Computationally mined microsatellites in chloroplast genome of *Pellia endiviifolia*

Asheesh Shanker

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

Abstract: The availability of organelle genome sequences of bryophytes provides opportunity to mine this data. Therefore in this study microsatellites in chloroplast genome sequence of *Pellia endiviifolia* (Accession number: NC_019628), downloaded from the National Center for Biotechnology Information (NCBI) in fasta format, were identified. The sequence was mined with the help of MISA, a Perl script, to detect microsatellites. In total, 16 perfect microsatellites were identified in 120.546 kb sequence mined. An average length of 14.94 bp was calculated for mined microsatellites with a density of 1 SSR/7.09 kb. Depending on the repeat units, the length of microsatellites ranged from 12 to 18 bp. Tetranucleotides (7, 43.75%) were the most frequent repeat type, followed by mononucleotide (3, 18.75%) repeats. Dinucleotide, trinucleotide and pentanucleotide repeats were found with equal frequency (2, 12.5%). Interestingly, hexanucleotide repeats were completely absent in chloroplast genome of *Pellia endiviifolia*.

Introduction

Bryophytes are non-vascular plants and have been broadly classified into liverworts, mosses and hornworts. Phylogenetic analysis using organelle genome sequences inferred paraphyly of bryophytes and liverworts as early diverging land plants (Shanker 2013; Shanker 2013a; Shanker 2013b).

Microsatellites also known as simple sequence repeats (SSRs) consist of short repeat motifs (1-6 bp) and are found in DNA sequences. These repeats are present in both coding and non-coding regions of DNA sequences (Katti et al. 2001; Shanker et al. 2007). Due to abundance and ability to associate with many phenotypes, microsatellites have been widely used as molecular markers in many plant genomes. Earlier microsatellites were identified in organelle genomes of plants including bryophytes (Shanker 2013c; Shanker 2013d; Villarreal et al. 2012). However the distribution of microsatellites in chloroplast genome of *Pellia endiviifolia*, a liverwort, has not been known.

Computational approaches offer rapid and economical microsatellites extraction using sequences from public databases (Shanker et al. 2007a). Therefore in present study the chloroplast genome of *Pellia endiviifolia* was mined to understand the distribution of microsatellites.

Materials and Methods

Chloroplast genome sequence of *Pellia endiviifolia*

Chloroplast genome sequences of few bryophytes (Shanker 2012; Shanker 2012a) are available at National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov). The chloroplast genome sequence of *Pellia endiviifolia* was recently sequenced (Grosche et al. 2012). It was downloaded from NCBI in FASTA and GenBank format (NC_019628, 120546 bp).

Detection of SSRs

The chloroplast genome sequence of *Pellia endiviifolia* was mined using MISA, a Perl script (available at http://pgrc.ipkgatersleben.de/misa/misa). MISA detect microsatellites in FASTA formatted nucleotide sequence and generates output along with statistical data in two separate files. Minimum length criteria of microsatellites were considered as ≥12 for mono, di, tri and tetranucleotide, ≥15 for pentanucleotide and ≥18 for hexanucleotide repeats. Based on the presence of repeats in coding and non-coding regions, identified microsatellites were classified into coding and non-coding microsatellites. The information of coding and non-coding regions was taken from GenBank file of the chloroplast genome.

Results and Discussion

In this study perfect chloroplastic microsatellites or SSRs (cpSSRs) with a minimum length of 12 bp was identified in chloroplast genome sequence of *Pellia endiviifolia*. The length of the identified SSRs ranged from 12 to 18 bp. Hexanucleotide repeats were not detected in chloroplast genome sequence of *Pellia endiviifolia*. The distribution of mined SSRs is presented in Fig. 1. It is evident from Fig. 2 that the majority of SSRs were found in non-coding region of the genome.

The chloroplast genome sequence of *Pellia endiviifolia* harbors only 16 SSRs showing density of 1 SSR/7.09 kb in 120.546 kb sequence mined. Tetranucleotides (7, 43.75%) were the most frequent repeat type, followed by mononucleotide (3, 18.75%) repeats. Dinucleotide, trinucleotide and pentanucleotide repeats were found with equal frequency (2, 12.5%). Out of all mined SSRs 7 (43.75%) lie in coding and 9 (56.25%) lie in non-coding region of the genome. Information of mined SSRs motif, their length, start-end position and the region in which they lie is presented in Table 1.

The density of SSRs in chloroplast genome of *Pellia endiviifolia* found to be lower (1 SSR/7.09 kb) than the density of cpSSRs in *Anthoceros formosae* (1 SSR/2.4 kb; Shanker 2013c) and *Aneura mirabilis* (1 SSR/5.68 kb; Shanker 2013d). Moreover it is lower than the density of EST-SSRs in barley, maize, wheat, rye, sorghum and rice (1 SSR/6.0 kb; Varshney et al. 2002), cpSSRs of rice (1SSR/6.5 kb; Rajendrakumar et al. 2007) and cpSSRs density in family Solanaceae (1 SSR/1.26kb; Tambarussi et al. 2009). Contrary to this the density of cpSSRs in *Pellia endiviifolia* found to be higher than the density of 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). Due to different parameters including minimum length of SSRs taken, the amount of data analyzed and genomic composition of the sequence might be the cause of variation in SSR density.

ISSN 0945-3466

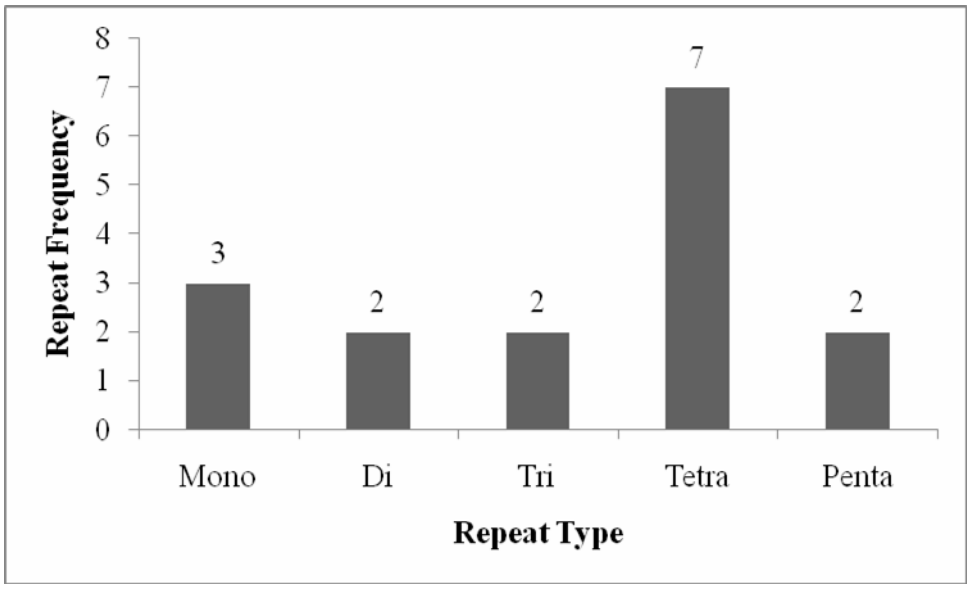

Figure 1. Frequency distribution of various repeat types.

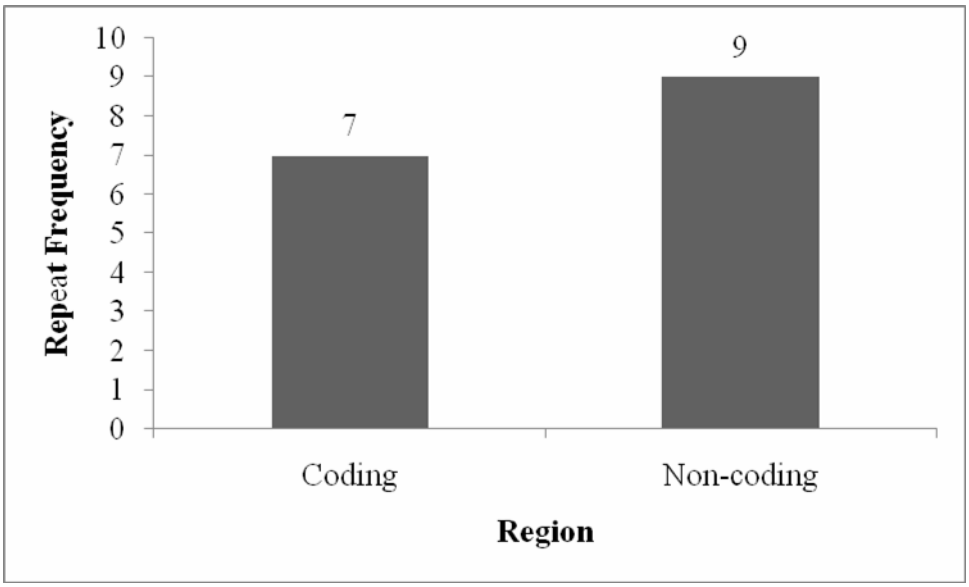

Figure 2. Distribution of identified cpSSRs in coding and non-coding regions.

Table 1: Information of mined SSRs in chloroplast genome sequence of *Pellia endiviifolia*.

S. No.	Motif	Length	Start	End	Region
1.	(TTAT)3	12	17490	17501	Non coding
2.	(AT)6	12	22587	22598	Non coding
3.	(TA)9	18	23142	23159	Non coding
4.	(ATGGT)3	15	30771	30785	Non coding
5.	(ATTTT)3	15	33586	33600	Coding
6.	(ATCA)3	12	38298	38309	Coding
7.	(TTTG)3	12	64263	64274	Coding
8.	(T)12	12	85847	85858	Non coding
9.	(ACT)4	12	86865	86876	Non coding
10.	(AGGT)3	12	88605	88616	Coding
11.	(TAAT)3	12	93860	93871	Non coding
12.	(C)17	17	97093	97109	Coding
13.	(AATA)3	12	97329	97340	Coding
14.	(CTAC)3	12	114437	114448	Coding
15.	(AGT)4	12	116179	116190	Non coding
16.	(A)12	12	117197	117208	Non coding

Similar to the cpSSRs in *Anthoceros* (Shanker 2013c) and *Aneura* (Shanker 2013d) most of the cpSSRs identified in *Pellia endiviifolia* lie in non-coding region. However the proportion of the coding SSRs is only slightly lower than non-coding SSRs. These cpSSRs can be useful to develop SSR markers and for other purposes. Moreover this study provides scientific base for phylogenetics, evolutionary genetics and diversity studies on different *Pellia* species in future.

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.

Grosche C, Funk HT, Maier UG, Zauner S (2012) The chloroplast genome of *Pellia endiviifolia*: gene content, RNA-editing pattern, and the origin of chloroplast editing. Genome Biol Evol 4: 1349-1357.

ISSN 0945-3466

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

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 (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

Shanker A (2013c) Identification of microsatellites in chloroplast genome of *Anthoceros formosae*. Archive for Bryology 191: 1-6.

Shanker A (2013d) Mining of simple sequence repeats in chloroplast genome of a parasitic liverwort: *Aneura mirabilis*. Archive for Bryology 196: 1-4.

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, 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.

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.

Villarreal JC, Forrest LL, McFarland K, Goffinet B (2012) Chloroplast, mitochondrial, and nuclear microsatellites from the southern Appalachian hornwort, *Nothoceros aenigmaticus* (Dendrocerotaceae). Am J Bot 99: e88-90.