Indian Journal of Biotechnology
Vol 13, January 2014, pp 34-40

Assessment of chloroplast microsatellite from pine family (*Pinaceae*) by using bioinformatics tools

Ertugrul Filiz[1]* and Ibrahim Koc[2]

[1]Department of Crop and Animal Production, Cilimli Vocational School, Duzce University, Duzce, Turkey
[2]Department of Molecular Biology and Genetics, Gebze Institute of Technology, Kocaeli, Turkey

Received 10 January 2013; revised 18 August 2013; accepted 5 September 2013

Microsatellites, also known as simple sequence repeats (SSRs), are repeating sequences of 2-6 base pairs of DNA and affect chromatin organization, regulation of gene activity, DNA repair, DNA recombination, etc. Chloroplast DNA (cpDNA) has been used extensively in plant studies at different taxonomic levels. Therefore, the aim of this study was to understand the distrubution of microsatellites in the coding and non-coding regions of organellar genomes (cpDNAs) of major species of pine family (*Pinaceae*), *viz.*, *Cathaya argyrophylla*, *Cedrus deodara*, *Larix decidua*, *Picea morrisonicola*, *P. sitchensis*, *Pinus contorta*, *P. gerardiana*, *P. koraiensis*, *P. krempfii*, *P. lambertiana*, *P. monophylla*, *P. nelsonii*, *P. thunbergii*, *Pseudotsuga sinensis* var. *wilsoniana.* 1623 cpSSRs were identified in pine species with an average frequency of 9.79 cpSSR per kb, of which 584 (22.5%) were in genic regions. Mononucleotide repeats were the most abundant cpSSRs (52.4%) in these species, followed by trinucleotide SSRs (37.3%), dinucleotide (5%), tetranucleotide (3.9%), pentanucleotide (0.8%), and hexanucleotide (0.6%). As expected, trinucleotide repeats are more common in coding regions, while other repeat motifs are abundant in non-coding cpDNA. G+C content of these species have closely similar frequency, ranging from 31.67 to 38.80%. Our analyses suggest that plastome database can be used for comparative genomics in different plant species.

Keywords: cpDNA, cpSSR, *Pinaceae*, pine, plastome

Introduction

The members of pine family (*Pinaceae*) are trees or shrubs and consist of 11 genera and more than 230 species. Based on the cone, seed and leaf morphology, the 11 genera are divided into four subfamilies: i) Pinoideae (*Pinus*), ii) Piceoideae (*Picea*), iii) Laricoideae (*Larix*, *Cathaya*, *Pseudotsuga*), and iv) Abietoideae (*Abies*, *Cedrus*, *Pseudolarix*, *Keteleeria*, *Nothotsuga*, *Tsuga*). Many members of conifers, such as, cedars, firs, hemlocks, larches, pines and spruces, are well-known for their commercial importance and are used as a source of timber, pulp and resins. They also play a very significant ecological role by producing large biomass and creating habitat for many other organisms[1]. Many species often form the dominant component of boreal, coastal and montane forests in the northern hemisphere[2].

Chloroplasts are dynamic organelles and have their own genome with expression of their genetic information[3]. The chloroplast genome (cpDNA) consists of homogeneous circular double stranded DNA molecules of 110-200 kb size, containing between 30-50 different RNA genes[4]. The nonrecombinant, uniparentally inherited nature of organelle genomes makes them potentially useful tools for evolutionary studies. Thus, the chloroplast genome is widely used in plant systematic studies to deduce plant phylogenies at different taxonomic levels[5,6]. Chloroplast genomes are excellent model systems in computational genomics[7].

Comparative genomic studies indicate that chloroplast genomes of plants are highly conserved in both gene order and gene content. In land plants, about 40-50% of each chloroplast genome contains non-coding spacer and regulatory regions[8]. Most cpDNAs include two identical regions in opposite orientations called the inverted repeat (IR), flanked by large single copy (LSC) and small single copy (SSC) regions[9]. Microsatellites, also known as simple sequence repeats (SSR) or short tandem repeats (STR), are composed of small motifs of 1 to 6 nucleotides repeated in tandem, which are widespread in both eukaryotic and prokaryotic genomes[10]. SSR markers have been useful for a variety of applications in plant

*Author for correspondence:
Tel: +90-380-6817312 ext. 7413; Fax: +90-380-6817313
E-mail: ertugrulfiliz@gmail.com

genetics and breeding because of their reproducibility, multiallelic nature, codominant inheritance, relative abundance, and good genome coverage[11]. Chloroplast SSRs (cpSSRs) have important and unique characteristics in comparison to nuclear microsatellites because chloroplast is characterized by haploidy, lack of recombination and uniparental inheritance[12]. Chloroplast genetic markers are potentially more effective indicators of subpopulations and differentiation than nuclear markers[13].

The main objective of the present study was to analyze the occurrence and distribution of cpSSRs in chloroplast genomes, both coding and non-coding regions, of the members of *Pinaceae* by using bioinformatics tools.

Material and Methods

Mining SSRs from cpDNA of Pine Species

All the chloroplast genome sequences of pine species (Table 1) were downloaded in FASTA format from http://www.ncbi.nlm.nih.gov/genom/ and all sequences were used for mining cpSSR data. cpSSRs were identified using MISA perl script (http://pgrc.ipk-gatersleben.de/misa/) with the criteria of selecting mononucleotide motifs with a minimum repeat length of 8, dinucleotide motifs with a minimum repeat length of 5 and trinucleotide, tetranucleotide, and pentanucleotide motifs with a minimum repeat length of 3. SSRs were searched in whole chloroplast genome as well as separate coding and non-coding regions for each species. A putative codon repeats (trinucleotide) in cpSSR regions were analyzed to compare different *Pinaceae* species.

Table 1—Comparison of general features of chloroplast genomes in *Pinaceae* species

Plant species	Plastome size (bp)	Acc. no.*	G+C content (%)
Cathaya argyrophylla	107122	NC_014589	38.78
Cedrus deodara	119299	NC_014575	31.67
Larix decidua	122474	NC_016058	38.78
Picea morrisonicola	124168	NC_016069	38.79
P. sitchensis	120176	NC_011152	36.25
Pinus contorta	120438	NC_011153	36.80
P. gerardiana	117618	NC_011154	37.97
P. koraiensis	117190	NC_004677	38.80
P. krempfii	116989	NC_011155	37.93
P. lambertiana	117239	NC_011156	34.06
P. monophylla	116479	NC_011158	37.70
P. nelsonii	116834	NC_011159	37.47
P. thunbergii	119707	NC_001631	38.50
Pseudotsuga sinensis var. *wilsoniana*	122513	NC_016064	38.76

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

Results and Discussion

Microsatellite Frequency and Distribution of cpDNA in Pine Species

Complete cpDNA sequence data were mined for chloroplast microsatellites from a total of 14 pine species and a detailed analysis of the frequency and distribution of all mono, di, tri, tetra, penta, and hexanucleotide repeats were obtained (Table 2). In the process, we identified 1623 cpSSRs, of which 584 (36%) were localized in genic regions and the 1039 (64%) cpSSRs were localized in intergenic regions. Conversely, most abundant cpSSRs of *P. koraiensis* were in genic regions (76.57%). In the present case, an average frequency of 9.79 cpSSR per kb was recorded, which is higher than those recorded for cpDNA of 14 Poaceae species (1.36 cpSSR per kb)[14], wheat ESTs (expressed sequence tags) (1.67 SSR per kb)[15], *Ricunus communis* ESTs (1.77 SSR per kb)[16], 6 *Poaceae* species ESTs (averaged 6 SSR per kb)[17], and *Solanaceae* species (1.26 cpSSR per kb)[6]. However, this value is lower than those found in loblolly pine ESTs (42.9 SSR per kb)[18]. When unit size repeat was analyzed, the mononucleotide type was the most abundant repeat in 11 of 14 cpDNAs studied. However, trinucleotide repeats were found most common in *L. decidua* and *C. deodara* (51.5 & 46%, respectively) and surprisingly mono and trinucleotide repeats equally occured in *P. morrisonicola* (42.6%). Analysis of cpSSRs in *Pinaceae* revealed that mononucleotide SSRs were the most common SSRs (52.4%) and they are closely followed by trinucleotide SSRs (37.3%), dinucleotide (5%), tetranucleotide (3.9%), pentanucleotide (0.8%), and hexanucleotide (0.6%) (Table 2). Mononucleotide repeats individually ranged from 6 to 42.2 % in genic regions, while it ranged from 14.8 to 49.2% in intergenic regions, which suggested the presence of different mutation rate or dynamics of microsatellites in both regions in these pine species.

For mononucleotide cpSSRs, A/T was the most frequent repeat in genic (31.9%) and intergenic (68.1%) region. This finding is in agreement with a study of different eukaryotic genomes, which revealed that the (A)/(T) motif was more abundant than the (G)/(C)[5,14,19]. G+C content of the different species had closely similar frequency, ranging from 31.67 to

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

	Mononucleotide		Dinucleotide		Trinucleotide		Tetranucleotide		Pentanucleotide		Hexanucleotide		
	G	I	G	I	G	I	G	I	G	I	G	I	T
C. argyrophylla	8.1	49.2	0.8	3.2	16.1	16.9	0.8	4.8	0	0	0	0	
	(10)	(61)	(1)	(4)	(20)	(21)	(1)	(6)	(0)	(0)	(0)	(0)	124
C. deodara	10.3	29.9	1.1	3.4	27.6	18.4	1.1	5.7	0	2.3	0	0	
	(9)	(26)	(1)	(3)	(24)	(16)	(1)	(5)	(0)	(2)	(0)	(0)	87
L. decidua	6	29.3	1	6.1	30.3	21.2	0	6.1	0	0	0	0	
	(6)	(29)	(1)	(6)	(30)	(21)	(0)	(6)	(0)	(0)	(0)	(0)	99
P. morrisonicola	8.2	34.4	1.6	8.2	22.9	19.7	0	4.1	0	0	0.8	0	
	(10)	(42)	(2)	(10)	(28)	(24)	(0)	(5)	(0)	(0)	(1)	(0)	122
P. sitchensis	9.7	38.8	1.9	3.9	16.5	25.2	0	2.9	0	0.9	0	0	
	(10)	(40)	(2)	(4)	(17)	(26)	(0)	(3)	(0)	(1)	(0)	(0)	103
P.contorta	16.2	39.6	0.9	4.5	17.1	16.2	0	4.5	0	0	0.9	0	
	(18)	(44)	(1)	(5)	(19)	(18)	(0)	(5)	(0)	(0)	(1)	(0)	111
P. gerardiana	15.1	42.8	0.8	3.9	17.5	16.7	0	2.4	0	0.8	0	0	
	(19)	(54)	(1)	(5)	(22)	(21)	(0)	(3)	(0)	(1)	(0)	(0)	126
P. koraiensis	42.2	14.8	3.1	1.6	28.1	3.9	2.3	1.6	0	1.6	0.8	0	
	(54)	(19)	(4)	(2)	(36)	(5)	(3)	(2)	(0)	(2)	(1)	(0)	128
P. krempfii	16.2	41.5	0.8	3.1	16.2	16.9	0	4.6	0	0	0.8	0	
	(21)	(54)	(1)	(4)	(21)	(22)	(0)	(6)	(0)	(0)	(1)	(0)	130
P.lambertiana	18.8	37.6	0.8	3.4	18.8	14.5	0	3.4	0	1.7	0.8	0	
	(22)	(44)	(1)	(4)	(22)	(17)	(0)	(4)	(0)	(2)	(1)	(0)	117
P. monophylla	19.8	37.3	0.8	4.8	16.7	17.5	0	1.6	0	0	0	1.6	
	(25)	(47)	(1)	(6)	(21)	(22)	(0)	(2)	(0)	(0)	(0)	(2)	126
P. nelsonii	17.5	36.8	0.8	3.5	19.3	15.8	0	2.6	0	2.6	0.8	0.8	
	(20)	(42)	(1)	(4)	(22)	(18)	(0)	(3)	(0)	(3)	(1)	(0)	114
P. thunbergii	26.8	31.5	1.6	2.4	23.9	12	2.4	1.6	0	0	0.8	0	
	(34)	(40)	(2)	(3)	(28)	(14)	(3)	(2)	(0)	(0)	(1)	(0)	127
P. sinensis.	10.9	31.1	0.9	4.6	24.8	20.2	0	2.7	0	0.9	0	0	
	(13)	(37)	(1)	(5)	(27)	(22)	(0)	(3)	(0)	(1)	(0)	(0)	109
Total													1623

G, Genic; I, Intergenic; T, Total number
Numbers within parentheses represent absolute number of microsatellites

38.80%. Among the dinucleotide repeats, (AT)/(TA) motif was the most common dinucleotide repeat with a frequency of 97.6%. Common distrubition of A/T and AT/TA cpSSR may have been affected selectively from A+T content of the *Pinus* family. Trinucleotide cpSSRs were the second most abundant repeats and the most frequent motif was (AAG)/(CTT) at 31%, (AAT)/(ATT) at 23.2% and (ATC/ATG) at 19.5%. (AAG/CTT) motif is also most common triplet in *Arabidopsis thaliana*[20]. For tetranucleotide cpSSRs, the most frequent motif was (AAAG/CTTT)-(ACCT/AGGT) at 22.2%, and (AAAT/ATTTT) at 17.5%. Among pentanucleotide cpSSRs, the most frequent motif was (AAGGG)/(CCCTT) at 25% and (AACCG)/(CGGTT) at 16.7%. For hexanucleotide cpSSRs, it (AATATG/ATATTC)-(ACCATC/ATGGTG) was at 22.2%. Interestingly, there is not any pentanucleotide repeats in the genic regions for all species, while 7 out of 9 hexanucleotide repeats (77.8%) were present in the genic regions. Only 261 out of 850 mononucleotide repeats (30.7%), 20 out of 85 dinucleotide repeats (23.5%), 339 out of 604 trinucleotide repeats (56.1 %), and 8 out of 63 tetranucleotide repeats (12.7%) were present in the genic region (Fig. 1).

SSRs affect chromatin organization, regulation of DNA metabolic processes, regulation of gene activity, etc. and their mutation rates are very high compared to the rates of point mutation at coding gene loci[21]. In general, intergenic cpSSR (64%) in the family *Pinaceae* were more abundant compared to genic cpSSR (36%), probably because higher polymorphism is associated with the intron regions in contrast to exons. The present results are consistent with the earlier studies on *Asteraceae*[22], *Fabaceae*[8], *Solanaceae*[23], and *Saccharum*[14]. The frequency of

plastome genes with cpSSR reached a maximum of 28.8% in *P. nelsonii*. While reviewing the largest absolute number of genes, we found that *P. koraiensis* has the highest number of genes (74) with cpSSR, followed by *P. thunbergii* (53 genes) (Table 3).

Tri and hexanucleotide repeats are reported to be the most common in the coding regions of eukaryotes[24] and, in many species, exons contain less dinucleotide and tetranucleotide SSRs, but have many more trinucleotide and hexanucleotide SSRs in comparison to other repeats[15,25]. This hypothesis is in agreement with the present results where we found haxanucleotide as the most abundant (77.6%) repeat, followed by trinucleotide (56.1%) in the genic regions of *Pinaceae* chloroplast genomes. Besides, microsatellites were identified and classified into Class I (n ≥ 20 nt), II (n = 12 to <20 nt) and III (n = 3 to <12 nt) based on the length of repeat motif.

Accordingly, in the present study, 18 (1.2%) class I, 150 (9.2%) class II and 1455 (89.6%) class III microsatellites were identified (Table 4). Mononucleotide cpSSRs were classified as only Class I and II microsatellites in the chloroplast genomes of *Pinaceae*; while mono and dinucleotide were present as Class II microsatellites in *P. monophylla*. And all types of cpSSRs were present as Class III microsatellites in *P. koraiensis*, *P. lambertiana,* and *P. nelsonii*. Lack of very long microsatellites is evident and indicates that selection operates in maintaining microsatellites within a certain range[26].

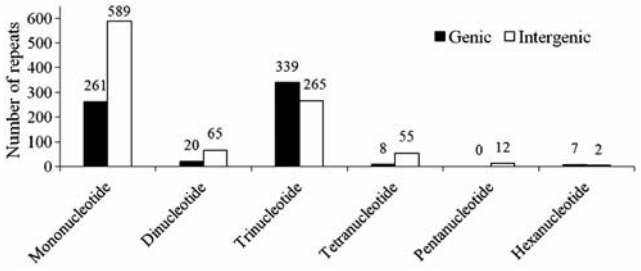

Fig. 1—Number of different repeats in genic and intergenic regions from plastomes of *Pinaceae* species.

Table 3—Frequency (%) of plastome genes with cpSSRs

Species	No. of plastome genes	Genes with cpSSR	% genes with cpSSR
C. argyrophylla	110	30	27.2
C. deodara	115	26	22.6
L. decidua	110	28	25.4
P. morrisonicola	116	31	26.7
P. sitchensis	99	26	26.3
P.contorta	110	28	25.4
P. gerardiana	110	31	28.2
P. koraiensis	315	74	23.5
P. krempfii	108	31	28.7
P. lambertiana	110	31	28.2
P. monophylla	111	31	27.9
P. nelsonii	111	32	28.8
P. thunbergii	203	53	26.1
P. sinensis.	113	30	26.5

Codon Repetitions in Coding DNA Sequences of Complete Genome

Coding DNA sequences of all the predicted peptides of *Pinaceae* chloroplast genomes were analyzed for the occurrence of the same codon (trinucleotide) (Table 5). The review of codon repeats revealed that *P. koraiensis* (99) had the most abundant codon repeats, followed by *P. morrisonicola* (85), *L. decidua* (84), and *P. thunbergii* (84), while *P. sitchensis* (51) had the less frequent codon repeats, followed by *C. argyrophylla* (52). In coding sequences, serine amino acid was found predominant (10%), followed by glutamic acid (9.6%) and methionine (9.2%). Surprisingly, no proline residue was observed in cpDNA of *Pinaceae*, while cysteine (0.3%), threonine (0.6%), and glutamine (0.6%) were the lowest frequent amino acids. *P. koraiensis* had strongly tyrosine and glutamic acid residues (12.1%); while *C. argyrophylla, C. deodara, P. morrisonicola,* and *L. deciduas* had glutamic acid repeats abundantly (19.2, 13.7, 11.8 & 11.8%, respectively). Other species had mostly serine, alanine, leucine, methionine, tryptophan, arginine, lysine, phenylalanine, tyrosine and glycine residues. The above results evidently imply that functional selection acts on amino acid repeats.

In protein-coding regions of all known proteins, 14% are proved to hold repeated sequences, with a 3 times higher abundance of repeats in eukaryotes as in prokaryotes[27]. Different amino acid repeats are inferred in different classes of proteins[28]. In plants, the most common codon repeats are lysine and arginine for *Arabidopsis* and sugarcane, respectively[29]. This is consistent with our data that some *Pinus* species have either lysine or arginine repeats abundantly (Table 5.). Furthermore, it was found that serine amino acid (10%) is the most abundant followed by glutamic acid (9.6%) and methionine (9.2%) in all *Pinus* species.

Microsatellites are not regularly distributed within coding and non-coding sequences and the DNA repair system affects in determining microsatellite distribution in different species[30,31]. The high mutation rates of microsatellites are explained by several mechanisms, such as, errors during recombination,

Table 4—Classification of microsatellites in chloroplast genomes of *Pinaceae* species

SSR category	Class I (n ≥ 20 nt)	Class II (n = 12 to <20 nt)	Class III (n = 3 to <12 nt)	SSR category	Class I (n ≥ 20 nt)	Class II (n = 12 to <20 nt)	Class III (n = 3 to <12 nt)
Cathaya argyrophylla				*P. koraiensis*			
Mononucleotide	4	18	49	Mononucleotide	1	20	52
Dinucleotide	-	-	5	Dinucleotide	-	-	6
Trinucleotide	-	-	41	Trinucleotide	-	-	41
Tetranucleotide	-	-	7	Tetranucleotide	-	-	5
Pentanucleotide	-	-	-	Pentanucleotide	-	-	2
Hexanucleotide	-	-	-	Hexanucleotide	-	-	1
Cedrus deodara				*P. krempfii*			
Mononucleotide	1	5	29	Mononucleotide	1	11	63
Dinucleotide	-	-	4	Dinucleotide	-	-	5
Trinucleotide	-	-	40	Trinucleotide	-	-	43
Tetranucleotide	-	-	6	Tetranucleotide	-	-	6
Pentanucleotide	-	-	2	Pentanucleotide	-	-	-
Hexanucleotide	-	-	-	Hexanucleotide	-	-	1
Larix decidua				*P. lambertiana*			
Mononucleotide	-	6	29	Mononucleotide	1	19	46
Dinucleotide	-	-	7	Dinucleotide	-	-	5
Trinucleotide	-	-	51	Trinucleotide	-	-	39
Tetranucleotide	-	-	6	Tetranucleotide	-	-	4
Pentanucleotide	-	-	-	Pentanucleotide	-	-	2
Hexanucleotide	-	-	-	Hexanucleotide	-	-	1
Picea morrisonicola				*P. monophylla*			
Mononucleotide	-	14	38	Mononucleotide	2	12	58
Dinucleotide	-	-	12	Dinucleotide	-	1	6
Trinucleotide	-	-	52	Trinucleotide	-	-	43
Tetranucleotide	-	-	5	Tetranucleotide	-	-	2
Pentanucleotide	-	-	-	Pentanucleotide	-	-	-
Hexanucleotide	-	-	1	Hexanucleotide	-	-	2
P. sitchensis				*P. nelsonii*			
Mononucleotide	3	5	42	Mononucleotide	1	12	49
Dinucleotide	-	-	6	Dinucleotide	-	-	5
Trinucleotide	-	-	43	Trinucleotide	-	-	40
Tetranucleotide	-	-	3	Tetranucleotide	-	-	3
Pentanucleotide	-	-	1	Pentanucleotide	-	-	3
Hexanucleotide	-	-	-	Hexanucleotide	-	-	1
Pinus contorta				*P. thunbergii*			
Mononucleotide	2	5	55	Mononucleotide	-	7	67
Dinucleotide	-	-	6	Dinucleotide	-	-	5
Trinucleotide	-	-	37	Trinucleotide	-	-	42
Tetranucleotide	-	-	5	Tetranucleotide	-	-	5
Pentanucleotide	-	-	-	Pentanucleotide	-	-	-
Hexanucleotide	-	-	1	Hexanucleotide	-	-	1
P. gerardiana				*Pseudotsuga sinensis*			
Mononucleotide	2	13	58	Mononucleotide	-	2	48
Dinucleotide	-	-	6	Dinucleotide	-	-	6
Trinucleotide	-	-	43	Trinucleotide	-	-	49
Tetranucleotide	-	-	3	Tetranucleotide	-	-	3
Pentanucleotide	-	-	1	Pentanucleotide	-	-	1
Hexanucleotide	-	-	-	Hexanucleotide	-	-	-

Table 5—Total occurrences of codon repeats in coding DNA sequences of chloroplast genomes in *Pinaceae* species

Codon	Encoded amino acid residue	Codon repeats in genic regions														Total no. of amino acid residues (%)
		C. argyrophylla	*C. deodara*	*P.morrisonicola*	*L.decidua*	*P. sitchensis*	*P. contorta*	*P. gerardiana*	*P. koraiensis*	*P. krempfii*	*P. lambertiana*	*P. monophylla*	*P. nelsonii*	*P .thumbergii*	*P. sinensis*	
		3	3	3	6	3	0	0	0	0	0	0	0	0	3	21 (2.1)
GCA/GCG/GCC/GCT	Alanine	3	3	9	0	6	6	3	3	3	3	3	3	6	6	57 (5.8)
GTA/GTG/GTC/GTT	Valine	0	6	0	3	0	0	3	0	0	0	0	0	0	3	15 (1.5)
CTA/CTG/CTC/CTT/TTA/TTG	Leucine	3	6	6	6	0	6	3	6	3	3	0	0	9	6	57 (5.8)
ATA/ATC/ATT	Isoleucine	3	3	3	0	3	3	3	3	3	3	6	6	6	3	48 (4.9)
TGC/TGT	Cysteine	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3 (0.3)
ATG	Methionine	6	6	6	9	6	6	6	6	6	6	6	6	6	9	90 (9.1)
TAC/TAT	Tyrosine	0	3	3	3	3	3	9	12	6	6	3	3	6	3	63(6.4)
TTC/TTT	Phenylalanine	0	3	3	6	3	0	6	6	6	6	6	9	9	6	69 (7)
TGG	Tryptophan	0	6	6	6	6	6	6	6	6	6	3	6	6	3	72 (7.3)
CCA/CCG/CCC/CCT	Proline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 (0)
TCA/TCT/TCC/AGC/AGT/TCG	Serine	6	9	9	9	6	6	6	6	6	6	9	9	6	6	99 (10)
ACA/ACG/ACC/ACT	Threonine	0	0	0	0	0	0	0	3	0	3	0	0	0	0	6 (0.6)
AAC/AAT	Asparagine	6	3	3	3	3	3	3	3	3	3	3	3	3	6	48 (4.9)
CAA/CAG	Glutamine	0	0	0	0	0	0	0	6	0	0	0	0	0	0	6 (0.6)
GAC/GAT	Aspartic acid	3	6	6	6	3	0	6	6	6	6	3	3	3	6	63 (6.4)
GAA/GAG	Glutamic acid	10	10	10	10	3	3	3	12	3	3	9	6	6	6	94 (9.5)
AAA/AAG	Lysine	0	3	9	6	3	3	3	9	6	6	6	6	9	6	75 (7.7)
CGA/CGG/CGC/CGT/AGA/AGG	Arginine	3	0	3	9	3	6	3	9	3	3	3	3	6	3	57 (5.8)
CAC/CAT	Histidine	6	3	3	3	0	3	3	3	3	3	3	3	3	3	42 (4.3)
Total occurrences of codon repeats (985)		52	73	85	85	51	54	66	99	63	66	63	66	84	78	

unequal crossing-over, and slippage during DNA replication or repair[32]. Owing to high microsatellite mutation rate, it is expected that coding regions have a low microsatellite density because they have important genetic information related to cell life[31]. In the present study, the comparison of *Pinaceae* species showed that genic regions have less cpSSRs, which is consistent with the existing hypothesis. SSR triplets are predominant over other repeats in coding regions because nontrimeric cpSSRs in coding region are suppressed by frameshift mutations[25]. cpDNA analysis of *Pinaceae* species revealed that trinucleotide repeats were second largest SSR motif and these repeats could affect *Pinaceae* plastome organizations, including high mutation rates, because most of the triplet repeats were located in genic region (56.1%), in contrast to intergenic regions of the chloroplast genomes. Thus, our results corroborate the above hypothesis[25].

In conclusion, we have demonstrated the distribution and organization of cpSSRs in the chloroplast genomes of some pine species. The information obtained from the present study has contributed to understand phylogenetic relationships between the *Pinaceae* species and could also become scientific basis for further studies on phylogeny, population genetics and evolutionary biology in *Pinaceae* family.

References

1 Krutovsky K V, Troggio M, Brown G R, Jermstad K D & Neale D B, Comparative mapping in the *Pinaceae*, *Genetics*, 168 (2004) 447-461.

2 Liston A, Gernandt D S, Vining T F, Campbell C S & Pinero D, Molecular phylogeny of *Pinaceae* and *Pinus*, *Acta Hortic*, 615 (2003) 107-114.

3 Bausher M G, Singh N D, Lee S B, Jansen R K & Daniell H, The complete chloroplast genome sequence of *Citrus*

sinensis (L.) Osbeck var. 'Ridge Pineapple': Organization and phylogenetic relationships to other angiosperms, *BMC Plant Biol*, 6 (2006) 21.

4 Sugiura M, The chloroplast genome, *Plant Mol Biol*, 19 (1992) 149-168.

5 Rajendrakumar P, Biswal K A, Balachandran S M, Srinivasarao K & Sundaram R M, Simple sequence repeats in organellar genomes of rice: Frequency and distribution in genic and intergenic regions, *Bioinformatics*, 23 (2007) 1-4.

6 Tambarussi E V, Melotto-passarin D M, Gonzalez S G, Brigati J B, de Jesus F A *et al*, *In silico* analysis of simple sequence repeats from chloroplast genomes of *Solanaceae* species, *Crop Breed Appl Biotechnol*, 9 (2009) 344-352.

7 De Las Rivas J, Lozano J J & Ortiz A R, Comparative analysis of chloroplast genomes: Functional annotation, genome-based phylogeny, and deduced evolutionary patterns, *Genome Res*, 12 (2002) 567-583.

8 Saski C, Lee S B, Daniell H, Wood T C, Tomkins J *et al*, Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes, *Plant Mol Biol*, 59 (2005) 309-322.

9 Cui L, Leebens-Mack J, Wang L, Tang J, Rymarquis L *et al*, Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach, *BMC Evol Biol*, 6 (2006) 13.

10 Field D & Wills C, Long polymorphic microsatellites in simple organisms, *Proc R Soc Lond (B) Biol Sci*, 263 (1998) 209-215.

11 Powell W, Polymorphism revealed by simple sequence repeats, *Trends Plant Sci*, 1 (1996) 215-222.

12 Birky C W, Uniparental inheritance of mitochondrial and chloroplast genes: Mechanisms and evolution, *Proc Natl Acad Sci USA*, 92 (1995) 11331-11338.

13 Petit R J, Duminil J, Fineschi S, Hampe A, Salvini D *et al*, Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations, *Mol Ecol*, 14 (2005) 689-701.

14 Melotto-Passarin D M, Tambarussi E V, Dressano K, De Martin V F & Carrer H, Characterization of chloroplast DNA microsatellites from *Saccharum* spp. and related species, *Genet Mol Res*, 10 (2011) 2024-2033.

15 Morgante M, Hanafey M & Powell W, Microsatellite are preferentially associated with conrepetitive DNA in plant genomes, *Nat Genet*, 30 (2002) 194-200.

16 Qiu L, Yang C, Tian B, Yang J B & Liu A, Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.), *BMC Plant Biol*, 10 (2010) 278.

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

18 Bérubé Y, Zhuang J, Rungis D, Ralph S, Bohlmann J *et al*, Characterization of EST-SSRs in loblolly pine and spruce, *Tree Genet Genomes*, 3 (2007) 251-259.

19 Tóth G, Gaspari Z & Jurka J, Microsatellites in different eukaryotic genomes: Survey and analysis, *Genome Res*, 10 (2000) 967-981.

20 Depeiges A, Goubely C, Lenoir A, Cocherel S, Picard G *et al*, Identification of the most represented repeated motifs in *Arabidopsis thaliana* microsatellite loci, *Theor Appl Genet*, 91 (1995) 160-168.

21 Li Y C, Korol A B, Fahima T, Beiles A & Nevo E, Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review, *Mol Ecol*, 11 (2002) 2453-2465.

22 Timme R, Kuehl E J, Boore J L & Jansen R K, A comparative analysis of the *Lactuca* and *Helianthus* (*Asteraceae*) plastid genomes: Identification of divergent regions and categorization of shared repeats, *Am J Bot*, 94 (2007), 302-312.

23 Daniell H, Lee S B, Grevich J, Saski C, Quesada-Vargas T *et al*, Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other *Solanaceae* genomes, *Theor Appl Genet*, 112 (2006) 1503-1518.

24 Goulding S E, Olmstead R G, Morden C W & Wolfe K H, Ebb and flow of the chloroplast inverted repeat, *Mol Gen Genet*, 252 (1996) 195-206.

25 Metzgar D, Bytof J & Wills C, Selection against frameshift mutations limits microsatellite expansion in coding DNA, *Genome Res*, 10 (2000) 72-80.

26 Nauta M J & Weissing F J, Constraints on allele size at microsatellite loci: Implications for genetic differentiation, *Genetics*, 143 (1996) 1021-1032.

27 Marcotte E M, Pellegrini M, Yeates T O & Eisenberg D A, A census of protein repeats, *J Mol Biol*, 293 (1999) 151-160.

28 Alba M M, Santibanez-Koref M F & Hancock J M, 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 (1999) 789-797.

29 Li Y C, Korol A B, Fahima T & Nevo E, Microsatellites within genes: Structure, function and evolution, *Mol Biol Evol*, 21 (2004) 991-1007.

30 Wilder J & Hollocher H, Mobile elements and the genesis of microsatellites in dipterans, *Mol Biol Evol*, 18 (2001) 384-392.

31 Oliveira E J, Pádua J G, Zucchi M I, Vencovsky R & Vieira M L C, Origin, evolution and genome distribution of microsatellites, *Genet Mol Biol*, 29 (2006) 294-307.

32 Strand M, Prolla T A, Liskay R M & Petes T D, Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair, *Nature (Lond)*, 365 (1993) 274-276.