant of the graph (when both axes are extended to 100%), indicating a strong tendency to use A or U in the third codon position. This tendency is strongest in liverwort. The photosynthesis genes show high homology among the three chloroplast genomes (see Table 2), higher overall GC contents and higher GC preferences in the third codon position than do other genes (Table 4). Thus, photosynthesis genes can also be considered to constitute a distinct and single group based on GC content (they are plotted in higher positions in Figure 1A−C).

It has been reported that nuclear genes in higher plants tend to use G or C in the third position of codons. Especially in most monocot nuclear genes, the GC content of the third codon position is more than 80% (17). The fact that the chloroplast genome tends to use A or U in the third codon position suggests the chloroplast genome is derived from different origins than the nuclear genomes or has been affected by different evolutional forces.

The *ndh* sequences are placed in the lower sequence homology group, separate from the group of photosynthesis genes. The GC content of these genes also supports this classification since the relationship between GC content and codon usage among these gene is more similar to that of ribosomal protein genes than that of photosynthesis genes (Fig. 1D−F). These observations suggest that *ndh* products have different functions or locations than photosynthesis proteins.

Plots for the GC content and codon usage for the conserved ORFs also show that these genes fall into lower homology group along with most of the chloroplast genes (Fig. 1F−I). However, dots representing some unique ORFs are dispersed in these plots. ORFs with exceptional values in GC content and GC usage in the third codon position are thought to be unfunctional. Conversely, conserved ORFs showing similar features to photosynthesis genes may encode photosynthesis proteins. These include ORF29/29/29, IRF170/168/167 and ORF216/IRF196/203. The the other conserved ORFs similar to ribosomal protein genes or *ndh* sequences are ORF31/31/31 and ORF321/313/320.

Some chloroplast genes contain introns. Introns from rice and tobacco are, in most cases, larger than those of liverwort (Table 5). The rice RNA polymerase gene, *rpoC1*, and ORF216 lack introns and liverwort IRF167 lacks the second intron. The GC contents in introns is lower than that of exons.

Spacer regions

The size of the chloroplast genomes in rice, tobacco and liverwort are different. These differences in genome size are largely accounted for by differences in the inverted repeat regions (IRs) and spacer regions. As shown in Table 6, most of the spacer regions in liverwort are shorter than those in rice and tobacco. The liverwort spacers are highly rich in A and T, and although promoters and terminators are included in the spacer regions between genes, little homology exists between the spacers in rice and tobacco and the spacers in liverwort (data not shown). It is therefore assumed that the structure of chloroplast promoters and terminators in liverwort are quite different from those found in the chloroplasts of higher plants.

REFERENCES

1. Palmer,J.D. (1985) Annu. Rev. Genet., **19**, 325−354.
2. Sugiura,M. (1989) Annu. Rev. Cell Biol. **5**, 51−70.
3. Shinozaki,K., Ohme,M., Tanaka,M., Wakasugi,T., Hayashida,N., Matsubayashi,T., Zaita,N., Chunwongse,J., Obokata,J., Shinozaki,K.Y., Ohto,C., Torazawa,K., Meng,B.Y., Sugita,M., Deno,H., Kamogashira,T., Yamada,K., Kusuda,J., Takaiwa,F., Kato,A., Tohdoh,N., Shimada,H. and Sugiura,M. (1986) EMBO J., **5**, 2043−2049.
4. Ohyama,K., Fukuzawa,H., Kohchi,T., Sano,T., Sano,S., Shirai,H., Umezomo,K., Shiki,Y., Takeuchi,M., Chang,Z., Aota,S., Inokuchi,H. and Ozeki,H. (1988) J. Mol. Biol. **203**, 281−298.
5. Hiratsuka,J., Shimada,H., Whittier,W.F., Ishibashi,T., Sakamoto,M., Mori,M., Kondo,C., Honji,Y., Sun,C.R., Meng,B.Y., Li,Y.Q., Kanno,A., Nishizawa,Y., Hirai,A., Shinozaki,K. and Sugiura,M. (1989) Mol. Gen. Genet., **217**, 185−194.
6. Edward,K. and Koessel,H (1981) Nucl. Acids Res., **9**, 2853−2869.
7. Shimada,H. and Sugiura,M. (1989) Curr. Genet. **16**, 293−301.
8. Howe,C.J. (1985) Curr. Genet., **10**, 139−145.
9. Howe,C.J., Barker,R.F., Bowman,C.M. and Dyer,T.A. (1988) Curr. Genet., **13**, 343−349.
10. Kato,A., Takaiwa,F., Shinozaki,K. and Sugiura,M. (1985) Curr. Genet., **9**, 405−409.
11. Sijben-Mueller,G., Hallick,R.B., Alt,J., Westhoff,P. and Herrman,R.G. (1986) Nucl.Acids Res. **14**, 1029−1044.
12. Fearnley,I.M., Runswick,M.J. and Walker,J.E. (1989) EMBO J., **8**, 665−672.
13. Nixon,P.J., Gounaris,K., Coomber,S.A., Hunter,C.N., Dyer,T.A. and Barber,J. (1989) J.Biol.Chem., **264**, 14129−14135.
14. Wu,M., Nie,Z.Q. and Yang,J. (1989) Plant Cell, **1**, 551−557.
15. Mayers,S.R., Cook,K.M. and Barber,J. (1990) FEBS Lett. **262**, 49−54.
16. Yokoi,F., Vassileva,A., Hayashida,N., Torazawa,K., Wakasugi, T. and Sugiura,M. (1990) FEBS Lett. in press.
17. Campbell,W.H. and Gorwi,G. (1990) Plant Physiol. **92**, 1−11.
18. Hallick,R.B. and Bottomley,W. (1983) Plant Mol. Biol. Repr. **1**, 38−43.
19. Hallick,R.B. (1989) Plant Mol. Biol. Repr. **7**, 266−275.
20. Torazawa,K., Hayashida,N., Obokata,J., Shinozaki,K. and Sugiura,M. (1986) Nucl. Acids Res., **14**, 3143.