Junctions of the large single copy region and the inverted repeats in *Spinacia oleracea* and *Nicotiana debneyi* chloroplast DNA: sequence of the genes for tRNA^{His} and the ribosomal proteins S19 and L2

Gerard Zurawski, Warwick Bottomley¹ and Paul R.Whitfeld¹

DNAX Research Institute, 1450 Page Mill Road, Palo Alto, CA 94304, USA and ¹Division of Plant Industry, CSIRO, GPO Box 1600, Canberra, ACT 2601, Australia

Received 31 May 1984; Revised and Accepted 4 August 1984

ABSTRACT

This work describes the organization, at the nucleotide sequence level, of genes flanking the junctions of the large single copy regions and the inverted repeats of <u>Spinacia oleracea</u> (spinach) and <u>Nicotiana debneyi</u> chloroplast DNAs. In both genomes, <u>trnH1</u>, the gene for tRNA-His(GUG) is located at the extremity of the large single copy region 3' to <u>psbA</u>, the gene for the 35 kd Photosystem 2 protein. Both <u>psbA</u> and <u>trnH1</u> are transcribed towards the inverted repeat. In spinach, the first 48 codons of <u>rps19</u>, the gene for the last 44 codons lie in the large single copy region at the end opposite to that carrying <u>trnH1</u>. The gene for a protein homologous to the <u>E</u>. <u>coli</u> ribosomal protein L2, <u>rpl2</u>, is in the inverted repeat immediately 5' to <u>rps19</u> and, like <u>rps19</u>, is transcribed towards the large single copy region. In <u>N</u>. <u>debneyi</u>, but not in spinach, <u>rpl2</u> is interrupted by a 666 bp insertion. The gene for tRNA-IIe(CAT), <u>trn11</u>, is located in the same direction as rpl2.

INTRODUCTION

The chloroplast DNA of most higher plants is arranged as a 130-150 kb circle with a large and a small single copy region separated by two repeat units which are arranged in inverted orientation (1). The 20-26 kb repeats encode genes for ribosomal RNA as well as some tRNA genes (1,2). In spinach (<u>Spinacia oleracea</u>) and <u>Nicotiana debneyi</u> (an Australian relative of <u>N. tabacum</u>) <u>psbA</u>^{*}, the gene for the 35 kd Photosystem 2 protein, is located in the large single copy region adjacent to the right-hand inverted repeat (3,4).

Preliminary experiments had indicated there might be an open reading frame traversing the inverted repeat - large single copy region junction. Because of the implications of this observation, we have now analysed the two junction regions in detail and here we report the identity, structure, and organization of the genes located in these regions.

MATERIALS AND METHODS

Plasmids pSocS1, pSocS15, pSocS24 (bearing spinach chloroplast DNA <u>Sal1</u> fragments 5, 6 and 9, respectively), pSocB5, pSocB11 (bearing spinach chloroplast DNA <u>BamH1</u> fragments 5 and 11, respectively) have been described previously (5). The source of cloned <u>N</u>. <u>debneyi</u> chloroplast DNA was pNdcB76, a plasmid bearing a 5 kb <u>BamH1</u> fragment (3). Procedures for plasmid purification, for the isolation and end-labelling of restricted DNA fragments and for sequence analysis are those described previously (3).

RESULTS

The Gene for $tRNA_{GUG}^{His}$ is at the Extremity of the Large Single Copy Region

Figure 1 is a representation of the circular spinach chloroplast DNA molecule detailing the various features discussed below. In particular we shall refer to the right-hand junction of the large single copy region as that adjacent to <u>psb</u>A. This junction maps in Sal6 (<u>Sal1</u> fragment 6) and Bam5 (<u>BamH1</u> fragment 5) while the left-hand junction maps in Sal9 and Bam11. The precise location of the junctions was defined by sequencing parts of Bam5 and Bam11 according to the strategy outlined in Figures 2a and 2b. Comparison of the sequences showed the two to be identical up to position 1563 in Bam11 (Figure 3) and position 144 in Bam5 (Figure 4a). Beyond these points the sequences bear no similarity. The sequences of the left and right junctions are 5'ACG/ATT and 5'ACG/GGC respectively.

In spinach chloroplast DNA the gene for $tRNA_{GUG}^{His}$ (<u>trn</u>H1) maps in the large single copy region close to its rightward extremity (7). Inspection of the nucleotide sequence distal to psbA revealed that <u>trn</u>H1 (shown as the



Figure 1. Location of the junctions of the large single copy region with the left and right inverted repeat in spinach chloroplast DNA. The extent of the inverted repeats is denoted by the shading between the two strands. Sall sites (6) are marked by arrowheads and the position of BamH1 fragments 5 and 11 is indicated. Protein coding regions identified in (3) and in this paper are shown as black bars on the appropriate DNA strand. Designation of genes is defined in the text. The polarity of the DNA strand represented by the outer circle is clockwise.



Figure 2. Partial restriction maps of the large single copy region inverted repeat junctions in spinach and N. debneyi chloroplast DNA. (a) Right-hand junction in spinach in Bam 5. The coordinates are nucleotides relative to the <u>Smal</u> site 5' to <u>psbA</u> (3). (b) Left-hand junction in spinach in Bam 11. (c) Right-hand junction in N. debneyi. The dotted lines through <u>rps19</u> and <u>rps19'</u> indicate the position of the junctions. Genes transcribed from within the inverted repeat towards the large single copy region are indicated by solid bars, those transcribed towards the inverted repeat by open bars. Designation of genes is defined in the text. Arrows indicate the origin and extent of individual sequencing runs. Arrows commencing in empty and filled circles indicate sequencing runs utilizing, respectively, 3' and 5' end-labelled DNA fragments. Arrows commencing at bars represent dideoxy sequencing runs.

complementary sequence in Figure 4a) lies 144 bp 3' to the <u>psb</u>A translation stop codon. Both <u>psb</u>A and <u>trn</u>H1 are transcribed towards the inverted repeat. All but the last base (position 144, Figure 4a) of <u>trnH1</u> lies in the large single copy region.

In <u>N</u>. <u>debneyi</u> chloroplast DNA, <u>trn</u>H1 (shown as the complementary sequence in Figure 4b) is located 360 nucleotides distal to <u>psb</u>A. Although in this case we have not directly identified the junction of the inverted repeat and the large single copy region by also sequencing across the left-hand junction, the location of the right-hand junction can be deduced as being at position 59 (Figure 4b) by comparison of the <u>N</u>. <u>debneyi</u> sequence with the spinach and <u>N</u>. <u>tabacum</u> (8,8a) sequences (see below). Thus <u>trn</u>H1 lies entirely within the large single copy region, its 3¹ terminal

GATCCCCG	GCTAA	trnI GCAT	CCAI	16601	GAAT	GGT1	AAAG	CGCC	CAAC	TCAT	AATI	10800	BAATI	CGT	66TT	CAAT	TCCT	ACTO	GATE	SCACE	SCCA4	TGGG	ACCI	CTCC	ATA	GTCI	ATTE	GAAT	120
TEECTCT	STATO	AATE	GAAT	гстси	ATCAT	CCAT	TACAT		GAATI	GGTE	STOGI	TATA	TCAT	TATÇA	TAAC	ATA	GAAG	CAGT	AAGAA	ACTAG	CATI	ICTT/	116	AAAC'	CGAA	CTC	TAGE	BAAG	240
AAAATAGA	ATTI	TGGA	1664	AATC	AAATA	ATGCA	AGTAI	TTA	CAGAC	CAAAA	16TA1	TCA	STTA	, TGG0	38AAA	AAĢ	CAAT	TAC	TCT	AATGI	CGA	TCA	GAT	CAAC	AGG/	CGG/	AATA	AAGC	360
ATTEGET	CGAAC		CTT	1661	TCAA	664	AATAG	CTA	GAAT	AGTO	CATCO	BACT	CC66	GAAA	16661	AGA	46GA1	6666	ACCT	ATTA	6664	CAT	CAA'	TGCA	TAC	GAC	TATE	BATCA	466
TTACACT	10441	ICGAI	STTA	TICT	ATTC		ICTI	GAA	16444		ACTT	-		TAC		TAGC	npi 2 met ATG	ala 606	ile ATA	his CAT	1eu TTA	ţăc	lys AAA	thr	ŧĉi	thr.		AGC	573
THEOCT	(G	thr		359	ard			- VAN			Á	Ă					tur			his	his					Ċ		
thr ars ACA CGC	asn AAT	ala GGA	ala GCC	val GTT	GAC	AGT	≰ln CAA	¥a1 GTG	lys AAA	Ser TCC	asn AAT	CCA	ars CGA	asn AAT	asn AAT	leu TTG	ile Atc	TCT	⊴1y 66A	⊈ln CAG	CGT	CGT	TĞT	¶l¥ GGT	lvs AAA	∎ly GGT	CGT	asn Aat	657
ala ars GCC AGA	sly GGA	ile ATC	ile ATT	thr ACC	ala GCA	ars AGG	his CAT	ars Aga	si) 666	⊈1¥ 66A	\$1¥ 66T	his Cat	1 y s AA6	ars Cot	leu CTA	tyr TAC	ars CGT	lys AAA	ile ATA	asp Gat	Phe TTT	ars AGA	ars CGG	asn AAT	⊈lu GAA	lys AAA	asp GAC	ile ATA	747
tur slu	ars lys	ile	val	thr	ile	≰lu	tyr	asp	970	asn	ars	asn	ala	twr	ile		leu	ile	e his	tyr	alv	asp	alv	≴lu	lys	ars	tyr	ile	
TAT GGT	AAA G	ATC	GTA	ACT	ĂTĂ	GAA	TÁC	GAC	CCT	AAT	ĊĠĂ	AAT	6CĂ	TAC	ĂŤŤ	TGT	ĊŤĊ	ĀTĀ	CAC	TAT	GGG	GAT	66Ç	GAG	AAG	ÂGA	TAT	ATT	837
leu his TTA CAT	600 000	ars Aga	⊈1¥ 666	ala GCT	ile ATA	ile ATT	≰ly GGA	asp Gat	thr ACC	ile ATT	¥al GTT	ser TCT	\$1 ¥ 66 T	thr ACA	≨lu GAA	val GTT	ĉċŝ	ile ATA	lys AAA	aet ATG	⊈ly GGA	asn AAT	ala GCC	leu CTA	669	100			
TGAACTA	TTGA	TTA	CGTA	ATTG	GAAA1	TAACI	CAATI	ragg'	TTTA	GAC	GAAA	CCTA	GAAA	CGA1	ICAC1	GATI	CCAAI	ITTG	AGTAC	сстст	GCAC	GAT	GAC	CTCA	CAG	AAA(TGAA	GAGT	(129)
AACGGCA	GCAAG	TGAT	TTGA	GTTC	AGTAC	STTC	CTCAT	ATA	AAATI	ATTO	GACTO	CTAG	AGAT	TAG	TAATA	TGG	GAAG	GACA	AAATI	GTTI	CAAC	CACO	GAC	46AAI	:066/	AGCO	0000	TTCT	(249)
TTCAAAG	AGAGO	GAGGA	40666	GTTA	TTCAC	CATT	TCAT	TGA	rggto	AGAG	36C6/	AATT	GAAAG	STTA	GCAG	6T G G (BAATI	ICTA	AAGA1	TTCCC	C066	GGAA	AAA	FAGA	ATG	CTC	TAC	STTAC	(369)
CCATAAT	ATGTO	36AA(STAT	CGAC	BTAAT		ATAGA	AGTC/	ATTCO	GTCI	GAA1	TOCT	ACATE	BAAGA	ACAT	TAAB	CAG	TGA	COGAA	AC666	BAAGA		166A'	IGTA	SAAGA	TCAT	AACA	TGAG	(489)
TOMITLO	0CH0		JOHT	ICH II		41411	LCHU	CHI	0100	HCI	CHI	IC IN		thr	35P	met.	.HIU PTO	1.014	alv.	the	ala	ile	his.	250	ile	alu	ile	thr	(807)
AGAAGCT	TGTA	CAGTI	TTGG	GAAG	BGGTI	TTTG	ATTG	ATCA	AAAGA	AGA	ATCT	ACTT	CA	ACC	GAT	ATG	ĊĊĊ	ŤŤĂ	ĞĞČ	ACG	ĞČČ	ĂŤĂ	CAT	AAC	ĂŤĂ	GAA	ĂŤĊ	ACA	966
leu sly CTT GGA	lys ars AGG	aja GGT	51y 60a	⊴ln C A A	leu TTA	ala GCT	ars AGA	ala GCA	ala GCG	sly GGT	ala GCT	val GTA	ala GCG	lys AAA	leu CTG	ile ATT	ala GCA	1 y s 888	slu GAG	⊴1¥ 666	195	ser TCG	ala GCC	thr ACA	leu TTA	195 888	leu TTA	614 100	1056
	A			ser																									
ser sly TCT 666	⊈lu GAG	val GTC	ars CGT	1eu TTG	ile ATA	ser TCC	195 AAA	asn AAC	TOC	ser TCA	ala GCA	thr ACA	val GTC	⊴ 1¥ 66A	sln CAA	val 616	41¥ 666	asn AAT	val GTT	sly GGA	val GTG	asn AAC	sin CAG	lys AAA	AGĢ	leu TTG	⊈ 1¥ 861	ars Aga	1146
-11-							••••																net						
GCC GGA	TCT	AAG	CGT	TGG	CTA	GGT	AAG	CGT	CCT	GTA	GTA	AGA	GGA	GTA	GTT	ATG	AAC	ĉĊŤ	GTA	GAC	CAC	ĉĊĊ	ACG	666	6TG	616	AAG	GGA	1236
slu sly pro	sin	leu	val	alu	lys	asn ala	pro leu	≰ln	919	leu	sly	val	ile	leu	his	leu	şlu	slu	slu	val	≸lu	lvs	alv	ile	asn	ile	val	ile	
A CCC	CAA	TTG	GTA	GAA	AAA	GCC AA	CTA	CAA	CCC	CTT	666	GTT	ATC	CTG	CAC	TTG	GAA	GAA	GAA	GTA	GAA	AAA	GGA	ATA	AAT	ATA	616	ATA	1326
ile leu ATT TTA	1eu TTA	⊁he TTC	val GTC	asp GAC	val GTA	val Sta	asn [.] AAT	ars AGG	1 ys Aga	¶lu GAA	asn AAT	≰lu GAA	asn A <u>a</u> t	ars Ağa	ile ATT	ser Agt	Phe TTC	phe TTC	val GTÇ	phe TTT	thr ACA	ŢÅA,				AAAI	AGGA	GTAA	1413
0	rps 11			Ľ					0					н			lys		'	asp		A 4	AAA	thr	111				
TTAACT GCT	net GTG	thr ACA	CGT	ser TCA	leu CTA	1ys AAA	195 AAA	asri AAT		phe TTT	val GTA	ala GCG c	asn AAT	his Cat	leu TTA	leu TTA	ars Aga A	195 AAA	ile ATC	⊈lu GAG T	1 ys AAG	leu CTT	asn AAC	195 AAA C	AAG a	ala GCG	⊴lu GAA	195 AAA	1503
⊴lu ile GAA ATA	ile ATA	val GTA	thr ACT	t r p T 66	ser TCA	ars CGG	ala GCA	ser	thr ACC	ile ATT	ile ATA	222	thr ACA	net ATG	ile ATT	⊈1 y 66C	his CAT	thr ACG	ile ATT	ala GCT	ile ATC	his CAT	asn AAT	⊈1y 66A	lys ars AGA	≰lu GAA	his CAC	leu TTG	1593
					C										c	t									A	1	b		
CCC ile	tyr TAT	ile ATA	thr ACA S	əsp Gat	CGT	aet ATG	val GTA	≝19 661 c	his CAT C	lys AAA	leu TTA	⊈ly 66C a	⊈lu GAA	Phe TTT	ala GCA	ĉćą	thr ACT	leu CTC t a	asn AAT	Fhe TTC	879 C66	#1 y 66A	his Cat	ala GCA	lys AAA	asn AAC G	85P Gat	asn AAT	1683
ars lys ser	875 797	816 197	т і .	617				AT7																					1700
GA	501		(HA	311		THA	0116		1461	HHIF	TRAI	AHA	1 1 1 1 1 1	10016	10 I I F	TLA	I I I A	I AA	1 00AL	50 I A f	HU ()	1.416	9999		IAAA	INRA	IMAUF	MAAU	1/48
AAAGCAGA	AAABGCAGAATTIGATAGAGAGGGTTACTCGATATCTATGTCTGTGACAAAGCCCGAAGAGTAATTGATCAGATCCGTGCGTTCCTACGCCGAAACACTTATGATATTAGAACTTCA 191 TGCCCTATCGAGCATGTTATCCCATTITTAAATTGATTT 1957						1918																						

Figure 3. Nucleotide sequence of part of the Bam 11 fragment of spinach chloroplast DNA. The sequence commences at the BamH1 site within the inverted repeat and reads through the left junction (marked by an arrow head) into the large single copy region. Coding sequences for trnl (underlined), rpl2 and rps19 are shown. The sequence is compared to two other sequences. The region from the BamH1 site to the beginning of the rps19 coding region (coordinates 1 - 1420) is shown compared to the spinach sequence (including insertions of 14, 6, 666 and 10 nucleotides and

6550

a deletion, indicated as dashes, of 6 nucleotides). The <u>rps</u>19 coding sequence (coordinates 1420-1698) is shown compared to the analogous sequence from <u>N</u>. <u>tabacum</u> as determined by Sugita and Sugiura (8). Differences that are neutral with respect to protein coding are indicated in lower case type. The bracketed numbers are relative to the first nucleotide of the <u>N</u>. <u>debneyi rpl</u>2 insertion.

nucleotide being 4 bp from the junction (Figure 4b).

There are two nucleotide differences between $tRNA_{GUG}^{His}$ from spinach and <u>N</u>. <u>debneyi</u>, with the <u>N</u>. <u>debneyi</u> sequence being identical to that found in maize chloroplasts (9). Both changes occur in loop regions of the cloverleaf structure. Comparison between the 144 bp spinach and the 360 bp <u>N</u>. <u>debneyi</u> <u>psb</u>A-<u>trn</u>H1 intervals reveals almost complete divergence except for approximately 60 bp distal to <u>psb</u>A which can form a stable stem and loop structure in the mRNA and may function as the <u>psb</u>A transcription terminator (3).

The Gene for the S19 Ribosomal Protein Spans the Left Junction of the Large Single Copy Region and the Inverted Repeat

Sugita and Sugiura have reported the sequence of a gene in <u>N</u>. <u>tabacum</u> chloroplast DNA that codes for a protein homologous to <u>E</u>. <u>coli</u> ribosomal protein S19 (8). They mapped the gene to the left margin of the large single copy region (8,8a). Analysis of the spinach chloroplast DNA Bam11 sequence spanning the left inverted repeat - large single copy region junction shows an open reading frame commencing at a GTG codon at position 1420 and ending at a translation stop codon at position 1696 (Figure 3). There is 89% homology between this nucleotide sequence (called <u>rps</u>19) and the tobacco S19 gene sequence. There is 92% homology between the predicted amino acid sequences of the S19 proteins of spinach and tobacco and 58% homology between the spinach and <u>E</u>. <u>coli</u> S19 proteins (Figure 5). The spinach S19 protein has a molecular weight of 10,577.

As the boundary between the spinach chloroplast left inverted repeat and large single copy region is at the 49th codon of <u>rps</u>19, the 44 C terminal codons lie in the large single copy region (Figure 3). For reasons given below, we think that, in <u>N</u>. <u>debneyi</u> also, <u>rps</u>19 spans the junction of the left inverted repeat and the large single copy region.

The trnH1 Non-coding Strand Specifies a Protein whose N Terminus is Identical to the S19 Protein

As the inverted repeats have identical sequences, a region coding for the same 48 N terminal amino acids as occur in S19 must also be present at the end of the right-hand inverted repeat in spinach. Examination of the d.

<u>Figure 4</u>. Nucleotide sequence of the <u>rps19'</u> and <u>trnH1</u> coding regions from (a) Bam 5 fragment of spinach chloroplast DNA and (b) the analogous sequence from <u>N</u>. <u>debneyi</u> chloroplast DNA. Numbering starts within the inverted repeat at the translation initiation codon for <u>rps19'</u> and ends at the complement of the translation stop codon (boxed) for <u>psbA</u> (3). The arrows indicate the junction between the right-hand inverted repeat and the large single copy region. The underlined sequence is the complementary strand of <u>trnH1</u>.

sequence at the right-hand junction shows that the 48 codons of <u>rps</u>19 within the inverted repeat continue in phase with a further 66 codons specified partly by the non-coding strand of <u>trn</u>H1 and partly by the <u>trn</u>H1-<u>psb</u>A interval (Figure 4a). This open reading frame therefore encodes a protein (ca. MW 12,513) whose N terminus is that of S19 and whose C terminus is entirely different to S19. We refer to this protein as S19' and to the open reading frame as <u>rps</u>19'.

In <u>N</u>. <u>debneyi</u> the sequence at the right-hand junction (Figure 4b) has an open reading frame of 27 codons starting with a GTG. The first 20 predicted amino acids, with the exception of a lysine/arginine difference, are identical to the first 20 N-terminal amino acids of spinach S19' and S19 (Figure 4a, b). Six of the remaining 7 codons are specified by the non-coding strand of <u>trn</u>H1 and the predicted amino acid sequence bears no similarity to the spinach S19' sequence. The molecular weight of the S19' protein derived from the <u>N</u>. <u>debneyi</u> sequence is 3135, a value very different from that of the putative spinach S19' protein.

In view of the overall similarity of the <u>N</u>. <u>debneyi</u> and <u>N</u>. <u>tabacum</u> (8a) sequences, it is reasonable to assume that in the former, as in <u>N</u>. <u>tabacum</u> and in spinach, <u>rps</u>19 is located close to the junction of the large single copy region and the left-hand inverted repeat. The fact that a sequence almost identical to part of <u>rps</u>19 occurs in the region of the right-hand junction in <u>N</u>. <u>debneyi</u> indicates that that sequence must Lie

а	60
Spinach	MTRSLKKNPFVANHLLRKIEKLNKKAEKEIIVTWSRASTIIP TMIGHTIÄIHNGREHLPI
Tobacco	**************************************
Soybean	**************************************
Maize	*** $K \times [\times \times \times \times \times K \times X \times X \times \times X \times X \times X \times X \times$
<u>E. COTT</u>	P*****G**IDL*****************************
	90
Spinach	YITDRMVGHKLGEFAPTLNFRGHAKNDNKSRR
Tobacco	****S*********************************
<u>E. coll</u>	F V**E**********************************
L.	
D	
N debnevi	MAINLINIJIJJJINNGAV UVNJNPRNNLI JGURRGGNARNARGIIIARNRGGGNRR
\underline{E} . \underline{coli}	*VVKC*PTSPGR*HVVKVVNPELHKGKPFAPLLEKNSKSG***NN*R**T**I*****Q
Spinach	LYRKIDFRRNEKDIYGKIVTIEYDPNRNAYICLIHYGDGEKRYILHPRGAIIGDTIVSGT
E. coli	A**IV**K**KDG*PAVVERL*****S*N*A*VL*K***R***A*K*LKA**Q*Q**V
Spinach	EVPIKMGNALPLTDMPLGTAIHNIEITLGRGGQLARAAGAVAKLIAKEGKSATLKLPSGE
<u>N. debneyi</u>	**************************************
<u>E. coli</u>	DAA**P**T**MRNI*V*STV**VEMKP*K*****S**TYVQIV*RD*AYV**R*R***
Sninach	236 VPL ISKNCSATVONVONVOVNKPLOPACSKPWLOKPDVVPCVVMNDVDHPTCVVKCCPD
N. debnevi	**S***********************************
<u>E</u> . <u>coli</u>	M*KVEAD*R**L*E***AEHMLRV**KA*AA**R*VRPT*R*TA****H*G*GHE*RNF
Sninach	I VEKAL ODI ČVTI HI FEEVĖKOTNTVTTI I EVOVVNOKENENOTSEEVET
N. debnevi	****ND********************************
E. coli	GKHPV T*W**QTKGKKTRSNKRTDKF*VRRRSK

Figure 5. Amino acid sequences of the S19 and L2 ribosomal proteins. (a) Comparison of the S19 sequences from spinach, tobacco (8), soybean (11) and maize (9) chloroplasts and from E. coli (10). (b) Comparison of the L2 sequences from spinach and N. debneyi chloroplasts and from E. coli (ref. 13). Only partial sequences are available for S19 from soybean and maize. Amino acids identical to those of spinach are indicated with an asterisk. Two small gaps have been introduced into the sequences to maximize homology.

within the inverted repeat, and the point at which the <u>N</u>. <u>debneyi</u> and spinach S19' sequences become dissimilar (codon 20, Figure 4b) must mark the junction of the inverted repeat and the large single copy region. This leads us to conclude that, in <u>N</u>. <u>debneyi</u>, in contrast to the situation in <u>N</u>. <u>tabacum</u> (8a) <u>rps</u>19 spans the left-hand junction, the first 20 codons lying within the inverted repeat.

The Gene for a Spinach Chloroplast Protein Homologous to the E.coli Ribosomal Protein L2 is Immediately Proximal to rps19

In E. coli rps19 is part of the S10 operon which encodes, in order of

transcription, the genes for ribosomal proteins S10, L3, L4, L23, L2, S19, L22, S3, L16, L29, and S17 (12). Analysis of the sequence of Bam11 proximal to rps19 on spinach chloroplast DNA shows an open reading frame of 286 codons (position 538-1398, Figure 3) terminating 21 bp prior to the rps19 translation start codon. If translation begins at the first available ATG codon, the molecular weight of the predicted protein would be 31,832. This closely approximates the size of the 31,500 dalton E. coli L2 ribosomal proximal rps19 (Zurawski, protein whose codina reaion maps to coli L2 protein sequence (13: unpublished). Comparison of the E. Zurawski, unpublished) with the spinach 286 codon open reading frame shows an overall level of homology of 40%, the value for residues 40 to 200 being 53% (Figure 5b). On the basis of this homology we propose that the open reading frame is the gene for the chloroplast equivalent of the E. coli L2 protein and accordingly designate it rpl2. The product of rpl2 is probably the spinach chloroplast ribosomal protein L4 described by Dorne et which comigrates on 2-dimensional gels and cross-reacts al (14)immunologically with the L2 protein of E. coli.

The gene for tRNA^[le]_{CAT}, <u>trn</u>11, occurs 447 bp 5' to, and on the same coding strand as spinach chloroplast <u>rpl</u>2 (Figure 3). The sequence of this gene which has a methionine anticodon has been reported previously (15). There is no open reading frame of significant size in the <u>trn</u>11-<u>rpl</u>2 interval which, excluding a 14 bp insert, is 92% homologous between spinach and <u>N</u>. <u>debneyi</u> (Figure 3). The sequence of <u>trn</u>11 in the two species is identical. N. debneyi rpl2 has a 666 bp Intervening Sequence

Figure 3 shows the sequence of approximately 2 kb 5' to <u>rps</u>19' in the right-hand inverted repeat of <u>N</u>. <u>debneyi</u>. The sequence is 685 bp longer than the corresponding sequence in spinach, most of the difference being accounted for by a single 666 bp insertion in the coding region of L2 between codons 128 and 129 (spinach numbering). Translation of the sequence on either side of the insertion gives open reading frames which, when combined, predict a protein 97% homologous to the spinach L2 protein (Figures 3 and 5b). The only gross changes between the two L2 coding regions are a two-codon insertion in <u>N</u>. <u>debneyi</u> close to the N terminus and a TTA \rightarrow TGA change which removes 22 C terminal residues from the <u>N</u>. <u>debneyi</u> protein relative to spinach L2 (Figure 5b) and results in an 89 bp <u>rpl2-rps</u>19 interval compared to a 21 bp interval in spinach. There is no substantial open reading frame in the insertion.

DISCUSSION

Sequence Organization Around the Inverted Repeat - Large Single Copy Junctions

We have established that rps19, the gene for the S19 ribosomal protein, spans the left-hand inverted repeat - large single copy junction in spinach chloroplast DNA. The same arrangement applies also for rps19 in N. debneyi except that the exact location of the junction within the gene is different. Recently the sequence of the two junctions of the large single copy region with the inverted repeats of soybean chloroplast DNA was published (11). Examination of this sequence shows that rps19 in soybean also spans the left-hand junction, 23 N-terminal amino acids being encoded in the inverted repeat. In maize chloroplast DNA an open reading frame which can now be identified as part of rps19 was reported (9) as occurring in an EcoRI fragment derived entirely from within the inverted repeat region. The sequence of the 33 amino acids reported are 76% homologous to spinach S19 (Figure 5a). As yet, the location of this sequence relative to single copy region of maize has not been established. the large Hybridization studies with E. coli rps19 DNA indicate that in Chlamydomonas reinhardtii chloroplast DNA rps19 is located in the large single copy region (16) as it is in N. tabacum (8). Thus the border of the inverted repeat-large single copy region seems to be a movable feature and its location may be such that it includes, partially includes, or excludes the gene for the S19 protein.

A consequence of a protein coding region straddling a repeat unit - single copy junction is that there will be a second open reading frame in the genome, one end of which will be identical to part of the first, but whose other end will be completely different. Thus in spinach and N. debneyi we have identified S19' coding regions for 114 and 27 amino acids respectively, and a 52-codon S19' sequence can be identified in the soybean chloroplast DNA sequence published bv Spielmann and Stutz (11). Extensive restriction enzyme mapping (1,11,15) and our sequence data, which extends over 400 bp from the junction, indicate perfect sequence identity between the two inverted repeats of any particular genome. Assuming sequence identity, as the 5' end of rps19/rps19' is within the inverted repeat both rps19 and rps19' will be transcribed and (providing the mRNAs have equal stability) translated with equal efficiency. Although there is a high probability that S19' proteins are synthesised by chloroplasts, we think that the great variation in their size and their C-terminal sequences indicates they are non-functional.

The location of the gene for $tRNA^{His}$ with respect to the right-hand junction varies by only a few nucleotides between spinach, <u>N</u>. <u>debneyi</u> and soybean (11). In the latter two species it lies entirely within the large single copy region, terminating 5 nucleotides and 1 nucleotide, respectively, prior to the junction. In spinach, the 3' terminal residue is actually the first nucleotide of the inverted repeat. However, in maize, although the junction has not yet been exactly defined, <u>trn</u>H is located entirely within the inverted repeat (9).

Intervening Sequence in rpl2

The major difference between the spinach and <u>N</u>. <u>debneyi</u> sequences is the presence of the 666 bp insert in the coding region of <u>rpl2</u> of <u>N</u>. <u>debneyi</u>. Intervening sequences have been reported to occur in chloroplast tRNA genes (2), in the large ribosomal RNA gene from <u>Chlamydomonas</u> <u>reinhardtii</u> chloroplasts (2), and in the genes for the large subunit of ribulose bisphosphate carboxylase (17) and the elongation factor EF-TU in <u>Euglena</u> chloroplasts (18). However, this is the first report of an intervening sequence in a gene for a protein in higher plant chloroplasts although the presence of introns in what may prove to be protein-coding sequences in <u>Vicia</u> faba chloroplast DNA has recently been detected by R-loop analysis (19). Although we have not sequenced the <u>rpl2</u> gene from the left-hand inverted repeat of <u>N</u>. <u>debneyi</u> chloroplast DNA there is no reason to suspect that it does not also contain the insert and that both copies of <u>rpl2</u> are not functional.

The insertion sequence is shown in Figure 3 to lie between a leucine and a threonine codon. Allocation of this site was determined by matching the <u>N</u>. <u>debneyi</u> sequence with the spinach sequence. However the insertion sequence could start 1 bp to the right of this position without affecting the deduced L2 amino acid sequence. In the latter case the 5' splice site sequence would be GA/GTGCGG which matches the consensus sequence proposed for the 5' splice site of introns in protein-coding genes of <u>Euglena</u> <u>gracilis</u> chloroplast DNA (19a). There is no obvious resemblance between the sequence at the 3' splice site and the 3' consensus splice site of <u>E</u>. <u>gracilis</u> introns (19a) nor with the consensus splice site of introns in animal genes (20).

The intervening sequence in <u>rpl2</u> of <u>N</u>. <u>debneyi</u> does not bear any similarity to the introns of chloroplast tRNA genes (2). There is no direct repeat or inverted repeat sequence associated with the ends of the insert so it cannot be equated directly to a transposable element. Given the great homology between spinach and <u>N</u>. <u>debneyi</u> <u>rpl2</u>, the 666 bp intervening

sequence is probably of recent origin and its presence in other plant species can be readily determined by hybridization studies.

Chloroplast Ribosomal Proteins

Of the 50 or so chloroplast ribosomal proteins, 10-14 of the small subunit proteins and 5-8 of the large subunit proteins are synthesised in isolated chloroplasts and are therefore probably encoded in the chloroplast DNA (21,22). Genes for the following ribosomal proteins have now been identified in chloroplast DNA on the basis of their homology with E. coli sequences: S4 in maize (23); S19 in N. tabacum (8), N. debneyi and spinach; L2 in N. debneyi and spinach. None of these genes appears to be organized as part of a large operon analogous to the organization of many E. coli ribosomal protein genes (12). However, given the proximity of rpl2 and rps19 in spinach and N. debneyi chloroplast DNA it is likely that these genes comprise a single transcriptional unit which may even include the closely linked trn11. A 10 bp inverted repeat (positions 1701-1710 and 1715-1724, Figure 3) immediately beyond the TAA stop codon of rps19 probably defines the end of this transcriptional unit. There is a 60-codon open reading frame further downstream (position 1772-1951, Figure 3) but the predicted amino acid sequence has no homology with the sequence predicted by a similarly located open reading frame in N. tabacum (8) and it is therefore unlikely to be functional.

Irrespective of the nature of the <u>rpl</u>² and <u>rps</u>¹⁹ mRNAs, the peculiar arrangement of these genes in spinach and <u>N</u>. <u>debneyi</u> results in two copies of <u>rpl</u>² and one copy of <u>rps</u>¹⁹ per chloroplast chromosome. If L2 and S19 proteins are present in equimolar amounts in the ribosomes some adjustment must occur regarding the translation of the two coding regions. In this regard a GGAG sequence, which is complementary to the 3¹ end of 16S rRNA and which may serve to facilitate ribosome binding, is present proximal to the <u>rps</u>¹⁹ translation initiation codon but not to the <u>rpl</u>² initiation codon.

* The nomenclature for chloroplast genes used in this paper follows the recommendations proposed by R.B. Hallick and W. Bottomley (1983) Plant Mol. Biol. Reporter 1, 38-43.

REFERENCES

- 1. Bedbrook, J.R. and Kolodner, R. (1979) Ann. Rev. Plant Physiol. 30, 593-620.
- Whitfeld, P.R. and Bottomley, W. (1983) Ann. Rev. Plant Physiol. 34, 279-310.

3.	Zurawski, G., Bohnert, H.J., Whitfeld, P.R. and Bottomley, W. (1982) Proc. Natl. Acad. Sci. USA 79, 7699-7703.
4.	Driesel, A.J., Speirs, J. and Bohnert, H.J. (1980) Biochim. Biophys. Acta 610, 297-310.
5.	Whitfeld, P.R. and Bottomley, W. (1980) Biochem. Internat. 1, 172-178.
6.	Herrmann, R.G., Whitfeld, P.R. and Bottomley, W. (1980) Gene 8, 179-191.
7.	Driesel, A.J., Crouse, E.J., Gordon, K., Bohnert, H.J. and Herrmann, R.G. et al. (1979) Gene 6, 285-306.
8.	Sugita, M. and Sugiura, M. (1983) Nucleic Acids Res. 11, 1913-1918.
8a.	Sugita, M., Kato, A., Shimada, H. and Sugiura, M. (1984) Mol. Gen. Genet. 194, 200-205.
9.	Schwartz, Z., Jolly, S.O., Steinmetz, A.A., and Bogorad, L. (1981) Proc. Natl. Acad. Sci. USA 78, 3423-3427.
10.	Yaquchi, M. and Wittmann, H.G. (1978) FEBS Lett. 88, 227-230.
11.	Spielmann, A. and Stutz, E. (1983) Nucleic Acids Res. 11, 7157-7167.
12.	Lindahl, L., Archer, R., and Zengel, J.M. (1983) Cell 33, 241-248.
13.	Kimura, M., Mende, L., and Wittmann-Liebold, B. (1982) FEBS Lett.
	149, 304-312.
14.	Dorne, A.M., Eneas-Filho, J., Heizmann, P. and Mache, R. (1984)
	Mol. Gen. Genet. 193, 129-134.
15.	Kashdan, M.A. and Dudock, B.S. (1982) J. Biol. Chem. 257, 11191-11194.
16.	Watson, J.C. and Surzycki, S.J. (1983) Current Genetics, 7, 201-210.
17.	Stiegler, G.L., Matthews, H.M., Bingham, S.E. and Hallick, R.B. (1982) Nucleic Acids Res. 10, 3427-3444.
18.	Montandon, P.E. and Stutz, E. (1983) Nucleic Acids Res. 11, 5877-5892.
19.	Koller, B. and Delius, H. (1984) Cell 36, 613-622.
19a.	Koller, B., Gingrich, J.C., Stiegler, G.L., Farley, M.A., Delius,
	H., and Hallick, R.B. (1984) Cell 36, 545-553.
20.	Lewin, B. (1980) Cell 22, 324-326.
21.	Eneas-Filho, J., Hartley, M.R. and Mache, R. (1981) Mol. Gen. Genet. 184, 484-488.
22.	Schmidt, R.J., Richardson, C.B., Gillham, N.W. and Boynton, J.E. (1983) J. Cell Biol. 96, 1451-1463.
23.	Subramanian, A.R., Steinmetz, A., and Bogorad, L. (1983) Nucleic Acids Res. 11, 5277-5286.