Theor Appl Genet (2000) 101:724–732

D.H. Xu · J. Abe · M. Sakai · A. Kanazawa
Y. Shimamoto

Sequence variation of non-coding regions of chloroplast DNA of soybean and related wild species and its implications for the evolution of different chloroplast haplotypes

Received: 16 December 1999 / Accepted: 12 February 2000

Abstract Soybean chloroplast DNAs (cpDNAs) are classified into three types (I, II and III) based on RFLP profiles. Type I is mainly observed in cultivated soybean (*Glycine max*), while type II and type III are frequently found in both cultivated and wild soybean (*Glycine soja*), although type III is predominant in wild soybean. In order to evaluate the diversity of cpDNA and to determine the phylogenetic relationship among different chloroplast types, we sequenced nine non-coding regions of cpDNA for seven cultivated and 12 wild soybean accessions with different cpDNA types. Eleven single-base substitutions and a deletion of five bases were detected in a total of 3849 bases identified. Five mutations distinguished the accessions with types I and II from those with type III, and seven were found in the accessions with type III, independently of their taxa. Four species of the subgenus *Glycine* shared bases that were identical to those with types I and II at two of the five mutation sites and shared bases that were identical to those with type III at the remaining three sites. Therefore, the different cpDNA types may not have originated monophyletically, but rather may have differentiated from a common ancestor in different evolutionary directions. A neighbor-joining tree resulting from the sequence data revealed that the subgenus *Soja* connected with *Glycine microphylla* which formed a distinct clade from *Clycine clandestina* and the tetraploid cytotypes of *Glycine tabacina* and *Glycine tomentella*. Several informative length mutations of 54 to 202 bases, due to insertions or deletions, were also detected among the species of the genus *Glycine*.

Communicated by R. Hagemann

D.H. Xu · J. Abe · M. Sakai · A. Kanazawa · Y. Shimamoto (✉)
Laboratory of Plant Genetics and Evolution,
Faculty of Agriculture, Hokkaido University,
Sapporo 060-8589, Japan
e-mail: yshimamo@res.agr.hokudai.ac.jp
Fax: + 81-11-706-4933

Key words Soybean · Chloroplast DNA · Non-coding region · DNA sequencing

Introduction

The chloroplast genome of cultivated soybean, *Glycine max*, is approximately 151 kb in length, and includes an inverted repeat of approximately 23 kb (Palmer et al. 1983; Spielman and Stutz 1983). Its structure and gene order are similar to those found in the great majority of legumes, but the soybean genome lacks a 78-kb inversion found in mung bean and common bean (Palmer et al. 1988). Close et al. (1989) reported six cpDNA hyplotypes, designated plastome groups I to VI, in cultivated and wild soybeans by RFLP analysis with mung bean cpDNA clones. They found that group I was predominant in modern cultivars, and group III was frequent in primitive cultivars, weedy and wild soybean. Based on a network of the plastome groups constructed by RFLP data, Close et al. (1989) showed that groups I and II branched off from groups III to VI, breaking the subgenus *Soja* into two major groups, each with sub-groups. They assumed that group I most likely was established through at least two independent mutations, from group III via group II. Shimamoto et al. (1992) have also detected cpDNA RFLPs in soybean by using a sugarbeet cpDNA clone as a probe and classified their collection into three haplotypes (types I, II and III), each with a different combination of *Eco*RI and *Cla*I RFLPs detected with the probe. These three RFLP types appeared to be identical to those detected by Close et al. (1989), and the types I, II and III of Shimamoto et al. (1992) corresponded to the plastome groups I, II and III–VI of Close et al. (1989), respectively. Based on the results of sequence analysis, Kanazawa et al. (1998) verified that the mutational events characterizing the three types of Shimamoto et al. (1992) are two single-base substitutions: one in the non-coding region between *rps*11 and *rpl*36, and the other in the 3′ part of the coding region of *rps*3.

The geographical distribution of the three chloroplast tpyes (I, II and III) has been extensively surveyed for cultivated and wild soybeans collected from various regions of East Asia (Shimamoto et al. 1992, 1998; Abe et al. 1999; Shimamoto et al. 2000). These studies confirmed the finding of Close et al. (1989) and indicated a markedly different distribution pattern of the three types between *G. max* and *Glycine soja*. Type I which is the predominant type in *G. max*, was rarely observed in *G. soja*; it was found only in four geographically separated sites in Japan (Abe et al. 1999). All of the wild plants with type I had phenotypic characteristics of typical wild soybean. Abe et al. (1999) assumed that the wild plants with type I may be either derivatives of hybridization between cultivated and wild soybeans or relics of a direct progenitor of cultivated soybean with type I. On the other hand, cultivated soybeans with type III, the most predominant in *G. soja*, were furhter classified into three hyplotypes based on the RFLP profiles of mitochondrial DNAs (Shimamoto et al. 2000). This result indicated that some cytoplasmic haplotypes of cultivated soybean might derive from wild plants with the same cytoplasmic genome, which occurred in different regions. The findings mentioned above further prompt us to examine molecular differences of the different cpDNA types in more detail and to resolve their phylogenetic relationships.

Analyses of non-coding regions of cpDNA have been employed to elucidate phylogenetic relationship of different taxa (Olmstead and Palmer 1994). Compared with coding regions, non-coding regions may provide more informative characters in phylogenetic studies at the species level because of their high variability due to the lack of functional constraints. Non-coding regions of cpDNA have been assayed either by direct sequencing (Manen and Natall 1995; Jordan et al. 1996; Sang et al. 1997; Small et al. 1998; McDade and Moody 1999; Molvray et al. 1999) or by restriction-site analysis of PCR products (PCR-RFLP) (Demesure et al. 1996; Wolfe et al. 1997; Asmussen and Liston 1998; Cipriani et al. 1998; Friesen et al. 1999). We sequenced non-coding regions of cpDNA from cultivated and wild soybeans, as well as four species of the subgenus *Glycine*, in this study. Our purposes were to evaluate the variation of non-coding regions of the cpDNA of soybean and its related wild species, and to explore the phylogenetic relationship of the different cpDNA types.

Materials and methods

Plant materials

Seven soybean cultivars introduced from Japan and China and 12 wild soybean accessions collected from Japan, China, and South Korea were sequenced. These cultivars and accessions were selected in terms of the introduction/collection sites and the chloroplast types identified by RFLP profiles. Four species of the subgenus *Glycine*, *Glycine tabacina*, *Glycine tomentella*, *Glycine microphylla* and *Glycine clandestina*, were included in this study. Of the four species, *G. tabacina* and *G. tomentella* are tetraploid cytotypes which were collected in Taiwan (Table 1).

Sequence analysis

Nine non-coding regions (intergenic spacers), whose expected lengths were suitable for PCR and its subsequent sequencing analysis, were assayed (Fig. 1). These regions are scattered over the whole cpDNA genome: eight are located in the large single-copy region and the remaining one is located in the small single-copy region. Some of these regions, such as *trn*H-*psb*A, *trn*L-*trn*F and *atp*B-*rbc*L, have been extensively analyzed for phylogenetic stud-

Table 1 Species and accessions of the genus *Glycine* used in this study

Species	Accession	Origin	cpDNA type[a]
G. max	Waseohsaya	Japan, Ibaraki	I
	Aogozen	Japan, Chiba	II
	Chasengoku 13	Japan, Mie	III
	Hualaidou	China, Gansu	I
	Wuhuasiyuehuang	China, Guangdong	II
	Shichengqingpidou	China, Jiangxi	III
	Peking	China	III
G. soja	B09055	Japan, Kumamoto	I
	B09051	Japan, Oita	I
	B07009	Japan, Ehime	II
	B01076	Japan, Hokkaido	III
	B07108	Japan, Kochi	III
	N23302	China, Jiangxi	II
	N23339	China, Fujian	II
	N23239	China, Anhui	III
	B00097	S. Korea, Kyongsangnam-do	II
	B00055	S. Korea, Kyongsangnam-do	III
	B00104	S. Korea, Cheju Island	III
	B00100	S. Korea, Chungchongbuk-do	III
G. tabacina	D00018	Taiwan, Penghu Islands	
G. tomentella	E00009	Taiwan, Taitung	
G. microphylla	IL449PK#1	USDA	
G. clandestina	PI2557452	USDA	

[a] Three cpDNA types (I, II and III) were defined based on the RFLP profiles obtained when DNAs digested by *Eco*RI and *Cla*I were hybridized with a 10.9-kb cpDNA fragment (Shimamoto et al. 1992)

ies of plant species (Manen and Natall 1995; Sang et al. 1997; McDade and Moody 1999; Molvray et al. 1999). The primer pairs were designed on the basis of the sequence data of cpDNA from tobacco and soybean (Table 2). DNA extraction was carried out as described by Doyle and Doyle (1990). The PCR reaction mixture contained 30 ng of genome DNA, 0.25 μM of 5′ and 3′ end-primers, 100 μM of nucleotides, 1 u *Tag* polymerase, 1 × PCR buffer containing 50 mM of KCl, 10 mM of Tris-HCl pH 8.3, and 1.5 mM $MgCl_2$ in a total volume of 50 μl. Cycling consisted of three steps: (1) denaturation at 94°C for 5 min and annealing at 50–60°C (depending on the kinds of primers) for 5 min, (2) 30 cycles at 94°C for 1 min, 50–60°C for 2 min and 72°C for 2 min, and (3) a final extension at 72°C for 10 min. The PCR reaction was performed with a GeneAmp PCR System 9700 (Perkin Elmer). PCR products were directly subjected to sequencing with dRhodamine Terminator sequencing kit (PE Applied Biosystems) on an ABI 377 Sequencer followed the manufacturer's instructions.

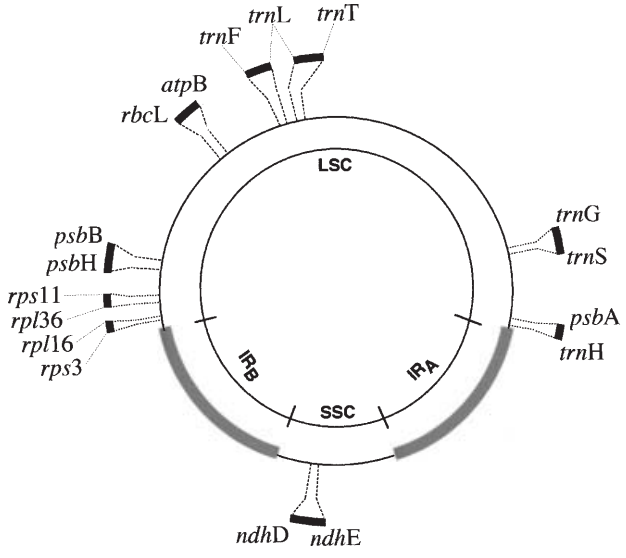

Fig. 1 Nine non-coding regions sequenced in this study

Data analysis

The DNA sequences were aligned using the program GENETYX-MAC (Version 8.0, Software Development Co., Ltd., 1995). Kimura's (1981) two-parameter estimates of the evolutionary distance were calculated using the DNADIST program of PHYLIP (version 3.573c; Felsenstein 1995). The phylogenetic tree was constructed with the neighbor-joining method (Satou and Nei 1987) based on the distance matrix by using the NEIGHBOR program of PHYLIP (version 3.573C; Felsenstein 1995).

Results

Sequence variation in the subgenus *Soja*

For each of the nine non-coding regions, a single PCR product with the same mobility in agarose-gel electrophoresis was observed in all of the 19 accessions tested, indicating that there is no large length polymorphism in the subgenus *Soja*. A total of 3849 bases were identified in the nine regions. The G + C content among the regions ranged from 11.6% for the region *rps*11–*rpl*36 to 32.3% for the regions *psb*B–*psb*H and *ndh*D–*ndh*E, with an average of 24.0% (Table 3).

Twelve mutations were detected in the nine non-coding regions in the subgenus *Soja* (Table 3). These include the mutations in the region *rps*11–*rpl*36 previously observed in 'Peking': a single-base substitution at the *Eco*RI site which discriminates types I and II from type III, and a deletion of one copy of direct repeats of five bases (AAAAT) (Kanazawa et al. 1998). The sequence variations observed were not due to a mismatch in the PCR amplification, since they appeared in repeated experiments. No mutation was detected in the regions *atp*B–*rbc*L and *ndh*D–*ndh*E, although the former region has been often sequenced in phylogenetic studies of plant species (Hodges and Arnold 1994; Manen and

Table 2 Primers used for amplification of nine non-coding regions of soybean cpDNA

Region	Primer sequence (5′ to 3′)[a]	Annealing temperature	Reference
*trn*H–*psb*A	f: TGATCCACTTGGCTACATCCGCC	60°C	Shinozaki et al. 1986 (tobacco)
	r: GCTAACCTTGGTATGGAAGT		Speilmann et al. 1983 (soybean)
*trn*S–*trn*G	f: GATTAGCAATCCGCCGCTTT	60°C	Shinozaki et al. 1986 (tobacco)
	r: TTACCACTAAACTATACCCGC		Shinozaki et al. 1986 (tobacco)
*trn*T–*trn*L	f: GGATTCGAACCGATGACCAT	60°C	This study (soybean)
	r: TTAAGTCCGTAGCGTCTACC		Shinozaki et al. 1986 (tobacco)
*trn*L–*trn*F	f: TCGTGAGGGTTCAAGTCC	56°C	Shinozaki et al. 1986 (tobacco)
	r: AGATTTGAACTGGTGACACG		Shinozaki et al. 1986 (tobacco)
*atp*B–*rbc*L	f: GAAGTAGTAGGATTGATTCTC	58°C	Shinozaki et al. 1986 (tobacco)
	r: CAACACTTGCTTTAGTCTCTG		Shinozaki et al. 1986 (tobacco)
*psb*B–*psb*H	f: AGATGTTTTTGCTGGTATTGA	56°C	Shinozaki et al. 1986 (tobacco)
	r: TTCAACAGTTTGTGTAGCCA		Shinozaki et al. 1986 (tobacco)
*rps*11–*rpl*36	f: GTATGGATATATCCATTTCGTG	50°C	Kanazawa et al. 1998 (soyben)
	r: TGAATAACTTACCCATGAATC		Kanazawa et al. 1998 (soybean)
*rpl*16–*rps*3	f: ACTGAACAGGCGGGTACA	50°C	Kanazawa et al. 1998 (soybean)
	r: ATCCGAAGCGATGCGTTG		Kanazawa et al. 1998 (soybean)
*ndh*D–*ndh*E	f: GAAAATTAAGGAACCCGCAA	56°C	Shinozaki et al. 1986 (tobacco)
	r: TCAACTCGTATCAACCAATC		Shinozaki et al. 1986 (tobacco)

[a] f, forward primer; r, reverse primer

Table 3 Mutations in the nine non-coding regions of cpDNA observed in the subgenus *Soja*

Regions	Length of sequence tested (bases)	CG content (%)	Mutation		
			Code	Position	Type
*trn*H–*psb*A	237	22.4			
			#1	6 bp downstream *psb*A	Transversion
			#2	130 bp downstream *psb*A	Transversion
*trn*S–*trn*G	477	26.0			
			#1	13 bp upstream *trn*S	Transition
			#2	181 bp upstream *trnS*	Transversion
*trn*T–*trn*L	588	21.1			
			#1	197 bp upstream *trn*T	Transversion
			#2	319 bp upstream *trn*T	Transversion
			#3	519 bp upstream *trn*T	Transversion
*trn*L–*trn*F	410	25.6			
			#1	202 bp downstream *trn*L	Transition
*atp*B–*rbc*L	538	25.5			
			Not detected		
*psb*B–*psb*H	576	32.3			
			#1	156 bp downstream *psb*B	Transversion
*rps*11–*rpl*36	268	11.6			
			#1[a]	35 bp downstream *rpl*36	Transversion
			#2	166 bp downstream *rpl*36	Deletion
*rpl*16–*rps*3	179	19.0			
			#1	75 bp upstream of *rpl*16	Transversion
*ndh*D–*ndh*E	576	32.3			
			Not detected		

[a] *Eco*RI polymorphism site detected previously (Kanazawa et al. 1998)

Table 4 Bases at the 11 mutation sites in the subgenus *Soja* and four species of the subgenus *Glycine*

Species	cpDNA type	Mutation sites											
		*trn*H–*psb*A		*trn*S–*trn*G		*trn*T–*trn*L			*trn*L–*trn*F	*psb*B–*psb*H	*rps*11–*rpl*36		*rpl*16–*rps*13
		#1	#2	#1	#2	#1	#2	#3	#1	#1	#1[b]	#2	#1
G. max	I (2[a])	C	C	A	A	C	C	G	A	A	A	AAAAT	G
	II (2)	C	C	A	A	C	C	G	A	A	A	AAAAT	G
	III-common (2)	A	C	A	C	A	C	T	A	C	C	AAAAT	G
	III-Peking (1)	A	C	A	C	A	C	T	A	C	C	−−−−−[c]	T
G. soja	I (2)	C	C	A	A	C	C	G	A	A	A	AAAAT	G
	II (4)	C	C	A	A	C	C	G	A	A	A	AAAAT	G
	III-common (3)	A	C	A	C	A	C	T	A	C	C	AAAAT	G
	III-B00055 (1)	A	C	T	C	A	C	T	T	C	C	AAAAT	G
	III-B07108 (1)	C	A	A	C	A	A	T	A	C	C	AAAAT	G
	III-B00100 (1)	A	C	A	C	A	C	T	A	C	C	−−−−−	G
G. tabacina		A	C	T	A	A	C	G	A	C	C	AAAAT	G
G. tomentella		A	C	T	A	A	C	G	T	C	C	AAAAT	G
G. microphylla		A	A	A	–	A	C	G	A	C	C	AAAAT	G
G. clandestina		A	C	T	A	A	C	G	T	C	C	AAAAT	G

[a] Number in parenthesis indicated the number of accessions tested
[b] Mutation involved in the *Eco*RI restriction site (Kanazawa et al. 1998)
[c] Deletion

Natall 1995; Small et al. 1998; McDade and Moody 1999). All of the mutations except for the deletion of five bases were single-base substitutions: three in the region *trn*T–*trn*L, two in each of the regions *trn*H–*psb*A and *trn*S–*trn*G, and one in each of the regions *trn*L–*trn*F, *psb*B–*psb*H, *rps*11–*rpl*36 and *rpl*16–*rps*3. The proportion of single-base substitutions in the non-coding regions was thus calculated as 2.9×10^{-3} per base. Nine of the eleven point mutations were either A/C or T/G substitutions, indicating that transversions were more prevalent than transitions in the soybean chloroplast genome.

Besides the mutation at the *Eco*RI site of *rps*11–*rpl*36, two point mutations, #1 of *trn*H–*psb*A and #1 of *trn*T–*trn*L, were respectievly involved in the recognition sites of two restriction enzymes (*Mfl*I and *Ssp*I). The bases at these two mutation sites can thus be easily determined by using the PCR-RFLP method (Xu et al., unpublished).

Relationship of the different cpDNA types

Five mutations, #2 of *trn*S–*trn*G, #1 and #3 of *trn*T–*trn*L, #1 of *psb*B–*psb*H and #1 of *rps*11–*rpl*36 (the *Eco*RI site), distinguished the accessions with types I and II from those with type III, independently of their taxa (*G. max* or *G. soja*) (Table 4). All of the accessions with types I and II had identical bases for all of the 12 mutation sites. The bases in mutation #1 of *trn*H–*psb*A were also different between accessions with types I and II and those with type III except for a Japanese wild accession (B07108), which possessed the same base (C) as the accessions with types I and II. The other seven mutations were detected only in the accessions with type III. Five of the nine accessions with type III have the same haplotype of non-coding regions and thus could be considered as a common one of type III. A Japanese wild accession (B07108) differed from the common haplotype by three mutations (#1 and #2 of *trn*H–*psb*A and #2 of *trn*T–*trn*L), being the most-remote type in the accessions with type III. A Chinese cultivar 'Peking' and a Korean wild accession (B00055) each differed from the common haplotype by two mutations. Another Korean wild accession (B00100) differed from the common haplotype by one mutation. The 12 mutations in the non-coding regions and the mutation at the *Cla*I site of the coding region of *rps*3, which differentiated type I from types II and III (Kanazawa et al. 1998), are schematically presented in Fig. 2.

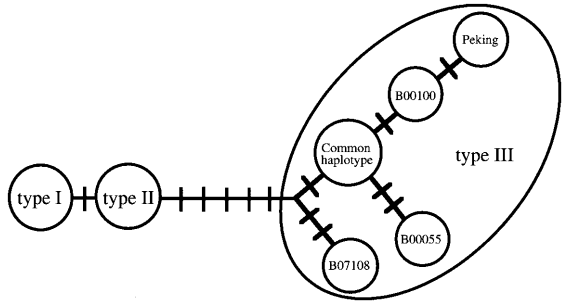

Fig. 2 Relationship of different cpDNA types (I, II and III) in the subgenus *Soja* as revealed by sequencing analysis of nine non-coding regions. *Eco*RI and *Cla*I restriction-site polymorphisms, which have been used for the classification of cpDNA types I, II and III (Kanazawa et al. 1998), are included in this figure. '*Common haplotype*' means the majority of type-III accessions

Comparison with species of the subgenus *Glycine*

Eight of the nine regions, except for *ndh*D–*ndh*E, were sequenced for four species of the subgenus *Glycine*. In contrast to the low variability in the subgenus *Soja*, many variations, including single-base and short-sequence substitutions and length polymorphisms, were found between the two subgenera and among the four species of the subgenus *Glycine*. The proportion of single-base substitutions between species of the subgenus *Soja* and species of the subgenus *Glycine* was, on average, 2.3×10^{-2} per base. On the other hand, the proportion of substitutions between the four species of the subgenus *Glycine* ranged from 5.0×10^{-3} per base between *G. tomentella* and *G. tabacina* to 2.5×10^{-2} per base between *G. tabacina* and *G. microphylla* with an average of 1.7×10^{-2} per base. An example of the sequence variation among the different species in the region *trn*H–*psb*A is presented in Fig. 3.

Fig. 3 Aligned sequences of the non-coding region *trn*H–*psb*A in five species of the genus *Glycine*. *Asterisks* indicated the same nucleotide as in *G. soja* and *dashes* indicate gaps. *M* means A or C

```
G. soja          C-TTATACTATGTAAAAATGATCTATATATAAAATTCTATCCTTTCGTTCGTTATCA------
G. tabacina      CT**********************--**C*******GT*****G**********------
G. tomentella    CT**********************--**C*******GT*****G**********------
G. microphylla   C-**********************--**A*******GC*****T**********------
G. clandestina   CT**********************--**C*******GT*****G**********TTCTTT

G. soja          -----TTCTTTTTTCCTTTTAAGATAGGAAAAATGCTAAAGAATCGCGAAAGAAATMAAAAAC
G. tabacina      -----**C******T***C***************T********T*T*********C******
G. tomentella    -----**C******T***C***************T********T*T*********C******
G. microphylla   -----**G******T***A***************C********C*T*********A******
G. clandestina   TTTTC**C******T***C***************T********T*T*********C******

G. soja          TTATGTATAAATATTTATAAATAAAAAGATTACTAATCAAATAAATAAAAGTAAAGGGGCAAT
G. tabacina      G***G*****C*****A******AAA**********************A************
G. tomentella    G***G*****C*****A******AAA**********************A************
G. microphylla   G***A*****A*****A******TTT**********************G************
G. clandestina   G***A*****C*****T******AAA**********************G************

G. soja          ATCAAAAAAGTTGATATTGCCTTTTACTTTCAAAAACTAATCTACCTTAAGATCMAAATT
G. tabacina      *****CTTTT**********C***************C*****CCTT******A*****
G. tomentella    *****CTTTT**********C***************C*****CCTT******A*****
G. microphylla   *****CTTTT**********C***************C*****ATAA******A*****
G. clandestina   *****AAAAG**********C***************C*****CCTT******A*****
```

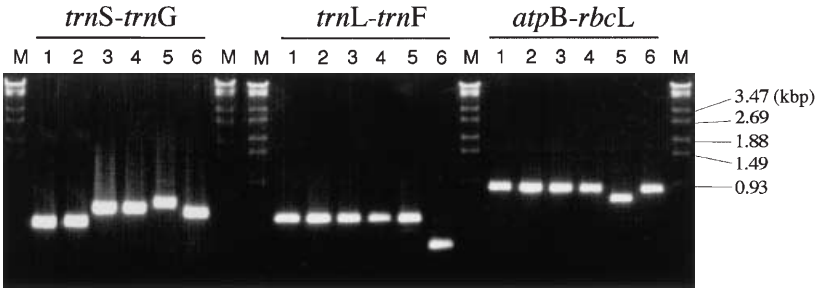

Fig. 4 Large length mutations detected in the three non-coding regions (*trn*S–*trn*G, *trn*L–*trn*F and *atp*B–*rbc*L) of cpDNA of the subgenus *Soja* and four species of the subgenus *Glycine*. *M*, λ/*Sty*I marker; *1 G. max; 2 G soja; 3 G. tabacina; 4 G. tomentella; 5 G. microphylla; 6 G. clandestina*. The PCR products were separated on a 1.5% agrose gel

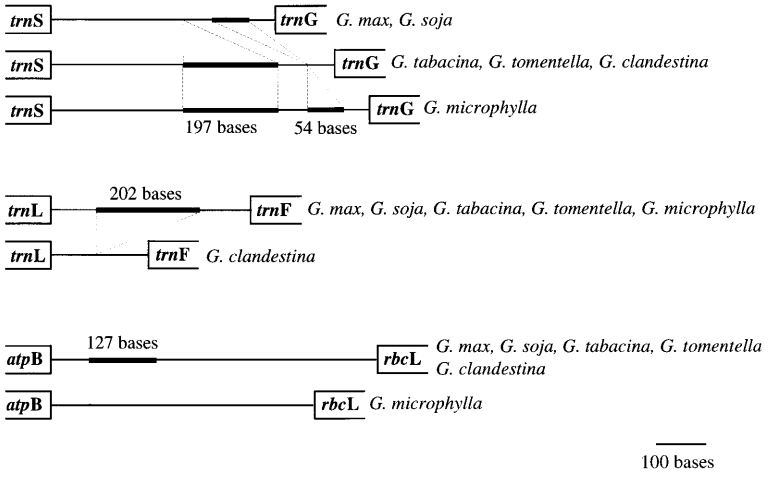

Fig. 5 Four large length mutations detected in three non-coding regions (*trn*S–*trn*G, *trn*L–*trn*F and *atp*B–*rbc*L) of cpDNA in the subgenus *Soja* and four species of the subgenus *Glycine*. The *heavy lines* indicate the deletion/insertion detected. A *scale bar* represents the length of non-coding regions

We compared the 12 mutation bases in the subgenus *Soja* with those at the corresponding mutation sites in the four species of the subgenus *Glycine* (Table 4). The four species of the subgenus *Glycine* shared the same bases with types I and II at two (#2 of *trn*S–*trn*G and #3 of *trn*T–*trn*L) of the five critical sites, which distinguished types I and II from type III, and shared the same bases with type III at the remaining three sites [#1 of *psb*B–*psb*H, #1 of *trn*T–*trn*L and #1of *rps*11–*rpl*36 (the *Eco*RI site)]. The four species also had the same base as was present in all of the accessions with type III except for a wild accession (B07108) at the mutation site #1 of *trn*H–*psb*A. Three of the six mutations found in the accessions with type III, #2 of *trn*H–*psb*A, #1 of *trn*S–*trn*G *and #1 of trn*L–*trn*F, were also observed in some species of the subgenus *Glycine*.

Four large length mutations were also detected in the six species including *G. max* and *G. soja* in three regions: *trn*S–*trn*G, *trn*L–*trn*F and *atp*B–*rbc*L (Fig. 4). Sequence analysis revealed a deletion of 202 bases in the region *trn*L–*trn*F in *G. clandestina*, and a deletion of 127 bases in the region *atp*B–*rbc*L in *G. microphylla*. In the region *trn*S–*trn*G, *G. tabacina*, *G. tomentella* and *G. clandestina* showed a deletion of 54 bases, whereas *G. max* and *G. soja* had a deletion of 197 bases in the region downstream from the deletion found in the former three species. Interestingly, *G. microphylla* possessed both of the two segments, producing the longest sequence in this region (Fig. 5).

A phylogenetic tree was constructed with the neighbor-joining method for the sequence data of five non-coding regions, *trn*H–*psb*A, *trn*T–*trn*L, *psb*B–*psb*H, *rpl*16–*rps*3 and *rps*11–*rpl*36, in which no large length mutation was detected (Fig. 6). Types I, II and III in the subgenus *Soja* had a well-supported clade and connected with *G. microphylla* which formed a distinct clade from the other species in the subgenus *Glycine*. *G. tabacina* and *G. tomentella* expressed a relatively close relationship. The four large length mutations were consistently overlaid on the tree obtained (Fig. 6).

Discussion

Features of the variations in the non-coding regions of the chloroplast genome in the subgenus *Soja*

A total of 12 mutations, 11 single-base substitutions and one deletion of five bases, were detected in the 3849 nucleotide bases identified in the nine non-coding regions for the 19 cultivated and wild soybean accessions tested in this study. The proportion of single-base substitutions in the non-coding regions in the subgenus *Soja* was estimated as 2.9×10^{-3} per base. This value is thus slightly lower than that between *G. tabacina* and *G. tomentella* obtained in this study (5.0×10^{-3}). The lower value in the subgenus *Soja* is not unexpected since the two species, *G. max* and *G. soja*, are often considered congener-

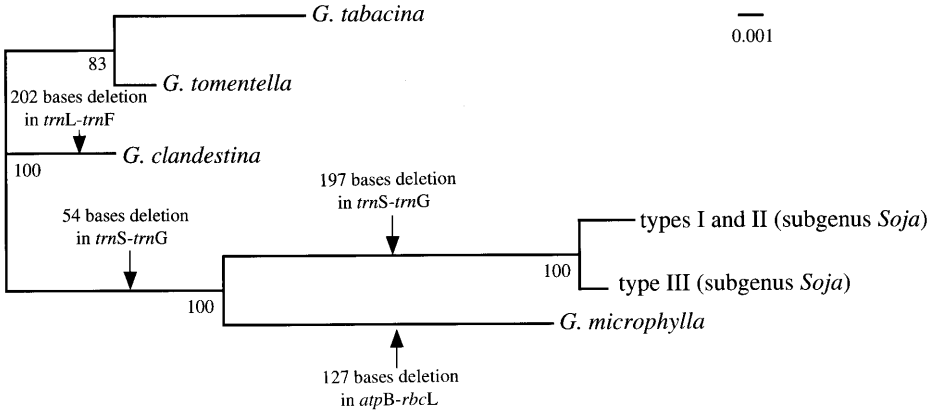

Fig. 6 A phylogenetic tree (neighbor-joining) of types I, II and III in the subgenus *Soja* and four species of the subgenus *Glycine* based on the sequence of five non-coding regions (*trn*H–*psb*A, *trn*T–trn*L*, *psb*B–*psb*H, *rps*11–*rpl*36 and *rpl*16–*rps*3) in which no large mutations were involved. *Numbers* indicate the percentages of grouping confidence calculated by bootstrap analysis. Arrows the size and locations of the four large length mutations. The *scale bar* represents the genetic distance

ic because of the lack of any constant reproductive barriers between them (Singh and Hymowitz 1988).

Of the 11 single-base substitutions detected in the subgenus *Soja*, nine were either A/C or T/G transversions. Morton and Clegg (1993) noted that the ratio of transversions to transitions is much lower in sequences with a lower A + T content in the non-coding regions of cpDNA. Morton (1995) analyzed the frequency of base substitutios for the two sequentially lined-up sequences with different A + T contents and found that the ratio of transversions to transitions increased with an increasing A + T content. The high ratio of transversions to transitions observed in this study may also be due to the characteristics of the sequences with a high A + T content.

A deletion of one copy of the two tandem repeats of five bases (AAAT) was observed. This may be an example of the deletion or duplication of short sequences which is caused by slipped-strand mispairing, a major source of variation among species in non-coding regions of chloroplast DNAs (Morton and Clegg 1993). Aldrich et al. (1988) indicated that this kind of deletion or duplication is usually in a direct orientation, and the repeated sequences are mostly rich in A + T content. Ham et al. (1994) found that 18 of the 19 putative insertions of more than three bases, which were observed in three non-coding regions of cpDNA of the *Crassulaceae* and related species, were parts of direct-repeat motifs. This kind of deletion or duplication was frequently observed when the species of the subgenus *Glycine* were included in the analysis.

Phylogenetic relationship of different cpDNA types and their origins

The sequencing of the non-coding regions of cpDNA demonstrates that types I and II were distinguished from type III. There are at least five base substitutions between types I, II and type III. This result was further supported by the data obtained from an expanded analysis of the four mutations, #1 of *trn*H–*psb*A, #1 of *trn*T–*trn*L, #3 of *trn*T–*trn*L and #1 of *psb*B–*psb*H, by using the PCR-RFLP method with more materials (Xu et al., unpublished). These studies suggest that chloroplast types I and II were not derived monophyletically from type III as previously assumed (Close et al. 1989; Kanazawa et al. 1998).

A comparison with the four species of the subgenus *Glycine* further supports the finding mentioned above. The species of the subgenus *Glycine* shared bases that were identical to those with types I and II at two of the five critical sites, and shared bases that were identical to those with type III at the remaining three sites. This indicated that types I, II and type III might be derived from a common ancestor in different evolutionary directions unless parallel mutations occurred.

The same is true for the two wild accessions (B00055 and B07108) tested in this study. The Korean accession (B00055) possessed two unique bases at the mutation sites #1 of *trn*S–*trn*G and #1 of *trn*L–*trn*F, both of which were the same as those in some of the species of the subgenus *Glycine*. In addition, the Japanese accession (B07108) has a unique base at the mutation site #2 of *trn*H–*psb*A, which was in common with that of *G. microphylla* (Table 4). The sequence diversity of the chloroplast genome in the subgenus *Soja* might have partly originated from a variability that pre-existed in an ancestral population from which the two subgenera had diverged.

Phylogenetic relationship between the subgenera *Soja* and *Glycine*, and among the four species of the subgenus *Glycine*

In contrast to the low cpDNA variation within the subgenus *Soja*, a great deal of variation, including single-base and short-sequence substitutios and insertion/deletion of short sequences, was observed between the subgenera *Soja* and *Glycine* and among the four species of the subgenus *Glycine*. The proportion of single-base substitutions between the four species of the subgenus *Glycine* ranged from 5.0×10^{-3} between *G. tabacina* and *G. tomentella* to 2.5×10^{-2} between *G. tabacina* and *G. microphylla*. The latter value is almost the same as the average of estimates between the species of two subgenera.

In addition, the phylogenetic tree obtained with the neighbor-joining method for the sequence data also revealed that *G. microphylla* formed a distinct clade from the other three species of the subgenus *Glycine* and connected with the clade of the subgenus *Soja*. Thus, *G. microphylla* appears to have branched from the other species examined in this study at an early stage of evolution in the subgenus *Glycine*. On the other hand, *G. tabacina* and *G. tomentella* expressed a relatively close relationship. The accessions of these two species were tetraploid cytotypes collected in Taiwan. More particularly, the tetraploid cytotype of *G. tabacina* in Penghu Islands is non-stoloniferous (Tateishi and Ohashi 1992), and was found to belong to the same plastome group (A) with *G. tomentella* and *G. clandestina* (Doyle et al. 1990a, b). Taking these findings into consideration, the topology of the phylogenetic tree obtained in this study is almost in accordance with that of Doyle et al. (1990b) obtained by RFLP analysis.

The four large length mutations were separately located in different branches of the phylogenetic tree (Fig. 6). Although the mechanisms of such large mutations remain to be determined, these large mutations may provide informative markers for tracing the phylogeny among species of the genus *Glycine*. As represented by the two length mutations in the region *trn*S–*trn*G, *G. tabacina, G. tomentella* and *G. clandestina* had the deletion of 54 bases, and *G. max* and *G. soja* had the deletion of 197 bases in the region downstream from that found in the former three species, whereas *G. microphylla* possessed both of the two segments (Fig. 5). The sequences in the region *trn*S–*trn*G of the subgenus *Soja* and those of the three species of the subgenus *Glycine* may therefore have been derived from an ancestral sequence common to *G. microphylla* via the deletions of 55 and 197 bases, respectively. This result supports the phylogenetic relationship on the neighbor-joining distance tree. Presently, 16 species are classified in the subgenus *Glycine* (Hymowtz et al. 1998). Our ongoing study with a greater number of species of the subgenus *Glycine* will give us a better understanding of the phylogenetic relationship and evolutionary process between the subgenera *Soja* and *Glycine.*

Acknowledgements This research was supported by the Japan Society for the Promotion of Science (JSPS). We are grateful to Professor Y. Sano (Faculty of Agriculture, Hokkaido University) and Professor J. Gai (Soybean Research Institute, Nanjing Agricultural University) for their considerable assistance. The experiments of this study comply with the current laws of Japan.

References

Abe J, Hasegawa A, Fujushi H, Mikami T, Ohara M, Shimamoto Y (1999) Introgression between wild and cultivated soybean of Japan revealed by RFLP analysis for chloroplast DNAs. Econ Bot 53:286–291

Aldrich J, Cherney BW, Merlin E, Christopherson L (1988) The role of insertions/deletions in the evolution of the intergenic region between *psb*A and *trn*H in the chloroplast genome. Curr Genet 14:137–146

Asmussen CB, Liston A (1998) Chloroplast DNA characters, phylogeny, and classification of *Lathyrus* (Fabaceae). Am J Bot 85:387–401

Cipriani G, Testolin R, Gardner R (1998) Restriction-site variation of PCR-amplified chloroplast DNA regions and its implication for the evolution of *Actinidia*. Theor Appl Genet 96:389–396

Close PS, Shoemaker RC, Keim P (1989) Distribution of restriction site polymorphism within the chloroplast genome of the genus *Glycine*, subgenus *Soja*. Theor Appl Genet 77:768–776

Demesure B, Comps B, Petit RJ (1996) Chloroplast DNA phylogeography of the common beach (*Fagus sylvatica* L.) in Europe. Evolution 50:2515–2520

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12:13—15

Doyle JJ, Doyle JL, Grace JP, Brown AHD (1990a) Reproductively isolated polyploid races of *Glycine tabacina* (*Leguminosae*) had different chloroplast genome donors. Systematic Bot 15:173–181

Doyle JJ, Doyle JL, Brown AHD (1990b) A chloroplast-DNA phylogeny of the wild perennial relatives of soybean (*Glycine* subgenus *Glycine*): congruence with morphological and crossing groups. Evolution 44:371–389

Felsenstein J (1995) PHYLIP (Phylogeny inference package) version 3.57c. University of Washington Press, Seatle

Friesen N, Pollner S, Bachmann K, Blattner FR (1999) RAPDs and noncoding chloroplast DNA reveal a single origin of the cultivated *Allium fistulosum* from *A. altaicum* (*Alliaceae*). Am J Bot 86:554–562

Ham RC, Hart H, Mes THM, Sandbrink JM (1994) Molecular evolution of noncoding regions of the chloroplst genome in the *Crassulaceae* and related species. Curr Genet 25:558—566

Hodges SA, Arnold ML (1994) Columbines: a geographically widespread species flock. Proc Natl Acad Sci USA 91:5129–5132

Hymowtz T, Singh RJ, Kollipara KP (1998) Biosystematics of the genus *Glycine*, 1997. Soybean Genet Newslett 25:114

Jordan WC, Courtney MW, Neigel JE (1996) Low levels of intraspecific genetic variation at a rapidly evolving chloroplast DNA locus in North American duckweeds (*Lemnaceae*). Am J Bot 83:430–439

Kanazawa A, Tozuka A, Shimamoto Y (1998) Sequence variation of chloroplast DNA that involves *Eco*RI and *Cla*I restriction site polymorphisms in soybean. Genes Genet Syst 73:111–119

Kimura M (1981) Estimation of evolutionary distance between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Manen JF, Natall A (1995) Comparison of the evolution of ribulose-1, 5-bisphosphate carboxylase (*rbc*L) and *atp*B—*rbc*L non-coding spacer sequences in a recent plant group, the tribe Rubieae (Rubiaceae). J Mol Evol 41:920–927

McDade LA, Moody ML (1999) Phylogenetic relationship among Acanthaceae: evidence from noncoding *trn*L–*trn*FF chloroplast DNA sequences. Am J Bot 86:70—80

Molvray M, Kores PJ, Chase MW (1999) Phylogenetic relationships within *Korthalsella* (*Viscaceaek*) based on nuclear ITS and plastid *trn*L-F sequence data. Am J Bot 86:249–260

Morton BR (1995) Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. Proc Natl Acad Sci USA 92:9717–9721

Morton BR, Clegg MT (1993) A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near *rbc*L in the grass family (*Poaceae*). Curr Genet 24:357–365

Olmstead RG, Palmer JD (1994) Chloroplast DNA systematics: a review of method and data analysis. Am J Bot 81:1205–1224

Palmer JD, Singh GP, Pilly DTN (1983) Structure and sequence evolution of three legume chloroplast DNAs. Mol Gen Genet 190:13–19

Palmer JD, Osorio B, Thompson WF (1988) Evolutionary significance of inversions in legume chloroplast DNAs. Curr Genet 14:65–74

Saitou N, Nei M (1987) The neighbor-joing method: a new method of reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sang T, Crawford DJ, Stuessy TF (1997) Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). Am J Bot 84:1120–1136

Shimamoto Y, HAsegawa A, Abe J, Ohara M, Mikami T (1992) *Glycine soja* germplasm in Japan: isozymes and chloroplast DNA variation. Soybean Genet Newslett 19:73–77

Shimamoto A, Fukushi H, Abe J, Kanazawa A, Gai J, Gao Z, Xu D (1998) RFLPs of chloroplast and mitochondrial DNA in wild soybean, *Glycine soja*, growing in China. Genet Res Crop Evol 45:433–439

Shimamoto Y, Abe J, Gao Z, Gai J, Thseng F (2000) Characterizing the cytoplasmic diversity and phyletic relationship of Chinese landraces of soybean, *Glycine max*, based on RFLPs of chloroplast and motochondrial DNA. Genet Res Crop Evol (in press)

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, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Singh RJ, Hymowitz T (1988) The genomic relationships among *Glycine max* (L.) Merr. and *G. soja* Sieb. and Zucc. as revealed by pachytene chromosome analysis. Theor Appl Genet 76:705–711

Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. Am J Bot 85:1301–1315

Spielmann A, Stutz E (1983) Nucleotide sequence of soybean chloroplast DNA regions which contain the *psb*A and *trn*H genes and cover the ends of the large single-copy region and one end of the inverted repeats. Nucleic Acids Res 11:7157–7167

Spielmann A, Ortiz W, Stutz E (1983) The soybean chloroplast genome: construction of a circular restriction site map and location of DNA regions encoding the genes for rRNAs, the large subunit of the ribulose-1, 5-bisphosphate carboxylase and the 32 kDa protein of the photosystem II reaction center. Mol Gen Genet 190:5–12

Tateishi Y, Ohashi H (1992) Taxonomc studies on *Glycine* of Taiwan. J Jpn Bot 67:127–147

Wolfe AD, Elisens WJ, Watson LE, Depamphilis CW (1997) Using restriction-site variation of PCR-amplified cpDNA genes for phylogenetic analysis of Tribe Cheloneae (*Scrophulariaceae*). Am J Bot 84:555–564