INTERNATIONAL JOURNAL OF ENVIRONMENT

Volume-3, Issue-3, Jun-Aug 2014 ISSN 2091-2854

Received: 31 May Revised:7 July Accepted:20 August

COMPUTATIONAL MINING OF MICROSATELLITES IN THE CHLOROPLAST GENOME OF *PTILIDIUM PULCHERRIMUM*, A LIVERWORT

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

Abstract

Microsatellites also known as simple sequence repeats (SSRs) are found in DNA sequences. These repeats consist of short motifs of 1-6 bp and play important role in population genetics, phylogenetics and also in the development of molecular markers. In this study chloroplastic SSRs (cpSSRs) in the chloroplast genome of *Ptilidium pulcherrimum*, downloaded from the National Center for Biotechnology Information (NCBI), were detected. The chloroplast genome sequence of *P. pulcherrimum* was mined with the help of a Perl script named MISA. A total of 23 perfect cpSSRs were detected in 119.007 kb sequence mined showing density of 1 SSR/5.17 kb. Depending on the repeat units, the length of SSRs found to be 12 bp for mono and tri, 12 to "22 bp for di, 12 to 16 bp for tetra nucleotide repeats. Penta and hexanucleotide repeats were completely absent in chloroplast genome of *P. pulcherrimum*. Dinucleotide repeats were the most frequent repeat type (47.83%) followed by tri (21.74%) and tetranucleotide (21.74%) repeats. Out of 23 SSRs detected, PCR primers were successfully designed for 22 (95.65%) cpSSRs.

Keywords: Microsatellites, Simple Sequence Repeats, Chloroplast, Bryophytes, Liverwort

Introduction

Bryophytes are the simplest and earliest land plants. These are broadly classified into liverworts, mosses and hornworts (Alam, 2014). For bryophytes only a small number of organelle genome sequences are available (Shanker, 2012; Shanker, 2012a) which helps in the elucidation of evolutionary relationship among these plants. Phylogenetic analysis based on mitochondrial and chloroplast genome sequences of bryophytes showed liverworts as the earliest diverging lineage and hornworts as sister group to vascular plants (Shanker, 2013; Shanker, 2013a; Shanker, 2013b). In the recent past, studies were conducted to detect microsatellites in organelle genome sequences of bryophytes (Shanker, 2013c; Shanker, 2013d).

Microsatellites also known as simple sequence repeats (SSRs) are found in DNA sequences. These repeats consist of short repeat motifs of 1-6 bp and are present in both coding and non-coding regions of DNA sequences (Shanker *et al.,* 2007). SSRs have been widely used as molecular markers in plant genomes (Gupta *et al.,* 2003; Jakobsson *et al.*, 2007; Blair and Hurtado, 2013). Apart from this, a database named MitoSatPlant has been developed which provides information about mitochondrial SSRs of green plants (Kumar *et al.*, 2014). However, there are available chloroplast genome sequences of bryophytes for which we do not have detailed information of SSRs and the chloroplast genome of *P. pulcherrimum* is one of them.

The plastid genome of *P. pulcherrimum* has been known to be the first plastid genome sequenced for any bryophyte using next generation technology (Forrest *et al.*, 2011). To get the sequence the researchers prepared a shotgun library from total genomic DNA of *P. pulcherrimum* which has been subjected to highthroughput sequencing. Bioinformatic approaches have been used for the assembly and annotation of plastid reads. Moreover, a combined analysis has been conducted for nuclear, mitochondrial and plastid contigs to screen microsatellite markers using msatCommander (Faircloth, 2008). Despite these efforts the detailed information of cpSSRs in *P. pulcherrimum* is not known.

Since bioinformatic approaches offer rapid and economical SSR extraction using sequences deposited in public databases (Shanker *et al.*, 2007a). Therefore, the present analysis was conducted using bioinformatic approach to identify cpSSRs in *Ptilidium pulcherrimum*. Additionally, the distribution of these repeats in coding and non-coding regions of chloroplast genome was analyzed. Attempt was also made to design PCR primer pairs for mined cpSSRs.

Materials and Methods

Chloroplast genome sequence of *Ptilidium pulcherrimum*

The complete chloroplast genome sequence of *P. pulcherrimum* (NC_015402, 119007 bp; Forrest *et al.*, 2011) was downloaded from NCBI (www.ncbi.nlm.nih.gov) in FASTA and GenBank format.

Mining of chloroplastic simple sequence repeats

To mine SSRs in chloroplast genome sequence of *P. pulcherrimum* a Perl script named MISA (available at http://pgrc.ipkgatersleben.de/misa/misa) was used. MISA takes FASTA formatted DNA sequence file as an input and generates information of perfect and compound SSRs, if detected. In perfect SSR same repeating motif is present without interruptions, e.g., $(CTA)_8$. However, two or more SSRs are found adjacent to one another in compound SSRs, e.g., $(GAC)_8(GA)_{16}$ (Bachmann and Bare, 2004). The length of SSRs in this study was defined as ≥12 for mono, di, tri and tetranucleotide, ≥15 for pentanucleotide and ≥18 for hexanucleotide repeats. Maximum difference between two compound SSRs was taken as 0. The GenBank file contains information of coding and non-coding regions of chloroplast genome. Therefore, based on the presence of mined repeats in respective regions of chloroplast genome, these cpSSRs were classified as coding, non-coding and coding-non-coding (few bases of coding-non-coding SSRs occur in coding as well as in non-coding regions or vice-versa) SSRs.

Primer designing for mined SSRs

Primer 3 (http://bioinfo.ut.ee/primer3-0.4.0/) with default parameters of GC content, melting temperature, primer and PCR product size was used to design PCR primers for mined SSRs. SSR flanking regions of 200 base pair were considered to design primers.

Results and Discussion

The present analysis deals with the mining of SSRs in chloroplast genome sequence of *P. pulcherrimum* considering a minimum length of 12 bp. Only 23 perfect SSRs were detected with length variation from 12 to 22 bp. Compound SSRs, penta and hexanucleotide repeats were totally absent in chloroplast genome sequence of *P. pulcherrimum*. The frequency of various repeat motifs (mono-tetra) identified is presented in Fig. 1. Additional information of mined

SSRs motif, their length, start-end position and the region in which they lie is presented in Table 1. It is evident from this table that out of total cpSSRs detected, 4 (17.39%) found in coding, 17 (73.91%) in non-coding and only 2 (8.69%) in coding-non-coding regions. Generally SSRs are abundant in non-coding regions of a genome (Hancock, 1995; Shanker, 2013c) and the results of the present study showed consistency with it.

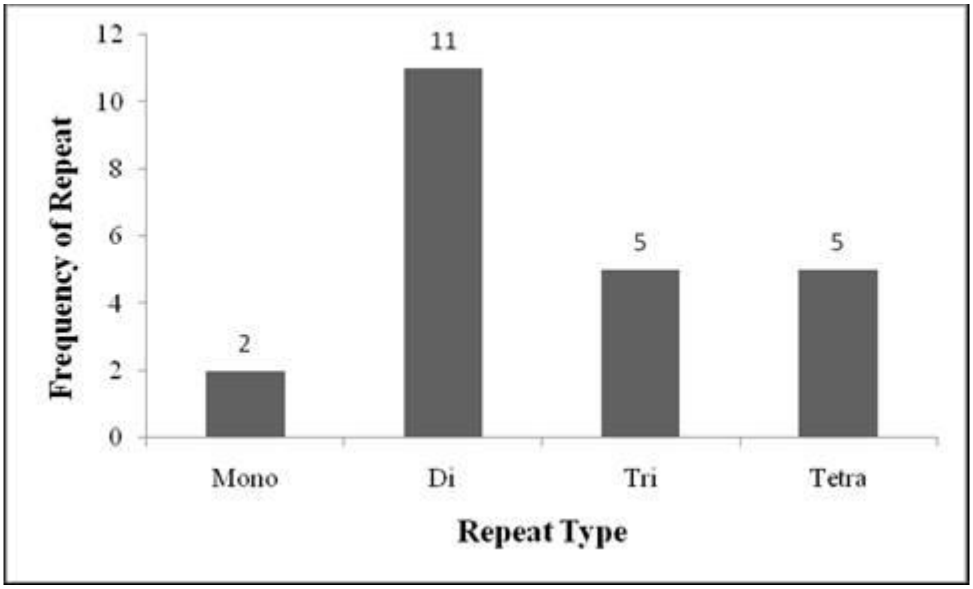

Fig. 1. Frequency distribution of mono-tetra repeats identified

Dinucleotides were the most frequent repeat (11, 47.83%) followed by tri and tetra nucleotide repeats, both present with equal frequency (5, 21.74%). Mononucleotide repeats (2, 8.7%) were the least abundant in chloroplast genome sequence of *P. pulcherrimum*. PCR primers were successfully designed for 22 (95.65%) cpSSRs identified. A list of designed PCR primers, their length, product size etc. is presented in Table 2.

Table 1: SSR motif, their length and other details of mined cpSSRs in *P. pulcherrimum*

S. No.	MOTIF	LENGTH	START	END	REGION
1	(AT)7	14	5575	5588	Non coding
2	(TA)11	22	5590	5611	Non coding
3	(AT)6	12	23045	23056	Non coding
4	(AT)7	14	24572	24585	Non coding
5	(TTAA)3	12	37528	37539	Non coding
6	(TA)6	12	37605	37616	Non coding
7	(ATA)4	12	47904	47915	Non coding
8	(ATTT)4	16	48748	48763	Non coding
9	(T)12	12	50293	50304	Non coding
10	(AT)6	12	50668	50679	Non coding
11	(T)12	12	60189	60200	Coding
12	(ATT)4	12	63777	63788	Non coding
13	(TAA)4	12	68491	68502	Non coding
14	(AT)9	18	72396	72413	Coding-Non coding
15	(TA)7	14	75132	75145	Non coding
16	(AT)6	12	78395	78406	Coding-Non coding
17	(AGGT)3	12	87159	87170	Coding
18	(TA)6	12	92349	92360	Non coding
19	(TAA)4	12	93635	93646	Non coding
20	(AT)7	14	94071	94084	Non coding
21	(AATA)3	12	95824	95835	Coding
22	(TAA)4	12	102754	102765	Non coding
23	(CTAC)3	12	112861	112872	Coding

Table 2: PCR primers designed for mined SSRs along with additional information

S. No.	MOTIF	START	END	LEFT / RIGHT PRIMER	PRIMER LENGTH (bases)	Tm	GC%	PRODUCT LENGTH (bases)
1	(AT)7	5575	5588	CCATTAAGGCCCCCAAGCT	20	60.325	55	261
				TCCATCGTATTATAGACAACCCAT	24	57.072	37.5	
2	(TA)11	5590	5611	CCATTAAGGCACCCCAAGCT	20	60.325	55	261
				TCCATCGTATTATAGACAACCCAT	24	57.072	37.5	
3	(AT)6	23045	23056	TCATTCTGGTTCAACTTCTCCT	22	57.018	40.909	235
				GGAACCATTACTTTTCCTTCTCCC	24	59.294	45.833	
4	(AT)7	24572	24585	GTTCGAATCCTTCCGTCCCA	20	59.752	55	222
				TAGCTACGCGCAAAGTTCCA	20	60.038	50	
5	(TTAA)3	37528	37539	GAGGGTCGTCTCTTGAAAACCT	22	59.963	50	245
				TGAACCGATGACTTACGCCT	20	59.107	50	
6	(TA)6	37605	37616	GAGGGTCGTCTCTTGAAAACCT	22	59.963	50	245
				TGAACCGATGACTTACGCCT	20	59.107	50	
7	(ATA)4	47904	47915	TCCCCCTCAGATTGAGCTGA	20	59.957	55	177
				GCCCAAATAGTTTATGAGGTTGGT	24	59.29	41.667	
8	(ATTT)4	48748	48763	TTCATTGTGTCTTCGTTTAACAGA	24	57.088	33.333	185
				AGGGATTCGAACCCTCGGTA	20	60.033	55	
9	(T)12	50293	50304	GGTGCAGAGACTCAAAGGGA	20	59.313	55	213
				CGATTTTCATCGCGGCTAAA	20	57.27	45	
10	(AT)6	50668	50679	TAGCCGGGATAGCTCAGTTG	20	58.959	55	247
				TGCCCGAGAGTTGGATAGGT	20	60.325	55	
11	(T)12	60189	60200	TCCACCTTTTGAGATTTATGCTATTT	26	57.406	30.769	181
				TCAAAGTTGCCTCAATCCAAGC	22	59.703	45.455	
12	(ATT)4	63777	63788	TTGCTCCGTGTAAACATCAAATT	23	57.554	34.783	250
				AGGAACCTAATGACAATGTCGT	22	57.447	40.909	
13	(TAA)4	68491	68502	ACTTTCGGAACACCAATAGGCA	22	60.225	45.455	249
				TGTCACCGGGATCATAGTATCG	22	59.182	50	
14	(AT)9	72396	72413	CGTAAACAAGGTATTTCGGGTCC	23	59.628	47.826	185
				AGTCACACACTCCCATAATCCA	22	58.819	45.455	
15	(TA)7	75132	75145	CCGCGGAAGACCAAGAAACA	20	60.883	55	238
				TGGTGGTTTGTTCTAATCCGA	21	57.227	42.857	
16	(AT)6	78395	78406	GCGAACCAATACGAATGATGACT	23	59.444	43.478	179
				CAAGAGCTCAAGGACGTGGT	20	59.966	55	
17	(AGGT)3	87159	87170	CTGTACCCGAAACCGACACA	20	59.968	55	189
				TCTTACGACTTTGCGGGGAC	20	60.038	55	
18	(TA)6	92349	92360	Primer not found				
19	(TAA)4	93635	93646	TGATCTTGCAACTTCGGTGGA	21	59.928	47.619	150

				ATTTTGCGAAAAACGGGTTGT	21	58.105	38.095	
20	(AT)7	94071	94084	AGTGCCACTATTTTTGCGAGT	21	58.496	42.857	214
				TTGGGGTGATGGAAGTCGTG	20	59.964	55	
21	(AATA)3	95824	95835	TTTCTCGTGGTCCAGCATCC	20	60.036	55	222
				ACCTGGTACTAGTGGTTTTGCA	22	59.56	45.455	
22	(TAA)4	102754	102765	TGTGATAGGAAATGTGGTGGTT	22	57.619	40.909	156
				TGGTCCAGTTATCGCTTCGA	20	58.821	50	
23	(CTAC)3	112861	112872	TCTTACGACTTTGCGGGGAC	20	60.038	55	189
				CTGTACCCGAAACCGACACA	20	59.968	55	

The mined SSRs represent a density of 1 SSR/5.17 kb in 119.007 kb sequence mined. The density of cpSSRs in *P. pulcherrimum* found to be higher than the density of cpSSRs in *Aneura mirabilis* (1 SSR/5.68 kb; Shanker, 2013d) and *Pellia endiviifolia* (1 SSR/7.09 kb; Shanker, 2014), rice (1 SSR/6.5 kb; Rajendrakumar *et al.*, 2007), 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.*, 2007). However, the density of cpSSRs in *P. pulcherrimum* found to be lower than the cpSSRs density in *Anthoceros formosae* (1 SSR/2.4 kb; Shanker, 2013c), family Solanaceae (1 SSR/1.26kb; Tambarussi *et al.*, 2009). The selection of SSR detection tools, parameters taken (e.g. minimum length of SSRs) and amount of data analyzed might be the cause of variations in SSR density.

The higher occurrence of dinucleotide repeats in this study shows inconsistency with earlier studies of cpSSRs in bryophytes. In the recent past mononucleotides were found to be abundant in the chloroplast genome of *Anthoceros formosae* (Shanker, 2013c), *Aneura mirabilis* (Shanker, 2013d) and *Pellia endiviifolia* (Shanker, 2014). However, in rice dinucleotide repeats were found to be the most abundant repeats in genic and intergenic regions but in the mitochondrial genome (Rajendrakumar *et al.*, 2007). The absence of pentanucleotide repeats in chloroplast genome of *P. pulcherrimum* is in agreement with cpSSRs studies on *Anthoceros formosae* (Shanker, 2013c). Moreover, the absence of hexanucleotide repeats shows consistency with cpSSRs studies on *Aneura mirabilis* (Shanker, 2013d) and *Pellia endiviifolia* (Shanker, 2014). It was suggested that the abundance of these repeats attributed to the evolutionary processes that fine tune distribution of SSR repeat types in genome (Lin and Kussell, 2012).

Conclusion

In silico mining of complete chloroplast genome sequence of *P. pulcherrimum* saves time, cost and provides sufficient number of SSRs for this liverwort. The designed PCR primers can be used to develop SSR markers. Once developed SSR markers will help in the diversity and phylogenetic analysis of *Ptilidium* species.

Acknowledgment

I am thankful to University Grants Commission, India for providing financial support in the form of Major Research Project (F.No. 42-138/2013).

References

Alam, A., 2014. Morphotaxonomy of three rare terricolous taxa of Jungermanniales occurring in Nilgiri Hills (Western Ghats) India. International Journal of Environment 3: 85-92.

Bachmann, L. and Bare, P.T.J., 2004. Allelic variation, fragment length analyses and population genetic model: A case study on *Drosophila* microsatellites. J. Zool. Syst. Evol. Res. 42: 215-222.

Blair, M.W. and Hurtado, N., 2013. EST-SSR markers from five sequenced cDNA libraries of common bean (*Phaseolus vulgaris* L.) comparing three bioinformatic algorithms. *Mol. Ecol. Resour.* 13: *688-695*.

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: *847-854*.

Faircloth, B.C., 2008. MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Mol. Ecol. Resour*. 8: *92-94*. DOI:10.1111/j.1471-8286.2007.01884.x

Forrest, L.L., Wickett, N., Cox, C.J. and Goffinet, B., 2011. Deep sequencing of *Ptilidium* (Ptilidaceae) suggests evolutionary stasis in liverwort plastid genome structure. *Pl. Ecol. Evol.* 144: *29-43*.

Gupta, P.K., Rustgi, S., Sharma, S., Singh, R., Kumar, N. and Balyan, H.S., 2003. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol. Genet. Genomics* 270: *315-323*.

Hancock, J.M., 1995. The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.* 41: *1038-1047*.

Jakobsson, M., Säll, T., Lind-Halldén, C. and Hallden, C., 2007. Evolution of chloroplast mononucleotide microsatellites in *Arabidopsis thaliana*. *Theor. Appl. Genet*. 114: *223-235*.

Kumar, M., Kapil, A. and Shanker, A., 2014. MitoSatPlant: Mitochondrial microsatellites database of Viridiplantae. *Mitochondrion*, http:// dx.doi.org/10.1016/j.mito.2014.02.002

Lin, W.H. and Kussell, E., 2012. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic Acids Res.* 40: *2399-2413*.

Rajendrakumar, P., Biswal, A.K., Balachandran, S.M., Srinivasarao, K. and Sundaram, R.M., 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., 2014. Computationally mined microsatellites in chloroplast genome of Pellia endiviifolia. *Archive for Bryology* 199: *1-5*.

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

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

Tambarussi, E.V., Melotto-Passarin, D.M., Gonzalez, S.G., Brigati, J.B., de Jesus, F.A., Barbosa, A.L., 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, R.K., 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*.