Identification of microsatellites in chloroplast genome of *Anthoceros formosae*

Asheesh Shanker
Department of Bioscience and Biotechnology, Banasthali University, Banasthali-304022, Rajasthan, India, Email: ashomics@gmail.com

Abstract: Microsatellites also known as simple sequence repeats (SSRs) are short repeat motifs (1-6 bp) found in DNA sequences Detection of microsatellites is important for the development of molecular markers and to study the mapping of traits of economic, medical or ecological interest. In the present study, chloroplast genome sequence of *Anthoceros formosae*, downloaded from the National Center for Biotechnology Information (NCBI) was mined with the help of MISA tool to detect SSRs in chloroplast genome (cpSSRs). A total of 67 SSRs were detected with a density of 1 SSR/2.4 kb in 161.162 kb sequence mined. Depending on the repeat units, the length of SSRs ranged from 12 to 18 bp for mono-, 14 to 46 bp for di-, 12 to 27 bp for tri-, 12 to 20 bp for tetra and 18 bp for hexa-nucleotide repeats. Mononucleotide repeats were the most frequent repeat type (35.82%) followed by dinucleotide repeats (25.37%). Penta-nucleotide repeats were not detected in chloroplast genome sequence of *Anthoceros formosae*.

Introduction

Bryophytes are small, herbaceous, non-vascular plants and grow on rocks, soil or as epiphytes on the trunks and leaves of forest trees. Molecular phylogenies of three extant bryophyte lineages, liverworts, mosses and hornworts,based on chloroplast and mitochondrial genome sequences showed that liverworts are the earliest diverging land plant clade and hornworts are the sister group to tracheophytes (Shanker 2013; Shanker 2013a; Shanker 2013b).

Microsatellites also known as simple sequence repeats (SSRs) are sequences consist of short repeat motifs (1-6 bp) and are present in both coding and non-coding regions of DNA sequences (Katti et al. 2001; Shanker et al. 2007). SSRs have been widely used as molecular markers in many plant genomes due to their abundance and ability to associate with many phenotypes. However, the importance of microsatellites in chloroplast genomes of bryophytes has not been completely understood.

Traditional molecular methods for SSR extraction are expensive and time-consuming. However computational approaches offer rapid and economical SSR extraction for sequences available in public databases (Shanker et al. 2007a). Therefore the present study was designed to understand the organization of SSRs in chloroplast genome of *Anthoceros formosae* and to know their distribution in coding and non-coding regions.

Materials and Methods

Retrieval of chloroplast genome sequence

Chloroplast genome sequences of several plants are available at National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov). However few of them belong to bryophytes (Shanker 2012; Shanker 2012a). The chloroplast genome sequence of *Anthoceros formosae* (NC_004543, 161162 bp) was downloaded from NCBI in FASTA and GenBank format.

Mining of SSRs

MISA, a Perl script (available at http://pgrc.ipkgatersleben.de/misa/misa) was used to identify SSRs in chloroplast genome sequence of *Anthoceros formosae*. MISA takes FASTA formatted sequence file as an input and generates information of mined SSRs, if detected, along with statistical data in two separate files. The length of SSRs were defined as ≥12 bp for mono, di, tri and tetranucleotide, ≥15 bp for pentanucleotide and ≥18 bp for hexanucleotide repeats. Identified SSRs were classified as coding and non-coding SSRs based on the presence of repeats in these coding and non-coding regions, as given in GenBank file format of the chloroplast genome.

Results and Discussion

The present analysis was conducted to identify perfect SSRs with a minimum length of 12 bp in chloroplast genome sequence of *Anthoceros formosae*. The length of the identified SSRs varied from 12 to 46 bp. Pentanucleotide repeats were totally absent in chloroplast genome sequence of *Anthoceros formosae*. The majority of the detected SSRs were found in non-coding region of the genome. The frequency of identified SSRs is presented in Fig. 1.

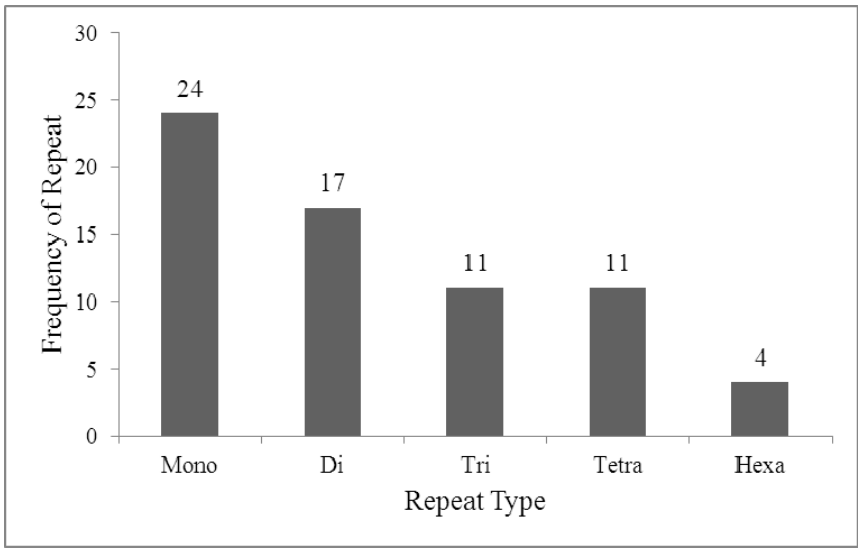

Figure 1. Frequency distribution of identified SSRs.

A total of 67 SSRs representing density of 1 SSR/2.4 kb were identified in chloroplast genome sequence of *Anthoceros formosae*. Information of mined SSRs motif, their length, start-end position and the region in which they lie is presented in Table 1. Mononucleotides were found to be most frequent repeat (24, 35.82%) followed by dinucleotide (17, 25.37%). Tri and tetranucleotide (11, 16.42%) repeats were found with equal frequencies. Hexanucleotides (4, 5.97%) were least abundant in chloroplast genome of *Anthoceros formosae*. Out of all mined SSRs 5 (7.46%) lie in coding, 62 (92.54%) lie in non-coding region of the genome.

The density of SSRs in chloroplast genome of *Anthoceros formosae* found to be higher (1 SSR/2.4 kb) than the density of EST-SSRs in barley, maize, wheat, rye, sorghum and rice (1 SSR/6.0 kb; Varshney et al. 2002), cotton and poplar (1 SSR/20 kb and 1 SSR/14 kb respectively; Cardle et al., 2000), Unigenes sequences of *Citrus* (1 SSR/12.9 kb; Shanker et al. 2007a) and cpSSRs of rice (1SSR/6.5 kb; Rajendrakumar et al. 2007) however, lower than the cpSSRs density in family Solanaceae (1 SSR/1.26kb; Tambarussi et al. 2009).

Generally SSRs are abundant in non-coding regions of a genome (Hancock 1995) and the results of the present study showed consistency with it. The abundance of mononucleotides in present analysis shows consistency with earlier studies of cpSSRs in rice (Rajendrakumar et al. 2007), Solanaceae species (Tambarussi et al. 2009), *Saccharum* spp. (Melotto-Passarin et al. 2011) and *Olea* species (Filiz and Koc 2012). The identified SSRs in chloroplast genome of *Anthoceros formosae* can be used to develop SSR markers and for other purposes.

References

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.

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

Hancock JM (1995) The contribution of slippage-like processes to genome evolution. J Mol Evol 41: 1038-1047.

Katti MV, Ranjekar PK and Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Ecol 18: 1161-1167.

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

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.

Shanker A, Singh A and Sharma V (2007) *In silico* mining in expressed sequences of *Neurospora crassa* for identification and abundance of microsatellites. Microbiol Res 162: 250- 256.

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

Shanker A (2012) Chloroplast genomes of bryophytes: a review. Archive for Bryology 143: 1-5.

Shanker A (2012a) Sequenced mitochondrial genomes of bryophytes. Archive for Bryology 146: 1-6.

Shanker A (2013) Paraphyly of bryophytes inferred using chloroplast sequences. Archive for Bryology 163: 1-5.

Shanker A (2013a) Inference of bryophytes paraphyly using mitochondrial genomes. Archive for Bryology 165: 1-5.

Shanker A (2013b). Combined data from chloroplast and mitochondrial genome sequences showed paraphyly of bryophytes. Archive for Bryology 171: 1-9

Tambarussi EV, Melotto-Passarin DM, Gonzalez SG, Brigati JB, de Jesus FA, Barbosa AL, Dressano K and Carrer H (2009) *In silico* analysis of simple sequence repeats from chloroplast genomes of Solanaceae species. Crop Breed Appl Biotech 9: 344-352.

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 Lett 7: 537-546.

ISSN 0945-3466

Table 1: Information of mined SSRs in chloroplast genome sequence of *Anthoceros formosae*.

S. No.	Motif	Length	Start	End	Region
1	(TATT)3	12	1014	1025	Non coding
2	(TA)20	40	14120	14159	Non coding
3	(ATT)5	15	15962	15976	Non coding
4	(CTTATT)3	18	16066	16083	Non coding
5	(T)12	12	20258	20269	Non coding
6	(A)18	18	22769	22786	Non coding
7	(A)15	15	23022	23036	Non coding
8	(TATT)3	12	23104	23115	Non coding
9	(TA)12	24	23664	23687	Non coding
10	(A)12	12	24670	24681	Non coding
11	(A)15	15	27238	27252	Non coding
12	(TTC)4	12	40937	40948	Non coding
13	(TTA)4	12	41710	41721	Non coding
14	(TAC)5	15	41726	41740	Non coding
15	(TAG)4	12	41851	41862	Non coding
16	(GTA)6	18	41865	41882	Non coding
17	(T)12	12	42193	42204	Non coding
18	(T)13	13	42994	43006	Non coding
19	(GTA)9	27	44322	44348	Coding
20	(AT)9	18	44892	44909	Non coding
21	(TA)9	18	44911	44928	Non coding
22	(A)12	12	45765	45776	Non coding
23	(TTTA)3	12	46805	46816	Non coding
24	(TAAA)3	12	46863	46874	Non coding
25	(ATAA)3	12	47448	47459	Non coding
26	(AT)7	14	49731	49744	Non coding
27	(TCTTTT)3	18	53791	53808	Non coding
28	(TTC)4	12	59470	59481	Non coding
29	(A)18	18	61852	61869	Non coding
30	(AT)20	40	62850	62889	Non coding
31	(T)15	15	64037	64051	Non coding
32	(AT)11	22	64183	64204	Non coding
33	(TA)10	20	64216	64235	Non coding
34	(A)12	12	68615	68626	Non coding
35	(AT)7	14	69529	69542	Non coding

S. No.	Motif	Length	Start	End	Region
36	(A)12	12	74760	74771	Non coding
37	(T)13	13	74841	74853	Non coding
38	(T)13	13	76704	76716	Non coding
39	(A)13	13	77462	77474	Non coding
40	(TA)23	46	80639	80684	Non coding
41	(ATTT)3	12	80768	80779	Non coding
42	(T)15	15	80995	81009	Non coding
43	(AAT)8	24	81251	81274	Non coding
44	(AT)11	22	81413	81434	Non coding
45	(TTTG)3	12	82068	82079	Coding
46	(A)13	13	83247	83259	Non coding
47	(TA)21	42	83329	83370	Non coding
48	(T)15	15	83595	83609	Non coding
49	(ATT)4	12	83738	83749	Non coding
50	(AT)8	16	85519	85534	Non coding
51	(T)13	13	90075	90087	Non coding
52	(T)14	14	90728	90741	Non coding
53	(AT)7	14	93166	93179	Non coding
54	(TA)7	14	93973	93986	Non coding
55	(AT)11	22	101156	101177	Non coding
56	(CTTT)3	12	101302	101313	Non coding
57	(A)12	12	103824	103835	Non coding
58	(AAAAAG)3	18	111063	111080	Non coding
59	(AGGT)3	12	119085	119096	Coding
60	(A)12	12	122860	122871	Non coding
61	(T)18	18	125535	125552	Non coding
62	(AAGA)5	20	126346	126365	Non coding
63	(TAA)6	18	129068	129085	Coding
64	(TA)7	14	131675	131688	Non coding
65	(T)12	12	145795	145806	Non coding
66	(CTAC)3	12	149568	149579	Coding
67	(CTTTTT)3	18	157586	157603	Non coding

ISSN 0945-3466