Plant Omics Journal POJ

POJ 8(1):24-29 (2015) ISSN:1836-3644

In silico analysis of simple sequence repeats (SSRs) in chloroplast genomes of *Glycine* species

Ibrahim Ilker Ozyigit[1], Ilhan Dogan[2], Ertugrul Filiz[3*]

[1]Marmara University, Faculty of Science and Arts, Department of Biology, 34722, Goztepe, Istanbul, Turkey
[2]Izmir Institute of Technology, Faculty of Science, Department of Molecular Biology and Genetics, 35430, Urla, Izmir, Turkey
[3]Duzce University, Cilimli Vocational School, Department of Crop and Animal Production, 81750, Duzce, Turkey

[*]Corresponding author: ertugrulfiliz@gmail.com

Abstract

Microsatellites, also known as simple sequence repeats, are short (1-6 bp long) repetitive DNA sequences present in chloroplast genomes (cpDNAs). In this work, chloroplast genomes of eight species (*Glycine canescens, G. cyrtoloba, G. dolichocarpa, G. falcata, G. max, G. soja, G. stenophita,* and *G. tomentella*) from *Glycine* genus were screened for cpSSRs by utilisation of MISA perl script with a repeat size of ≥10 for mono-, 5 for di-, 3 for tri-, tetra-, penta- and hexa-nucleotide, including frequency, distributions, and putative codon repeats of cpSSRs. According to our results, a total of 1273 cpSSRs were identified and among them, 413 (32.4%) were found to be in genic regions and the remaining (67.6%) were all located in intergenic regions, with an average of 1.04 cpSSRs per kb. Trinucleotide repeats (45%) were the most abundant motifs, followed by mononucleotides (36%) and dinucleotides (11.8%) in the plastomes of the *Glycine* species. In genic regions, trimeric repeats, the most frequent one reached the maximum of 70.7%. Among the other repeats, mono- and tetrameric repeats were represented in proportions of 25.7% and 3.6%, respectively. Interestingly, there were no di-, penta-, and hexameric repeats in coding sequences. The most common motifs found in all plastomes were A/T (97.8%) for mono-, AT/AT (98%) for di-, and AAT/ATT (41.5%) for trinucleotides. Among the chloroplast genes, *ycf1* had the highest number of cpSSRs, and *G. cyrtoloba* and *G. falcata* species had the maximum number of genes containing cpSSRs. The most frequent putative codon repeats located in coding sequences were found to be glutamic acid (21.2%), followed by serine (15.5%), arginine (8.3%) and phenylalanine (7.8%) in all species. Also, tryptophan, proline, and aspartic acid were not detected in all plastomes.

Keywords: *Glycine*, chloroplast genome, cpSSRs, *in silico* analysis, bioinformatic analysis.
Abbreviations: cpDNA_chloroplast DNA; SSR_simple sequence repeats; MISA_ MIcroSAtellite identification tool; EST_expressed sequence tags

Introduction

The genus *Glycine*, which is a member of Leguminosae family, includes two subgenera namely glycine and soja. The subgenus soja contains annual cultivated soybean *G. max* and its presumed wild progenitor *G. soja*, native to northeastern Asia. These species are diploid (2n=40) and interfertile (Hymowitz and Singh, 1987; Sakai et al., 2003; Carter et al., 2004). *Glycine* subgenus is composed of 16 wild perennial species, indigenous to Australia and Papua New Guinea (Hymowitz, 1970; Doyle et al., 2004). These species have various chromosome numbers, including diploid (2n=40), tetraploid (2n=80), aneudiploid (2n=38), and aneutetraploid (2n=78) (Singh and Hymowitz 1985). Chloroplasts have their independent genome encoding some proteins used in phtosythesis and many other metabolic activities. The size of chloroplast genomes are ranged from 110 to 200 kb, bearing about 30-50 different RNA and a few protein coding genes (Sugiura, 1995). Chloroplast genome structure is highly conserved in terrestrial plants and contains two inverted repeats (IR) with large-single-copy (LSC) and small single-copy (SSC) regions (Palmer, 1990). It is accepted that gene duplication is the main force for genetic variation and could lead to formation of new genes and gene functions (Schmidt and Davies, 2007). Gene duplications can also affect organelle evolution by positive selections in chloroplast genes (Erixon and Oxelman, 2008). Microsatellites or simple sequence repeats are tandem repeated motifs (1–6 bp long) and are distributed throughout the three eukaryoti genomes: nucleus, chloroplast and mitochondria (Powell et al., 1995; Soranzo et al., 1999). The mono-, di-, tri- and tetranucleotide repeats are accepted as main types of microsatellites (Ellegren, 2004) and length of the microsatellite is one of the most important factors affecting mutation rate. The potential utilization of SSRs in plant molecular genetics was first demonstrated at the beginning of 1990s in the *Glycine* subgenus *soja*, which is a subdivision of the annual cultivated soybean *G. max* and its presumed wild progenitor *G. soja* (Akkaya et al., 1992; Morgante and Olivieri, 1993). Lately, SSR markers have been effectively used in different *Glycine* taxa. In a study, population genetic structures of 77 wild Japanese soybean populations (*G. soja*) were analyzed using 20 microsatellite primers (Kuroda et al., 2006). SSR and SNP elements (single-nucleotide polymorphism) elements were used to analyze genetic diversity in *G. max* and *G. soja* from

Table 1. Details of chloroplast genomes for *Glycine* species.

Plant Species	Genome size (bp)	Accession Number*	G+C content (%)
Glycine canescens	152518	NC_021647	35.33
Glycine cyrtoloba	152381	NC_021645	35.31
Glycine dolichocarpa	152804	NC_021648	35.31
Glycine falcata	153023	NC_021649	35.33
Glycine max	152218	NC_007942	35.37
Glycine soja	152217	NC_022868	35.38
Glycine stenophita	152618	NC_021646	35.32
Glycine tomentella	152728	NC_021636	35.33

*Genbank database (http://www.ncbi.nlm.nih.gov/genome/).

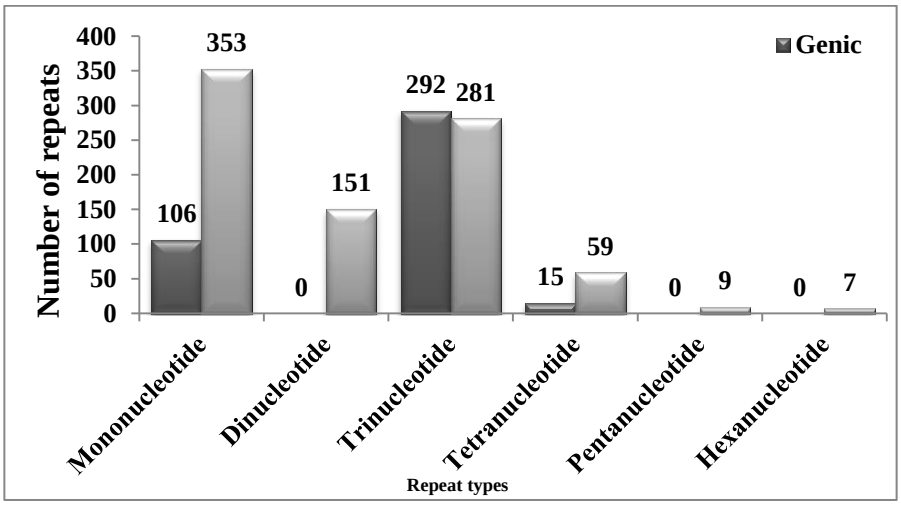

Fig 1. Total number of different repeats in genic and intergenic regions of all *Glycine* species.

four geographic regions of China (Li et al., 2010). In another study, the natural population structures and genetic diversities of 40 wild soybean (*G. soja*) populations obtained from China were analyzed by using 20 microsatellites (Guo et al., 2014). In addition to nuclear SSRs, chloroplast SSRs are commonly used in *Glycine* taxa studies for evaluation of population genetic structure and genetic diversity levels (Powell et al., 1996a; Shimamoto, et al., 2000; Xu et al. 2002; He et al., 2012). An interesting feature of cpSSRs is their non-recombinant, uniparentally inherited nature and they consist of typically mononucleotide motifs repeating 8 to 15 times (Navascués and Emerson, 2005). The main objective of this study was to identify SSRs in chloroplast genomes of *Glycine* species for estimating of their occurrence and distribution in both coding and noncoding regions. Also, putative amino acid repeats were investigated in coding SSR regions.

Results

Presence and frequency of cpSSRs

A total of eight chloroplast genomes of *Glycine* species were screened for existence of cpSSRs and a total of 1273 cpSSRs were identified, of which 413 (32.4%) were in genic regions and 860 (67.6%) were in intergenic regions (Table 2), based on the annotated coding sequences of *Glycine* species in NCBI genome database. Also, an average frequency of cpSSR was found to be 1.04 cpSSR per kb and G+C (%) frequency contents of the *Glycine* species were closely similar, ranged from 35.31 to 35.38 (Table 1). Data from the chloroplast genomes of *Glycine* species showed that total numbers of cpSSRs ranged from 153 (in *G. soja* and *G. stenophita*) to 174 (in *G. falcata*) (Table 2). Among the *Glycine* species, the highest numbers of cpSSRs (56) were detected in coding regions of *G. max* and *G. soja*, whereas the lowest number of cpSSRs (41) was detected in coding regions of *G. cyrtoloba*. There were no di-, penta-, and hexameric repeats in genic regions of all plastomes (Table 2). Presence of mono-, di-, tri-, and tetra-, and absence of penta- and hexameric repeats were confirmed in all *Glycine* species by our study. Also, penta- and hexameric repeats were not observed in plastomes of *G. canescens*, *G. cyrtoloba*, *G. dolichocarpa*, and *G. tomentella*, and *G. max* and *G. soja*, respectively. When the number of chloroplast genes containing cpSSRs was analyzed, it was found that *G. cyrtoloba* and G. *falcate* had the highest gene number (41) and frequency (32.28%), respectively (Table 3). Also, the maximum number of cpSSRs was found in *ycf1* gene (7-9 times) in all plastomes of *Glycine* species. Furthermore, four repeats were present in *rpoC1* and *ndhA* genes of *G. dolichocarpa* and *G. stenophita* species and three repeats were in *rpl16, ndhA, trnK, pasA, rpoC1, rpl16, clpP* and *ycf2* genes of various *Glycine* species.

Distribution and motif types of cpSSRs

Among the repeat types, the most abundant one was found to be tri- (45%), followed by mono- (36%), di- (11.8%), tetra- (5.8%), penta- (0.8%), and hexanucleotide (0.6%) (Fig. 1 and Table 2). While trimeric repeats (70.7%) were predominant in genic regions, monomeric repeats were found to be widespread in intergenic regions. Although, dimeric repeats were detected abundantly, they were not located in genic regions.

Table 2. Frequency (%) of the genic and intergenic cpSSRs for *Glycine* species based on motif size.

Species	Mono		Di		Tri		Tetra		Penta		Hexa		
	G.	I.	G.	I.	G.	I.	G.	I.	G.	I.	G.	I.	T.
Glycine canescens	8.2 (13)	28.3 (45)	0 (0)	11.9 (19)	22.7 (36)	22.7 (36)	1.2 (2)	4.4 (7)	0 (0)	0 (0)	0 (0)	0.6 (1)	159
Glycine cyrtoloba	3.2 (5)	32 (50)	0 (0)	15.2 (24)	21.7 (34)	22.9 (36)	1.2 (2)	3.2 (5)	0 (0)	0 (0)	0 (0)	0.6 (1)	157
Glycine dolichocarpa	8.5 (14)	28 (46)	0 (0)	10.5 (17)	21.9 (36)	23.3 (38)	1.2 (2)	5.4 (9)	0 (0)	0 (0)	0 (0)	1.2 (2)	164
Glycine falcata	8.1 (14)	31.7 (55)	0 (0)	9.7 (17)	20.8 (36)	21.9 (38)	1.1 (2)	4.4 (8)	0 (0)	1.8 (3)	0 (0)	0.5 (1)	174
Glycine max	9.7 (15)	27.2 (42)	0 (0)	12.9 (20)	25.8 (40)	19.4 (30)	0.6 (1)	3.2 (5)	0 (0)	1.2 (2)	0 (0)	0 (0)	155
Glycine soja	9.9 (15)	26.8 (41)	0 (0)	13.2 (20)	26.4 (40)	18.9 (29)	0.6 (1)	3 (5)	0 (0)	1.2 (2)	0 (0)	0 (0)	153
Glycine stenophita	9.9 (15)	21.6 (33)	0 (0)	10.5 (16)	22.9 (35)	24.2 (37)	1.9 (3)	7.2 (11)	0 (0)	1.2 2)	0 (0)	0.6 (1)	153
Glycine tomentella	9.5 (15)	25.9 (41)	0 (0)	11.4 (18)	22.3 (35)	23.4 (37)	1.2 (2)	5.7 (9)	0 (0)	0 (0)	0 (0)	0.6 (1)	158
Total													1273

G: genic I: intergenic, numbers within parentheses represent absolute number of microsatellites. Also, mono, di, tri, tetra, penta, and hexa show the repeat types.

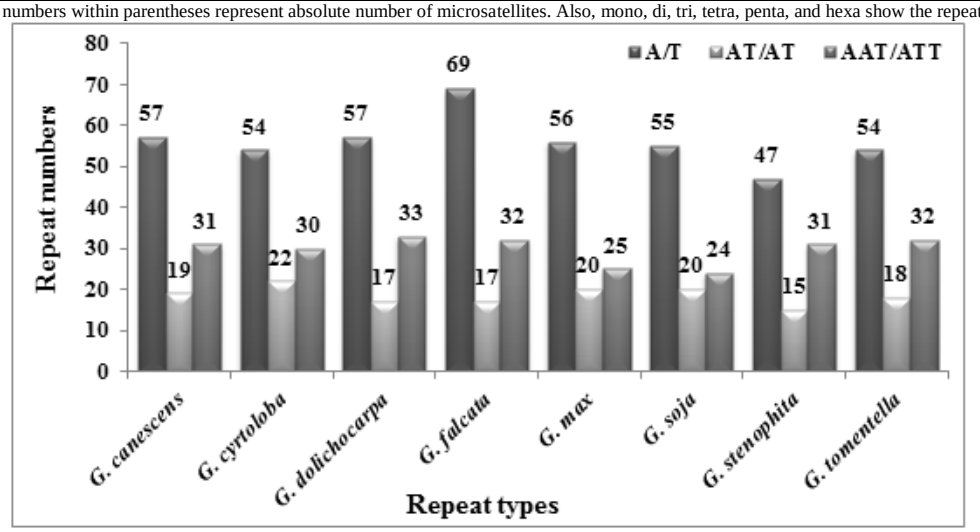

Fig 2. The most frequent motif types of all *Glycine* species for mono-, di-, and trimeric repeat types.

The motifs A/T (97.8%), AT/AT (98%) and AAT/ATT (41.5%) had the highest frequencies for mono-, di-, and and trimeric in all plastomes, respectively (Fig. 2). The highest numbers of motifs for tetra-, penta-, and hexameric repeats were: AAAT/ATTT and AAATTG/AATTTC in *G. canescens*, AAAT/ATTT and AGGGAT/ATCCCT in *G. cyrtoloba*, AAAT/ATTT, AAATTC/AATTTG and AAATTG/AATTTC in *G. dolichocarpa*, AAAT/ATTT, AAAAT/ATTTT, AAATC/ATTTG, AATAG/ATTCT and AAATTG/AATTTC in *G. falcata*, AATC/ATTG, AGAT/ATCT and AACAG/CTGTT in *G. max* and *G. soja*, AAAT/ATTT, AAATT/AATTT, AATAG/ATTCT and AGATAT/ATATCT in *G. stenophita*, and AAAT/ATTT and AAATTG/AATTTC in *G. tomentella*. Interestingly, *G. max* and *G. soja* had the same cpSSR motifs for all types. In genic regions, A/T (98.1%) and AAG/CTT (43.8%) motifs were prevalent for mono-and trimeric repeats in studied *Glycine* species, respectively.

Putative codon repeats

Trimeric repeats in genic regions were analyzed for putative amino acid codons (Table 4). A total of 853 putative codon repeats were identified and according to our results, glutamic acid (21.2%) was the predominant amino acid in triplets, followed by serine (15.5%), arginine (8.3%), and phenylalanine (7.8%) in *Glycine* species. The codon repeats were ranged from 102 (in *G. max* and *G. soja*) to 118 (in *G. tomentella*). Although, glutamic acid was the most abundant in six *Glycine* species (except for *G. max* and *G. soja*), glutamic acid and serine codons were in equal numbers (21) in *G. max* and *G. soja* species. Tryptophan, proline, and aspartic acid were absent in all plastomes of *Glycine* species. Interestingly, *G. max* and *G. soja* had the same putative codon repeats and numbers.

Discussion

Most of the genomic SSRs are located in nuclear genome and they can be classified into three types based on their locations in the genome: nuclear SSRs (nuSSRs), chloroplast SSRs (cpSSRs), and mitochondrial SSRs (mtSSRs) (Kalia et al., 2011). In this research, cpSSRs were screened in eight *Glycine* species by using bioinformatics tools and a total of 1273 cpSSR were identified with an average frequency of 1.04 cpSSR per kb. It is lower than that of 13 Poaceae species,

Table 3. Frequency (%) of chloroplast genes bearing cpSSRs.

Glycine Species	Number of Chloroplast Genes	Genes with cpSSR	Genes with cpSSR %
Glycine canescens	127	37	29.13
Glycine cyrtoloba	127	41	32.28
G. dolichocarpa	127	39	30.71
Glycine falcata	127	41	32.28
Glycine max	128	36	28.12
Glycine soja	131	36	27.48
Glycine stenophita	127	40	31.50
Glycine tomentella	127	38	29.92

Table 4. Total occurrences of codon repeats in coding DNA sequence of *Glycine* chloroplast genomes.

Codons	Encoded amino acid residue	*G. canescens*	*G. cyrtoloba*	*G. dolichocarpa*	*G. falcata*	*G. max*	*G. soja*	*G. stenophita*	*G. tomentella*	Total number of amino acid residues (%)
GGA/GGG/GGC/GGT	Glycine	6	6	6	6	6	6	6	6	48 (5.6)
GCA/GCG/GCC/GCT	Alanine	3	3	3	3	3	3	3	3	24 (2.8)
GTA/GTG/GTC/GTT	Valine	3	3	3	0	3	3	3	3	21(2.4)
CTA/CTG/CTC/CTT/TTA/TTG	Leucine	6	6	6	6	6	6	6	6	48 (5.6)
ATA/ATC/ATT	Isoleucine	3	3	3	3	3	3	3	3	24 (2.8)
TGC/TGT	Cysteine	0	3	0	0	3	3	3	0	12 (1.4)
ATG	Methionine	6	6	6	6	6	6	6	6	48 (5.6)
TAC/TAT	Tyrosine	6	6	6	6	6	6	6	6	48 (5.6)
TTC/TTT	Phenylalanine	9	6	9	9	9	9	6	9	66 (7.8)
TGG	Tryptophan	0	0	0	0	0	0	0	0	0 (0)
CCA/CCG/CCC/CCT	Proline	0	0	0	0	0	0	0	0	0 (0)
TCA/TCT/TCC/AGC/AGT/TCG	Serine	15	15	15	15	21	21	15	15	132 (15.5)
ACA/ACG/ACC/ACT	Threonine	3	3	3	3	3	3	3	3	24 (2.8)
AAC/AAT	Asparagine	3	3	3	3	3	3	3	3	24 (2.8)
CAA/CAG	Glutamine	3	6	6	6	6	6	6	6	45 (5.2)
GAC/GAT	Aspartic acid	0	0	0	0	0	0	0	0	0 (0)
GAA/GAG	Glutamic acid	27	18	24	24	21	21	21	24	180 (21.2)
AAA/AAG	Lysine	3	0	3	3	3	3	0	0	15 (1.8)
CGA/CGG/CGC/CGT/AGA/AGG	Arginine	6	10	6	6	13	13	10	6	70 (8.3)
CAC/CAT	Histidine	3	3	3	3	3	3	3	3	24 (2.8)
Total occurrences of codon repeats (853)		105	100	105	102	118	118	103	102	

(1.36 cpSSR per kb) (Melotto-Passarin et al., 2011), olive species (1.47 cpSSR per kb) (Filiz and Koc, 2012), major species of pine family (9.79 cpSSR per kb) (Filiz and Koc, 2014) and *Solanum lycopersicum* EST-SSRs (expressed sequence tags) (1.3 SSR per kb) (Gupta et al., 2010) which however, is higher than that of *Eucalyptus* EST-SSRs (0.37 SSR per kb) (Ceresini et al., 2005) and *Citrus* EST-SSRs (0.5 SSR per kb) (Palmieri et al., 2007). Trimeric repeats are seen more commonly in monocot plant species, whereas monomeric repeats are more common in dicot plant species (Lawson and Zhang, 2006). There is conflict between given data and our findings. It was found that trimeric repeats (45%) were the most common, followed by mono- (36%) and dimeric repeats (11.8%) in the present study. Similar results such as cpSSRs in olive species (Filiz and Koc, 2012) and Brasssicaceae family (Gandhi et al., 2010), and EST-SSRs in some cereal species (Varshney et al., 2002). were reported by previous studies. Trimeric repeats (70.7%) were predominant in coding regions of studied *Glycine* species. In higher eukaryotic genomes, tri- and hexanucleotides are more ample in coding regions (Metzgar et al., 2000) and this is consistent with our results. Based on motif analysis, A/T was predominant motif in monomeric repeats in *Glycine*-species. The results from previous studies such as cpSSRs in olive (Filiz and Koc, 2012), Poaceae (Melotto-Passarin et al., 2011), rice (Rajendrakumar et al. 2007) and *Eucalyptus* species (Rabello et al., 2005) were in agreement with our findings. For dimeric repeats, AT/AT was found to be ample motif in experimental *Glycine* species. Similar findings were reported in cpSSRs of rice (Rajendrakumar et al., 2007), EST-SSRs of *Citrus* (Shanker et al., 2007), SSRs of organelle genomes of major cereals (Rajendrakumaret al., 2008), cpSSRs of some Poaceae species (Melotto-Passarin et al., 2011), cpSSRs of Brassicaceae family (Gandhi et al., 2010) and cpSSRs of olive species (Filiz and Koc, 2012). AT repeats are plentiful in plant species; in contrast, AC repeats are common in animal species. Hence, this difference is important criteria for plant and animal genomes (Powell et al., 1996b) and our findings support this given information. For trimeric repeats, AAT/ATT was prominent in *Glycine* species and this data was consistent with 22 chloroplast genomes of algal species in Chlorophyta (Kuntal et al., 2010). Especially, the same cpSSR motifs existed in plastomes of *G. max* and *G. soja*. These species belong to subgenus *Soja* and it is presumed that *G. soja* is wild progenitor of *G. max* (Carter et al., 2004). Hence, it can be suggested that they have similar gene pool and the same cpSSRs motifs in their plastomes because of the phylogenetic relationship. In general, intergenic cpSSRs (67.6%) were more ample than genic cpSSRs (32.4%) in *Glycine* species. This data was in agreement with previous studies done with Asteraceae (Timme et al., 2007), Fabaceae (Saski et al.,

2008), Solanaceae (Daniell et al., 2006), Brassicaceae (Gandhi et al., 2010), Poaceae (Melotto-Passarin et al., 2011), Oleaceae (Filiz and Koc, 2012), Vitaceae (Jansen et al., 2006) and Theaceae (*Camellia*) families (Yang et al., 2013). It can be said that intergenic regions commonly include more cpSSRs in comparison with genic regions in higher plant species. From previous studies, five most frequent types of amino acid codons in the nuclear genomes of *Arabidopsis* and rice were identified. These are serine (27.5%), proline (11.9%), glycine (11.8%), glutamic acid (11.4%) and glutamine (6.2%) for *Arabidopsis*, and alanine (26.4%), glycine (22.4%), proline (13.1%), serine (10.1%) and arginine (5.8%) for rice.

Only tryptophan is not detected in *Arabidopsis*, while all amino acids are present in rice genome (Lawson and Zhang, 2006). In *Glycine* species, glutamic acid (21.2%) codons were predominant, followed by serine (15.5%) and arginine (8.3%). There is a similarity between our findings and the data from the studies done on rice and *Arabidopsis*. In these studies, it was found that glutamic acid in *Arabidopsis*, and serine and arginine in rice were observed abundantly. Furthermore, tryptophan was not detected in the plastomes of *Glycine* species. Different amino acid repeats are related with different classes of proteins. Acidic and polar amino acid repeats are connected with transcription factors and protein kinases while serine repeats are connected with membrane transporter proteins (Alba et al., 1999). Our findings may be associated with functional selections on amino acid repeats in the encoded proteins of *Glycine* plastomes. Also, previous data showed that SSR expansions in protein-coding regions can cause a gain-or-loss gene function with frameshift mutation (Li et al., 2004), suggesting that SSRs identified in this study may affect the gene structures and variations in *Glycine* species. Changes in the gene order and content in chloroplast genomes could be produced by gene duplication, gene or intron loss, transposition, inversion, and indels (Lee et al., 2007). Putative codon repeats in genic regions and cpSSRs in genes may be affected by genomic dynamics in evolution of *Glycine* chloroplast genomes. In Poaceae, cpSSRs are present in *ndh* (NADH dehydrogenase), *rps* (ribosomal proteins), *trn* (tRNA), and *rpl* (ribosomal proteins) gene clusters (Melotto-Passarin et al., 2011). Similarly, there are cpSSRs in *psaA*, *psaB*, and *ycf2* genes of *Nuphar advena* and *Ranunculus macranthus* (Raubeson et al., 2007). The maximum number of cpSSRs was found in *ycf1* gene in *Glycine* species and there is disagreement between our finding and the studies done previously.

Materials and Methods

The complete chloroplast genome sequences of eight *Glycine* species were retrieved from NCBI genome database (http://www.ncbi.nlm.nih.gov/genomes) (Table 1). The identifications of chloroplast microsatellites were done by MISA perl script (MIcroSAtellite identification tool) (http://pgrc.ipk-gatersleben.de/misa/), which is able to identify the number and distribution of perfect microsatellites as well as compound microsatellites (interrupted by a certain number of bases with a repeat size of ≥10 for mono-, 5 for di-, 3 for tri-, tetra-, penta- and hexa-nucleotide). The presence of repeats in genic and intergenic regions was determined by using the coding sequence annotation information available in the GenBank genome database (http://www.ncbi.nlm.nih.gov/genome/). In addition, cpDNA sequences were screened for G+C% content using Bioedit 7.2.5 version and coding cpDNA sequences were analyzed in predicted three-nucleotide repeats which accept encoding amino acids for putative codon repetitions.

Conclusion

In the present study, *in silico* analysis of cpSSRs in eight *Glycine* species were evaluated using MISA perl script. Also, putative codon repetitions were analyzed using tri-nucleotide repeats in coding regions. A total of 1273 cpSSRs were identified and tri-nucleotide (45%) was found to be the highest repeat type. The most frequent putative codon repeat was found as glutamic acid (21.2%) in coding regions. Among the chloroplast genes, *ycf1* was found to be contained the highest number of cpSSRs, and *G. cyrtoloba* and *G. falcata* species were included the maximum number of genes containing cpSSRs. In conclusion, our results could be a scientific basis for future cpDNA studies related with *Glycine* taxa.

References

Akkaya MS, Bhagwat AA, Cregan PB (1992) Length polymorphisms of simple sequence repeat DNA in soybean. Genetics. 132: 1131–1139.

Alba MM, Santibanez-Koref MF, Hancock JM (1999) Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. J Mol Evol. 49: 789–797.

Carter TE, Nelson R, Sneller CH, Cui Z (2004) Genetic diversity in soybean. Soybeans: Improvement, Production and Uses, eds Boerma HR, Specht JE (Am Soc Agron, Madison, WI), pp. 303–416.

Ceresini PC, Silva CLSP, Missio RF, Souza EC, Fischer CN, Guillherme IR, Gregorio I, Silva EHT, Cicarelli RMB, Silva MTA et al. (2005) Satellypus: analysis and database of microsatellites from ESTs of *Eucalyptus*. Genet Mol Biol. 28: 589–600.

Daniell H, Lee SB, Grevich J, Saski C, Queseda-Vargas T, Guda C, Tomkins J, Jansen RK (2006) Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. Theor Appl Genet. 112: 1503–1518.

Doyle JJ, Doyle JJ, Rauscher JT, Brown AHD (2004) Evolution of the perennial soybean polyploid complex (*Glycine* subgenus *Glycine*): a study of contrasts. Biol J Linn Soc. 82: 583–597.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev. 5: 435–445.

Erixon P, Oxelman B (2008) Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast *clpP1* gene. PLoS ONE 3: e1386.

Filiz E, Koc I (2012) *In silico* chloroplast SSRs mining of *Olea* species. Biodiversitas 13: 114–117.

Filiz E, Koc I (2014) Assessment of chloroplast microsatellite from pine family (Pinaceae) by using bioinformatics tools. Indian J Biotechnol. 13: 34–40.

Gandhi SG, Awasthi P, Bedi YS (2010) Analysis of SSR dynamics in chloroplast genomes of Brassicaceae family. Bioinformation 5: 16–20.

Guo J, Liu Y, Wang Y, Chen J, Li Y, Huang H, Chen J, Li Y, Huang H, Qiu L, WangY. (2012) Population structure of the wild soybean (*Glycine soja*) in China: implications from microsatellite analyses. Ann Bot. 110: 777–785.

Gupta S, Tripathi KP, Roy S, Sharma A (2010) Analysis of unigene derived microsatellite markers in family Solanaceae. Bioinformation 5: 113–121.

Hancock JM, Simon M (2005) Simple sequence repeats in proteins and their significance for network evolution. Gene. 345: 113–118.

Shuilian He, Yunsheng Wang, Sergei Volis, Dezhu Li, Tingshuang Yi (2012) Genetic diversity and population structure: implications for conservation of wild soybean (*Glycine soja* Sieb. et Zucc) based on nuclear and chloroplast microsatellite variation, Int. J Mol Sci. 13: 12608–12628.

Hymowitz T (1970) On the domestication of the soybean. Econ Bot. 23: 408–421.

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Bio. 6: 32.

Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK (2011) Microsatellite markers: an overview of the recent progress in plants. Euphytica 177: 309–334.

Kuntal H, Sharma V, Daniell H (2010) Microsatellite analysis in organelle genomes of Chlorophyta. Bioinformation. 8: 255–259.

Kuroda Y, Kaga A, Tomooka N, Vaughan DA (2006) Population genetic structure of Japanese wild soybean (*Glycine soja*) based on microsatellite variation. Mol Ecol. 15: 959–974.

Lee HL, Jansen RK, Chumley T W, Kim KJ (2007) Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. Mol Biol Evol. 24: 1161–1180.

Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 21: 991–1007.

Li YH, Li W, Zhang C, Yang L, Chang RZ, Gaut BS, Qiu LJ (2010) Genetic diversity in domesticated soybean (*Glycine max*) and its wild progenitor (*Glycine soja*) for simple sequence repeat and single-nucleotide polymorphism loci. New Phytologist. 188: 242–253.

Melotto-passarin DM, Tambarussi EV, Dressano K, De Martin VF, Carrer H (2011) Characterization of chloroplast DNA microsatellites from *Saccharum* spp. and related species. Genet Mol Res. 10: 2024–2033.

Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. The Plant J. 3: 175–182.

Navascués M, Emerson BC (2005) Chloroplast microsatellites: measures of genetic diversity and the effect of homoplasy. Mol Ecol. 14: 1333–1341.

Palmer JD (1990) Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet. 6: 115–120.

Palmieri DA, Novelli VM, Bastianel M, Cristofani-Yaly M, Astúa-Monge G, Carlos EF, Oliveira AC, Machado MA (2007) Frequency and distribution of microsatellites from ESTs of *Citrus*. Genet Mol Biol. 30: 1009–1018.

Powell W, Morgante M, McDevitt R, Vendramin GG, Rafalski JA (1995) Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. P Natl Acad Sci. 92: 7759–7763.

Powell W, Morgante M, Doyle JJ, McNicol JW, Tingey SV, Rafalski AJ (1996a) Genepool variation in genus *Glycine* subgenus *Soja* revealed by polymorphic nuclear and chloroplast microsatellites. Genetics 144: 792–803.

Powell W, Machray GC, Provan J (1996b) Polymorphism revealed by simple sequence repeats. Trends Plant Sci. 1: 215–222.

Rabello E, Nunes de Souza A, Saito D, Tsai SM (2005) *In silico* characterization of microsatellites in *Eucalyptus* spp.: Abundance, length variation and transposon associations. Genet Mol Biol. 28: 582–588.

Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM (2007) Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. Bioinformatics. 23: 1–4.

Rajendrakumar P, Biswal AK, Balachandran SM, Sundaram RM (2008) *In silico* analysis of microsatellites in organellar genomes of major cereals for understanding the phylogenetic relationships. In Silico Biology. 8: 9–18.

Saski C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG, Jansen RK (2005). Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Mol Biol. 59: 309–322.

Sakai M, Kanazawa A, Fujii A, Thseng FS, Abe J, Shimamoto Y (2003) Phylogenetic relationships of the chloroplast genomes in the genus *Glycine* inferred from four intergenic spacer sequences. Plant Syst Evol. 239: 29–54.

Schmidt EE, Davies CJ (2007). The origins of polypeptide domains. Bioessays. 29: 262–70.

Singh R. J., Hymowitz T (1985) The genomic relationships among six wild perennial species of the genus Glycine subgenus *Glycine* Willd. Theor Appl Genet. 71: 221–230.

Shanker A, Bhargava A, Bajpai R, Singh S, Srivastava S, Sharma V (2007) Bioinformatically mined simple sequence repeats in UniGene of *Citrus sinensis*. Sci Hortic. 113: 353–361.

Shimamoto Y, Abe J, Gao Z, Gai JY, Thseng FS (2000) Characterizing the cytoplasmic diversity and phyletic relationship of Chinese landraces of soybean, *Glycine max*, based on RFLPs of chloroplast and mitochondrial DNA. Genet Resour Crop Evol. 47: 611–617.

Soranzo N, Provan J, Powell W (1999) An example of microsatellite length variation in the mitochondrial genome of conifers. Genome. 42: 158–161.

Sugiura M, Hirose T, Sugita M (1998) Evolution and mechanism of translation in chloroplasts. Annu Rev Genet. 32: 437–459.

Timme R, Kuehl EJ, Boore JL, Jansen RK (2007) A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. Am J Bot. 94: 302–312.

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett. 7: 537–546.

Xu DH, Abe J, Gai JY, Shimamoto Y (2002) Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: Evidence for multiple origins of cultivated soybean. Theor Appl Genet. 105: 645–653.

Yang JB, Yang SX, Li HT, Yang J, Li DZ (2013) Comparative Chloroplast Genomes of *Camellia* Species. PLoS ONE. 8: e73053.