Crop Breeding and Applied Biotechnology 9: 344-352, 2009
Brazilian Society of Plant Breeding. Printed in Brazil

In silico analysis of Simple Sequence Repeats from chloroplast genomes of Solanaceae species

Evandro Vagner Tambarussi[1], Danila Montewka Melotto-Passarin[1], Simone Guidetti Gonzalez[1], Joice Bissoloti Brigati[1], Frederico Almeida de Jesus[1], André Luiz Barbosa[1], Keini Dressano[1], and Helaine Carrer[1*]

Received 10 June 2009

Accepted 25 September 2009

ABSTRACT - *The availability of chloroplast genome (cpDNA) sequences of* Atropa belladonna, Nicotiana sylvestris, N. tabacum, N. tomentosiformis, Solanum bulbocastanum, S. lycopersicum *and* S. tuberosum, *which are Solanaceae species, allowed us to analyze the organization of cpSSRs in their genic and intergenic regions. In general, the number of cpSSRs in cpDNA ranged from 161 in* S. tuberosum *to 226 in* N. tabacum, *and the number of intergenic cpSSRs was higher than genic cpSSRs. The mononucleotide repeats were the most frequent in studied species, but we also identified di-, tri-, tetra-, penta- and hexanucleotide repeats. Multiple alignments of all cpSSRs sequences from Solanaceae species made the identification of nucleotide variability possible and the phylogeny was estimated by maximum parsimony. Our study showed that the plastome database can be exploited for phylogenetic analysis and biotechnological approaches.*

Key words: cpSSR, plastome, nucleotide polymorphism, phylogenetic, molecular marker.

INTRODUCTION

Solanaceae is one of the most diverse plant families, which includes many important species such as tomato, tobacco, pepper, potato, eggplant and petunia. According to Bindler et al. (2006) the members of Solanaceae family have a variable genome size, ranging from less than 950 Mb (tomato) to approximately 3,000 Mb (pepper). Besides that, these members also differ from each other in habit, habitat and morphology (Knapp 2001, 2002). Estimates of species diversity in the family range from 9,000-10, 000 species, with about 2,000 of those being species of a large cosmopolitan genus Solanum (Knapp 2002).

Chloroplast DNA (cpDNA) has been widely used to infer plant phylogenies at different taxonomic levels. It is possible because of the maternal inheritance of the chloroplast genomes (plastomes), whereas nuclear genomes have biparental inheritance and polyploidy that may cause difficulties for phylogenetic analyses (Nishikawa et al. 2002). The plastome is an important source of genetic markers for phylogenetic analysis; population-level studies, genotyping and mapping that can be used for genomic characterization and inter-specific comparison (Raubeson et al. 2007).

Studies of chloroplast sequences from Solanaceae species have shown that those sequences are highly conservative among related species (Melotto-Passarin et al. 2008).

Many studies have identified chloroplast microsatellites in complete sequenced chloroplast genomes of different plants (Nishikawa et al. 2005). Microsatellites or Simple Sequence Repeats (SSRs) are ubiquitous, hypervariable, abundant and well distributed throughout the genomes of many eukaryotic

[1] Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Departamento de Ciências Biológicas, Av. Pádua Dias 11, C.P. 9, 13418-900, Piracicaba, SP, Brazil. *E-mail: hecarrer@esalq.usp.br

and prokaryotic species (Nishikawa et al. 2005, Rajendrakumar et al. 2007). The SSRs can be considered as non-coding repetitive DNA regions composed of small motifs of one to six nucleotides repeated in tandem (Oliveira et al. 2006).

Studies with SSR markers can also provide important information to identify conservation in motifs, investigate the genetic processes that occur in a plant population (Heywood and Iriondo 2003) and infer evolution and phylogenetic processes. Based thereon, it was hypothesized that the SSR of plastome of Solanaceae species could be used as genetic marker and not only an occurrence during plastid evolution within the family. Therefore the SSRs sequences of cpDNA from *Atropa belladonna*, *Nicotiana sylvestris*, *N. tabacum*, *N. tomentosiformis*, *Solanum bulbocastanum*, *S. lycopersicum* and *S. tuberosum*, which are Solanaceae species, were analyzed with the following objectives: 1) to assess the organization of SSRs in their genic and intergenic regions at Solanaceae species plastomes; 2) to analyze the utility and potential of SSRs to be integrated in Solanaceae phylogenetic analysis.

MATERIAL AND METHODS

Search for SSRs in chloroplast genome of Solanaceae species

The chloroplast DNA sequences of Solanaceae species from GenBank were used for the purpose of generating SSR data. These sequences included the chloroplast genomes of *Atropa belladonna* (GenBank Accession # AJ316582; Schmitz-Linneweber et al. 2002), *Nicotiana sylvestris* (GenBank Accession # AB237912; Yukawa et al. 2006), *N. tabacum* (GenBank Accession # Z00044; Shinozaki et al. 1986), *N. tomentosiformis* (GenBank Accession # AB240139; Yukawa et al. 2006), *Solanum bulbocastanum* (GenBank Accession # DQ347958; Daniell et al. 2006), *S. lycopersicum* (GenBank Accession # AM087200) and *S. tuberosum* (GenBank Accession # DQ231562; Chung et al. 2006). SSRs were identified and localized using the software FastPCR (Kalendar et al. 2009), which identifies mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats.

We have considered only those repeats, wherein the motifs were repeated as follows: mononucleotide repeats with a repeat length ≥8 nt; dinucleotide with a repeat length ≥5 nt; tri-, tetra-, penta-, hexanucleotide with a repeat length ≥3 nt for further analysis. Also, we have considered ≤9 nt interrupting the interrupted microsatellites type. The rationale for choosing a low cut-off value is that SSRs are often disrupted by single base substitution (Subramanian et al. 2003). The occurrence of repeats in genic and intergenic regions was identified based on the sequence annotation information available in GenBank database.

Phylogenetic reconstruction and sequence analysis

SSR sequences of plastomes of each Solanaceae species were concatenated and all of them were aligned using the multiple alignment algorithm CLUSTAL W (Thompson et al. 1994), with subsequent manual correction following the guidelines of Kelchner (2000). The aligned matrix was imported into MEGA4 software version 4.0 (Molecular Evolutionary Genetics Analysis) for phylogenetic analysis (Tamura et al. 2007). Maximum parsimony tree was obtained from the resulting matrix using heuristic search options. Searches with 1,000 replicates of random addition sequences (saving no more than 30 trees per replicate to reduce time spent swapping large islands of trees) with the tree bisection reconnection (TBR) branchswapping algorithm and MulTrees on (keeping the multiple equally most-parsimonious trees).

Internal support was assessed using 1,000 bootstrap replicates (Felsenstein 1985). Groups with bootstrap percentages of 90-100 were considered to be strongly supported, 80-89 moderately supported and 50-79 weakly supported. Only groups with bootstrap > 50 that are consistent with the strict consensus tree are shown.

A distinct phylogenetic analysis was estimated by maximum parsimony with the following species: *A. belladonna*, *N. sylvestris*, *N. tabacum*, *N. tomentosiformis*, *S. bulbocastanum*, *S. lycopersicum* and *S. tuberosum*, based on the unrooted tree method.

RESULTS AND DISCUSSION

Frequency and distribution of SSRs in genic and intergenic regions

Seven different plastomes from Solanaceae family, representing 1,245,917 bp, were mined for microsatellites. SSRs were analyzed for type, abundance, and distribution (genic and intergenic regions). A total of 1,567 SSRs were identified, with an average frequency of 1.26 SSR per kb, which is a low

frequency compared to the rice chloroplasts (6.5 SSR per kb) (Rajendrakumar et al. 2007) and is higher when compared with *Eucalyptus* ESTs (0.37 SSR per kb) (Ceresini et al. 2005) and *Citrus* ESTs (0.5 SSR per kb) (Palmieri et al. 2007). This was expected since the occurrence of particular microsatellite motifs and repeats (especially the non-trimeric ones) could have implications on how the gene coding region is transcribed, due to risks of frameshift mutations (Metzgar et al. 2000).

When unit size repeat were analyzed, the mononucleotide type was the most abundant repeat in all eight plastomes studied, a result that agrees with Rajendrakumar et al. (2007) in a study on rice chloroplasts, and differs from Cardle et al. (2000) in a study on *Arabidopsis*, who found dinucleotides as the most common, and from that of Varshney et al. (2002), who found trinucleotides as the most frequent in ESTs of some cereals, followed by dinucleotides. The results suggest that in organellar genomes like chloroplast the mononucleotides SSRs are the most abundant.

According to the type of repeat sequence, microsatellites are classified as: perfect (sequence with uninterrupted motifs), imperfect (repeat sequence with interrupted motifs), interrupted (small sequence within the repeated sequence that does not match the motif sequence) or composite (sequence contains two adjacent distinctive sequence-repeats). All these were found in chloroplast genomes of the Solanaceae species studied. Researches with microsatellite markers can also provide important information to identify conservation units, to investigate the genetic processes that occur in a population (Heywood and Iriondo 2003) and to understand evolution and phylogenetic processes.

The mononucleotide SSR type reaches the maximum of 72.56% in *N. tabacum*. Similar results were obtained in rice chloroplasts (Rajendrakumar et al. 2007). Among the other repeats, di-, tri- and tetranucleotides showed variations at frequencies for all species, ranging from 4.0% in *N. tabacum* to 8.2% in *N. tomentosiformis*, 12.5% in *N. sylvestris* to 18.5% in *S. bulbocastanum*, and 4.9% in *N. tabacum* to 11.8% in *S. bulbocastanum*, respectively. Penta- and hexanucleotide repeats were represented in proportions of 0.6 to 4.7% and 0.0 to 1.1%, respectively (Table 1).

In general the number of intergenic SSR in the Solanaceae family was more abundant than genic SSR (Table 2 and 3). It also occurs in rice chloroplast (Rajendrakumar et al. 2007), probably due to an associated lower polymorphism of coding regions in contrast to non-coding ones. In general, the adenine/timine-rich repeat motifs are most common in SSRs and were also the most abundant in the Solanaceae family analyzed.

Among the mononucleotide repeats, the A/T motif was found to be more abundant than C/G in exons in all taxa studied by Tóth et al. (2000), which is in agreement with our data, in which A/T motif was found to be more abundant than C/G either in genic (98 – 100%) and intergenic (95.4 – 100%) chloroplast regions and also agrees with rice chloroplasts (Rajendrakumar et al. 2007).

The AT/AT motif was the most common dinucleotide repeat in the different chloroplast genomes (75 – 84.6%). This motif was the most frequent in intergenic regions and in *Nicotiana* species genic regions, in agreement with rice (Rajendrakumar et al. 2007) and chloroplasts from other species (Powell et al. 1996). However dinucleotide repeats differ in the genic regions from the others species analyzed. In *A. belladonna* the GA/TC motif is frequent in genic regions, which also happen with *S. tuberosum* and *S.*

Table 1. Frequency (%) of chloroplast SSRs based on motif size for each species. Numbers within parentheses represent absolute number of microsatellites

	AB	NT	NS	NTO	SB	SL	ST
Mononucleotide	68.2 (150)	72.6 (164)	67.7 (130)	65.6 (128)	58.5 (114)	54.7 (110)	66.5 (107)
Dinucleotide	5.0 (11)	4.0 (9)	5.7 (11)	8.2 (16)	6.7 (13)	4.0 (8)	6.2 (10)
Trinucleotide	15.9 (35)	15.0 (34)	12.5 (24)	12.8 (25)	18.5 (36)	19.9 (40)	17.4 (28)
Tetranucleotide	8.2 (18)	4.9 (11)	8.3 (16)	10.3 (20)	11.8 (23)	15.4 (31)	9.3 (15)
Pentanucleotide	2.7 (6)	2.6 (6)	4.7 (9)	2.6 (5)	4 (8)	5 (10)	0.6 (1)
Hexanucleotide	0 (0)	0.9 (2)	1.1 (2)	0.5 (1)	0.5 (1)	1 (2)	0 (0)

*AB - *Atropa belladonna*; NT - *Nicotiana tabacum*; NS - *Nicotiana sylvestris*; NTO - *Nicotiana tomentosiformis*; SB - *Solanum bulbocastanum*; SL - *Solanum lycopersicum*; ST - *Solanum tuberosum*

bulbocastanum and *S. lycopersicum* that do not present dinucleotide repeats in the genic region. Among the dinucleotide repeats, the CG/CG repeat was extremely rare in genic and intergenic regions of the organellar genomes (Rajendrakumar et al. 2007) and in the Solanaceae family studied this motif does not occur.

For trinucleotides, the ATT/AAT and GAA/TTC motifs were the most common, and reached 51.4% in *A. belladonna* and 36% in *N. tomentosiformis*, respectively. Among intergenic regions, the trinucleotide ATT/AAT motif was the most abundant (36.4 – 70% of intergenic trinucleotide) and in the genic regions the most frequent was GAA/TTC (33.3 – 60% of genic trinucleotides). The most common tetranucleotide repeats were TTTA/TAAA (13.3 – 44.4%), GAAA/TTTC (3.1 – 25%) and AGAA/TTCT (0 – 20%). All chloroplast genomes studied present these three different tetranucleotide SSR motifs, except *N. tabacum* and *N. sylvestris* that do not present the third one. Among the intergenic tetranucleotides SSR motifs, the most frequent was the TTTA/TAAA repeat (16.7 – 50%). With respect to the genic regions there is no consensus among the species, while in *A. belladonna*, *S. lycopersicum* and *Nicotiana* species in general the most abundant motif was GAAA/TTTC (20 – 100%), in the *Solanum* species in general the AGAA/TTCT motif was the most common (20 – 25%). Excluding *A. belladonna* and *S. tuberosum* that do not present hexanucleotide SSR motifs, the remaining penta- and hexanucleotide repeats were represented by different motifs but they are rich in A/T nucleotides and were much more abundant in intergenic regions (Table 3).

Despite SSR motif types, abundance and mutation rates are different among species, with a wide range of genetic properties (Cruz et al. 2005), they seem to be similar in chloroplast genomes within the same genera in the Solanaceae family.

Phylogenetic Analysis

The phylogeny of some Solanaceae species was estimated based on the multialignment of concatenated SSRs.

The analysis includes 7 species of Solanaceae. Of the 8,517 characters from aligned SSR matrix, 6,415 are constant. Among the 2,102 variable characters, 1,122 are parsimony-uninformative and 980 are parsimony informative. The analysis of this data set identified the most parsimonious tree (Figure 1). The ingroup is divided into two main clades, one containing *S. tuberosum*, S. *bulbocastanum* and *S. lycopersicum*; another one composed by *Nicotiana* species. The *A. belladonna* was not grouped in a specific clade and this species has the most divergence in the SSR multialignment sequence, exhibiting transversal, transition and indel mutations. Although, when tobacco is compared with *A. belladonna,* indel mutations rarely are situaded in conserved domains. Moreover, all the 113 genes identified in *A. belladonna* chromosome plastid are found in tobacco plastids, arranged in identical order (Schmitz-Linneweber et al. 2002). Five clades had robust bootstrap support (95-100% bootstrap support; Figure 1).

Borisjuk et al. (1994) pointed out that the New World *Solanum* species is more related to S. lycopersicum than to other *Solanum* species. Nucleotide sequence analysis of a phylogenetically informative part of the 3' end of 25S rDNA confirmed a closer relationship between tomato and potato than between tomato and tobacco, previously detected by Southern hybridization with an intergenic spacer element fragment of *Solanum tuberosum* as hybridization probe (Borisjuk et al. 1994).

Examples abound of the importance of a phylogenetic framework for diverse areas of plant research (for review, see Daly et al. 2001). One obvious example is the value of placing model organisms in the appropriate phylogenetic context to obtain a better

Table 2. Frequency (%) of the genic and intergenic chloroplast SSRs based on motif size for each species. Numbers within parentheses represent absolute number of microsatellites

	Mono		Di		Tri		Tetra		Penta		Hexa	
Specie	Genic	Intergenic	Genic	Intergenic	Genic	Intergenic	Genic	Intergenic	Genic	Intergenic	Genic	Intergenic
A. belladonna	29.3 (44)	70.7 (106)	18.2 (2)	81.8 (9)	20 (7)	80 (28)	11.1 (2)	88.9 (16)	16.7 (1)	83.3 (5)	0 (0)	0 (0)
N. tabacum	34.1 (56)	65.8 (108)	44.4 (4)	55.5 (5)	41.2 (14)	58.8 (20)	9.1 (1)	90.9 (10)	0 (0)	100 (6)	0 (0)	100 (2)
N. sylvestris	37.7 (49)	62.3 (81)	27.3 (3)	72.7 (8)	41.7 (10)	58.3 (14)	37.5 (6)	62.5 (10)	11.1 (1)	88.9 (8)	0 (0)	100 (2)
N. tomentosiformis	37.5 (48)	62.5 (80)	25 (4)	75 (12)	48 (12)	52 (13)	25 (5)	75 (15)	20 (1)	80 (4)	0 (0)	100 (1)
S. bulbocastanum	32.5 (37)	67.5 (77)	15.4 (2)	84.6 (11)	33.3 (12)	66.7 (24)	21.7 (5)	78.3 (18)	25 (2)	75 (6)	0 (0)	100 (1)
S. lycopersicum	34.3 (45)	65.6 (86)	0 (0)	100 (8)	41.4 (12)	58.6 (17)	25 (4)	75 (12)	25 (1)	75 (3)	0 (0)	100 (1)
S. tuberosum	39.2 (42)	60.7 (65)	20 (2)	80 (8)	64.3 (18)	35.7 (10)	26.7 (4)	73.3 (11)	0 (0)	100 (1)	0 (0)	0 (0)

Table 3. Occurrence of individual SSR motifs among the 1,567 microsatellites identified in the different chloroplasts species from Solanaceae family (G – genic, I – intergenic)

Category	Motifs	*A. belladonna*		*N. tabacum*		*N. sylvestris*		*N. tomentosiformis*		*S. bulbocastanum*		*S. lycopersicum*		*S. tuberosum*	
		G	**I**	**G**	**I**	**G**	**I**	**G**	**I**	**G**	**I**	**G**	**I**	**G**	**I**
Mono	A/T	44	102	56	104	48	78	47	78	37	75	45	82	42	65
	C/G	0	4	0	4	1	3	1	2	0	2	0	4	0	0
Di	AG/CT	0	0	0	1	0	0	1	0	0	0	0	0	1	0
	AT/TA	1	8	3	4	2	7	3	10	1	10	0	6	0	8
	GA/TC	1	1	1	0	1	1	0	2	1	1	0	2	1	0
Tri	AAG/CTT	0	0	2	0	3	0	3	0	2	2	2	2	2	1
	AAC/GTT	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	AGA/TCT	0	0	1	1	0	0	0	0	1	2	0	0	1	1
	AGT/ACT	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	ATA/TAT	0	0	0	0	0	0	0	0	1	2	0	0	0	0
	ATT/AAT	1	17	3	10	0	10	1	6	1	14	3	8	3	7
	CAA/TTG	1	1	1	1	0	0	2	0	2	0	1	0	1	0
	CAG/CTG	0	0	0	1	0	0	0	0	0	0	0	1	1	0
	CAT/ATG	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	CCT/AGG	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CTC/GAG	0	0	0	1	0	0	0	1	0	1	0	0	0	1
	GAA/TTC	4	8	5	3	6	1	6	3	5	2	5	3	8	1
	GAT/ATC	0	0	0	2	0	1	0	1	0	0	0	1	0	0
	GCA/TGC	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	GGA/TCC	0	1	0	0	0	0	0	0	0	0	0	0	1	0
	GTA/TAC	0	0	0	0	0	0	0	0	0	1	0	1	0	0
	GTT/AAC	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	TAG/CTA	0	0	1	0	0	1	0	0	0	0	0	1	0	0
	TGA/TCA	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Tetra	AAGA/TCTT	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	AAGT/ACTT	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	AGAA/TTCT	0	2	0	0	0	0	1	0	1	2	0	1	1	2
	AGAT/ATCT	0	0	0	1	0	0	0	0	0	0	1	1	0	1
	AGTA/TACT	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	ATAT/TATA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ATTA/TAAT	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CAAA/TTTG	0	1	0	1	0	1	0	1	0	2	0	2	0	1
	CTAT/ATAG	0	0	0	0	0	0	1	0	0	1	0	0	0	0
	CTTA/TAAG	0	1	0	0	0	0	0	0	0	1	0	0	0	0
	CTTC/GAAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CTTT/AAAG	0	0	0	0	0	1	1	4	1	4	1	1	1	0
	GAAA/TTTC	1	1	1	1	3	3	1	4	0	1	1	1	1	2
	GGAA/TTCC	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	GGAG/CTCC	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	GGTT/AACC	0	0	0	1	0	1	0	1	1	0	1	0	0	0
	GTTT/AAAC	0	0	0	1	0	1	0	0	0	0	0	0	0	0
	TAGA/TCTA	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	TTAA/TTAA	0	1	0	1	0	1	0	0	0	0	0	1	0	0
	TTAC/GTAA	0	1	0	0	0	0	0	0	0	0	0	1	0	0
	TATT/AATA	0	0	0	1	0	0	0	0	0	2	0	0	0	1
	TTGG/CCAA	1	0	0	0	0	2	0	0	0	0	0	0	0	1
	TTTA/TAAA	0	8	0	3	1	4	1	5	1	3	0	4	0	2
	TTTC/GAAA	1	1	1	1	3	3	1	4	0	1	1	1	1	2

To be continued …

Table 3. Cont.

Category	Motifs	*A. belladonna*		*N. tabacum*		*N. sylvestris*		*N. tomentosiformis*		*S. bulbocastanum*		*S. lycopersicum*		*S. tuberosum*	
		G	I	G	I	G	I	G	I	G	I	G	I	G	I
Penta	AATTG/CAATT	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	ATATT/AATAT	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ATTCA/TGAAT	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	TAAAA/TTTTA	0	2	0	1	0	2	0	2	0	2	0	0	0	1
	TAAAC/GTTTA	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	TATTA/TAATA	0	0	0	1	0	1	0	0	0	1	0	2	0	0
	TTATT/AATAA	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	TTCAA/TTGAA	0	0	0	1	0	1	0	0	0	0	0	0	0	0
	TTTAA/TTAAA	0	0	0	0	0	2	0	0	0	0	0	0	0	0
	TTTAT/ATAAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	TTTCT/AGAAA	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	TTTTC/GAAAA	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	CTAAA/TTTAG	0	1	0	1	0	0	0	0	0	1	0	0	0	0
	CTTAT/ATAAG	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	CTTTT/AAAAG	0	1	0	1	0	0	0	0	1	1	1	0	0	0
	GAATT/AATTC	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	GTTTT/AAAAC	0	1	0	1	1	0	1	0	1	0	0	0	0	0
Hexa	AAGAAA/TTTCTT	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	AGAAAA/TTTTCT	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	CTTATT/AATAAG	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	CTTTTT/AAAAAG	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	GAAAAA/TTTTTC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	GTAAAA/TTTTAC	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	TCTATA/TATAGA	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	TTTATT/AATAAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	TTTGAA/TTCAAA	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	Total SSR	56	164	75	151	69	123	70	125	58	137	62	127	66	95

understanding of both patterns and processes of evolution. The fact that tomato and other species of the small genus *Lycopersicon* are actually embedded within the well-marked subclade *Solanum* (and, hence, are more appropriately referred to as species of *Solanum*; tomato has been renamed *Solanum lycopersicum*; e.g., Spooner et al. 1993, Olmstead et al. 1999) is a powerful statement that is important to geneticists, molecular biologists, and plant breeders for comparative genetic/genomic research and for crop improvement.

These results are also in agreement with previous data obtained by calculation of nucleotide substitution rates in cpDNA (Kawagoe and Kikuta 1991), which demonstrated that evolutionary separation between *Solanum* and *Lycopersicon* occurred considerably later than between these two genera and *Nicotiana* (as discussed by Borisjuk et al. 1994).

The phylogenetic analysis based on the SSRs presented here shows clearly that *Solanum*, as defined traditionally, is a paraphyletic taxon. A paraphyletic group contains some, but not all, of the descendants from a common ancestor. The members included are those that have changed little from the ancestral state; those that have changed more are excluded: a paraphyletic group contains the rump of conservative descendants from an ancestral species. Thus, the absorption of *Lycopersicon* by *Solanum* seems justifiable (Melotto-Passarin et al. 2008). As was expected, our results indicated the high identity between tomato and *Solanum* species, which have the same types and amounts of SSRs (Table 3). *A. belladonna* is the most divergent species of the Solanaceae family based on the SSRs of plastome sequence multialignment.

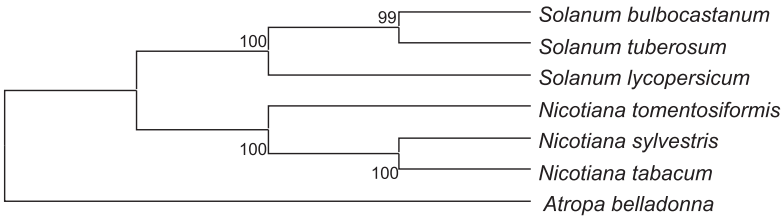

Figure 1. The maximum parsimonious unrooted tree obtained from SSRs of Solanaceae plastome species. Numbers indicate bootstrap percentages > 50%

The results presented here suggest that nucleotide sequence polymorphism of SSRs should be included in analysis with multiple chloroplast regions and can contribute to the resolution of phylogenetic studies of Solanaceae and related species. Daniell et al. (2006) sequenced the cpDNA of *S. bulbocastanum* and *S. lycopersicum* for a comparative analysis with other Solanaceae genomes. They affirmed that only four spacer regions are fully conserved (100% sequence identity) among all genomes; deletions or insertions within some intergenic spacer regions result in less than 25% sequence identity, underscoring the importance of choosing appropriate intergenic spacers for plastid transformation and providing valuable new information for phylogenetic utility of the chloroplast intergenic spacer regions. In general the number of intergenic SSR in the Solanaceae family are more abundant than genic SSR (Table 2 and 3), probably due to an associated lower polymorphism of coding regions in contrast to non-coding ones. These results show the importance of studying the potential of integration of cpDNA SSRs, which is more frequent in intergenic regions, in phylogenetic analysis.

Our results are in agreement with Daniell et al. (2006) and we assume that the SSRs represent marker position and not only isolated occurrence during plastid evolution within the family.

In conclusion, we have demonstrated a generic approach for assessing genetic variation in Solanaceae. Our study identified genes possessing SSRs in plastome (data not showed). The repeat motifs are not uniformly distributed across the Solanaceae plastomes but mostly confined to intergenic regions. A few of the genic SSRs have been found to be significantly different among Solanaceae plastomes, which may be helpful in specific PCR markers. The high sequence conservation of the cpDNAs will make it easier to design primers that may work even in relatively distant species. The SSR markers developed in the present study could also be useful in determining the maternal origin of Solanaceae species and in phylogenetic studies. SSR markers could be included in analysis with multiple chloroplast regions to improve the resolution of phylogenetic studies of the species studied.

ACKNOWLEDGMENTS

The authors would like to thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for financial support provided to Evandro Vagner Tambarussi (07/05795-0), Joice Bissoloti Brigati (06/59512-7), Frederico Almeida de Jesus (07/05797-3) and Simone Guidetti-Gonzalez (05/58117-4), CAPES for financial support provided to Danila Montewka Melotto-Passarin and CNPq for financial support provided to André Luiz Barbosa.

Análises *in silico* de Microssatélites nos genomas cloroplastidiais de Solanaceae

RESUMO - *A disponibilidae das sequencias completas dos genomas de cloroplastos (cpDNA) de* Atropa belladona, Nicotiana sylvestris, N. tabacum, N. tomentosiformis, Solanum bulbocastanum, S. lycopersicum, S. Tuberosum, *que pertecem a famlia Solanaceae, nos permite analisar a organização dos cpSSRs nas regiões gênicas e intergenicas. Em geral, o número de cpSSRs nas regiões intergênicas variam de 161 em* S. Tuberosum *a 226 em* N. Tabacum, *o número de cpSSrs foi maior do que cpSSRs das regiões gênicas. Em cpDNA, repetições de mono nucleotídeos foram os mais freqüentes nas espécies estudadas,*

mas também identificamos di, tri, tetra, penta e hexa nucleotídeos. Alinhamentos múltiplos de todas as sequencias de espécies de Solanaceaes auxiliaram a identificação da variabilidade e a filogenia foi estimada pela parsimonia maxima. Estes estudos mostram que o banco de dados de plastomas pode ser explorado para análises filogenéticas e abordagens biotecnológicas.

Palavras-chave: cpSSR, marcadores moleculares, plastomas, filogenética, polimorfismos de nucleotídeos.

REFERENCES

Bindler G, van der Hoeven R, Gunduz I, Plieske J, Ganal M, Rossi L, Gadani F and Donini P (2007) A microsatellite marker based linkage map of tobacco. **Theoretical and Applied Genetics 114**: 341-349.

Borisjuk N, Borisjuk L, Petjuch G and Hemleben V (1994) Comparison of nuclear ribosomal RNA genes among *Solanum* species and other Solanaceae. **Genome 37**: 271-279.

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D and Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. **Genetics 156**: 847-54.

Ceresini PC, Petrarolha Silva CLS, Missio RF, Souza EC, Fischer CN, Guillherme IR, Gregorio I, Silva EHT, Cicarelli RMB, Silva MTA, *et al.* (2005) Satellyptus: Analysis and database of microsatellites from ESTs of *Eucalyptus*. **Genetics and Molecular Biology 28**: 589-600.

Chung HJ, Jung JD, Park HW, Kim JH, Cha HW, Min SR, Jeong WJ and Liu JR (2006) The complete chloroplast genome sequences of *Solanum tuberosum* and comparative analysis with Solanaceae species identified the presence of a 241-bp deletion in cultivated potato chloroplast DNA sequence. **Plant Cell Report 25**: 1369-1379.

Cruz F, Pérez M and Presa P (2005) Distribution and abundance of microsatellites in the genome of bivalves. **Gene 346**: 241-247.

Daly DC, Cameron KM and Stevenson DW (2001) Plant systematics in the age of genomics. **Plant Physiology 127**: 1328-1333.

Kawagoe Y and Kikuta Y (1991) Chloroplast DNA evolution in potato (*Solanum tuberosum* L.). **Theoretical and Applied Genetics 81**: 13-20.

Daniell H, Lee SB, Grevich J, SAski C, Quesada-Vargas T, Guda C, Tomkins J and Jansen RK (2006) Complete chloroplast genome sequences of *Solanum bulbocastanum, Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. **Theoretical and Applied Genetics 112**: 1503-1518.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. **Evolution 39**: 783-791.

Heywood VH and Iriondo JM (2003) Plant conservation: old problems, new perspectives. **Biological Conservation 113**: 321-335.

Kahlau S, Aspinall S, Gray JC and Bock R (2006) Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. **Journal of Molecular Evolution 63**: 194-207.

Kalendar R, Lee D and Schulman AH (2009) FastPCR software for PCR primer and probe design and repeat search. **Genes, Genomes and Genomics 3**. Available at accessed on: www.biocenter.helsinki.fi/bi/Programs/fastpcr.htm.

Kelchner SA (2000) The evolution of noncoding chloroplast DNA and its application in plant systematics. **Annual Missouri Botanical Garden 87**: 482-498.

Knapp S (2001) Is morphology dead in Solanum taxonomy? In: van der Werden G, Barendse G and van den Berg R (eds.) **Solanaceae V**. University of Nijmegen, The Netherlands, p. 23-38.

Knapp S (2002) Floral diversity and evolution in the Solanaceae. In: Cronk QCB, Bateman RM and Hawkins JA (eds.) **Developmental genetics and plant evolution.** Taylor and Francis, London, p. 267-297.

Melotto-Passarin DM, Berger IJ, Dressano K, Morell VF, Oliveira GCX, Bock R and Carrer H (2008) Phylogenetic relationships in Solanaceae and related species based on cpDNA sequence from plastid *trnE-trnT* region. **Crop Breeding and Applied Biotechnology 8**: 85-95.

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

Nishikawa T, Salomon B, Komatsuda T, von Bothmer R and Kadowaki K (2002) Molecular phylogeny of the genus *Hordeum* using three chloroplast DNA sequences. **Genome 45**: 1157-1166.

Nishikawa T, Vaughan DA and Kadowaki K (2005) Phylogenetic analysis of *Oryza* species, based on simple sequence repeats and their flanking nucleotide sequences from the mitochondrial and chloroplast genomes. **Theoretical and Applied Genetics 110**: 696-705.

Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R and Vieira MLC (2006) Origin, evolution and genome distribution of microsatellites. **Genetic and Molecular Biology 29**: 294-307.

Olmstead RG, Sweere JA, Spangler RE, Bohs L and Palmer JD (1999) Phylogeny and provisional classification of the Solanaceae based on chloroplast DNA. In: Nee M, Symon DE, Jessup JP and Hawkes JG (eds.) **Solanaceae IV.** Advances in biology and utilization. The Royal Botanical Gardens, Kew, p. 111-137.

Palmieri DA, Novelli VM, Bastianel M, Cristofani-Yaly M, Astúa-Monge G, Carlos EF, Oliveira AC and Machado MA (2007) Frequency and distribution of microsatellites from ESTs of citrus. **Genetics and Molecular Biology 30**: 1009-1018.

Powell W, Machray GC and Provan J (1996) Polymorphism revealed by simple sequence repeats. **Trends Plant Science 1**: 215-222.

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

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade MH, Boore JL and Jansen RK (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. **BMC Genomics 8**: 1-27.

Schmitz-Linneweber C, Regel R, Du TG, Hupfer H, Herrmann RG and Maier RM (2002) The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation. **Molecular Biology and Evolution 19**: 1602-1612.

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, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H and Sugiura M (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. **EMBO Journal 5**: 2043-2049.

Spooner DS, Anderson GJ and Jansen RK (1993) Chloroplast DNA evidences for the interrelationships of tomatoes, potatoes and pepinos (Solanaceae). **American Journal of Botany 80**: 676-688.

Subramanian S, Mishra, RK and Singh L (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. **Genome Biology 4**: R13.

Tamura K, Dudley J, Nei M and Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. **Molecular Biology and Evolution 24**: 1596-1599.

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

Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. **Nucleic Acids Research 22**: 4673-4680.

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

Yukawa M, Tsudzuki T and Sugiura M (2006) The chloroplast genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: complete sequencing confirms that the *Nicotiana sylvestris* progenitor is the maternal genome donor of *Nicotiana tabacum*. **Molecular Genetics and Genomics 275**: 367-373.