Microsatellite analysis in organelle genomes of Chlorophyta

Himani Kuntal[1], Vinay Sharma[1]* & Henry Daniell[2]

[1]Department of Biosciences and Biotechnology, Banasthali University, Banasthali–304022 Rajasthan, India; [2]Department of Molecular Biology and Microbiology, University of Central Florida, College of Medicine, Orlando, Florida, USA; Vinay Sharma - Email: vinaysharma30@yahoo.co.uk; *Corresponding author

Abstract:
Simple Sequence Repeats (SSRs) or microsatellites constitute a significant portion of genomes however; their significance in organellar genomes has not been completely understood. The availability of organelle genome sequences allows us to understand the organization of SSRs in their genic and intergenic regions. In the present work, SSRs were identified and categorized in 14 mitochondrial and 22 chloroplast genomes of algal species belonging to Chlorophyta. Based on the study, it was observed that number of SSRs in non-coding region were more as compared to coding region and frequency of mononucleotides repeats were highest followed by dinucleotides in both mitochondrial and chloroplast genomes. It was also observed that maximum number of SSRs was found in genes encoding for beta subunit of RNA polymerase in chloroplast genomes and NADH dehydrogenase in mitochondrial genomes. This is the first and original report on whole genomes sequence analysis of organellar genomes of green algae.

Key words: SSRs, microsatellites, Chlorophyta, chloroplast, mitochondria, green algae.

Background:

Microsatellites also known as simple sequence repeats (SSRs) or simple sequence length polymorphisms (SSLPs) are small arrays of tandem repeats of one to six nucleotides that are interspersed throughout the genome [1]. They are ubiquitous and highly abundant in prokaryotes and eukaryotes, present even in the smallest bacterial genomes. Microsatellites can be found anywhere in the genome, both in coding and non-coding regions, may arise through an error in the process responsible for copying of the genome during cell division and are unavoidable products of genome replication [2]. Microsatellites can be classified as perfect, imperfect and compound repeats. SSRs were initially considered to be evolutionary neutral but now they are known to play an important role in genome evolution and hot spot of recombination due to their high mutability [3]. They are thought to be involved in gene expression, regulation and function as transcriptional activating elements. SSRs are inherently unstable and are inherited in a mendelian manner and so they can be used for checking genetic relationships [4]. SSRs are highly variable that is number of repeat units in the array is different in different members of a species [5]. SSRs act as powerful genetic markers as they are locus-specific, co-dominant, PCR-based and highly polymorphic [6]. Microsatellites are highly applied in the field of forensics for DNA fingerprinting, paternity studies, diagnosis and identification of disease and population studies [2, 7, 3, 8].

The growing numbers of completed genome sequences in eukaryotic organisms from fungi to humans have greatly assisted understanding of SSRs at the genome- wide level. One obvious observation from such studies have been that the distribution of SSRs in the genome is not random in several respects; differential distribution in terms of abundance of SSRs in between intronic and intergenic regions 5' and 3' UTRs, and

different chromosomes; and different species have different frequencies of SSR types and repeat units **[9].**

Most of the previous studies on microsatellites distribution are based on sequence databases in which coding or gene-rich regions are overrepresented and are generally nuclear- genome based **[10]** but in current sequenced genome revolution, the complete organellar genomes permit the determination of frequencies of SSRs at the whole genome level. A significant feature of organellar genomes is that they are uniparentally inherited and not perturbed by recombination **[11]** thus the observed variation may be relevant to understand their maternal mode of transmission. In present study, chloroplast and mitochondrial genomes of organisms belonging to Chlorophyta are analysed for SSRs identification and characterization to detect the degree of polymorphism in them.

Methodology:
Organelle genome sequence source
The mitochondrial and chloroplast genome sequence data, belonging to Chlorophyta, were retrieved from NCBI's Genome data bank www.ncbi.nlm.nih.gov. The plant species used in study are listed in **Table 1 (see supplementary material)** (chloroplast) and **Table 2 (see supplementary material)** (mitochondria) with their accession numbers. Selection of organisms was based on the availability of completely sequenced chloroplast or mitochondrial genomes.

Identification of SSRs
SSRs were identified using Perl script MISA http://pgrc.ipk-gatersleben.de/misa/misa.html, which detects perfect SSRs only. The frequency of SSRs according to size and type of constituting SSRs were determined. MISA considered identifying motifs of one to six nucleotides in size. The minimum repeat unit was defined as ten for mononucleotides, six for dinucleotide and five for all the higher order motifs including tri-, tetra-, penta-, and hexanucleotides. MISA also detects compound microsatellites. The maximal number of interrupting base pairs in a compound microsatellite was set to 100. The occurrence of repeats in genic and intergenic regions and functional categorization of sequences having microsatellites was identified based on the sequence annotation information available in GenBank database.

Discussion:
The results obtained from both organelles had not shown any relation between them, they were different in terms of frequency and type of microsatellite present. Chloroplast and mitochondrial genomes sequences are different from nuclear genome in terms of frequency and SSR patterns present in them.

Frequency of SSRs in chloroplast and mitochondrial genomes
The number of SSRs was varying in different organelle genomes of Chlorophyta. The frequency of SSRs is highest in members of family Trebouxiophyceae and Chlorophyceae in chloroplast genomes and, in Chlorophyceae and Pendinophyceae in case of mitochondrial genomes. In comparison to chloroplast genomes the frequencies of SSRs in mitochondrial genomes were much lower. The chloroplast genomes also contain compound microsatellites in larger number as compared to mitochondrial genomes. The

chloroplast genomes of members of Trebouxiophyceae and Chlorophyceae had represented the major compound SSRs containing genomes; in other family members either the number of compound SSR was much less or zero. In mitochondrial genomes no such family based characteristics were observed. The numbers of compound SSRs was consistently low in mitochondrial genomes. In chloroplast genome sequence of *Ostreococcus tauri, Nephroselmis olivacea* and mitochondrial genome sequence of *Chlamydomonas reinhardtii*, no SSRs were detected. In common algal species existing in both chloroplast and mitochondrial datasets, there was no relationship between the frequencies of simple and compound SSRs. Frequencies of SSRs were less in the Chlorophyta organelle genomes in comparison with organelle genomes of members of Streptophyta **[12, 13].**
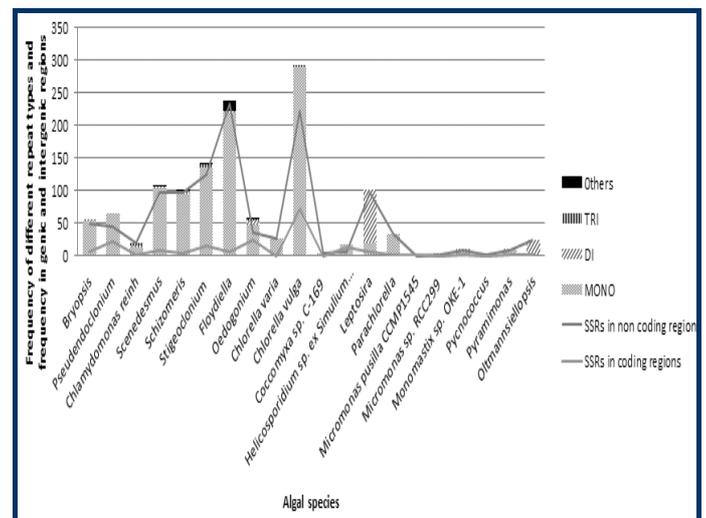

Figure 1: Frequency of different type of microsatellite repeats and abundance of SSRs in genic and intergenic regions of chloroplast genomes
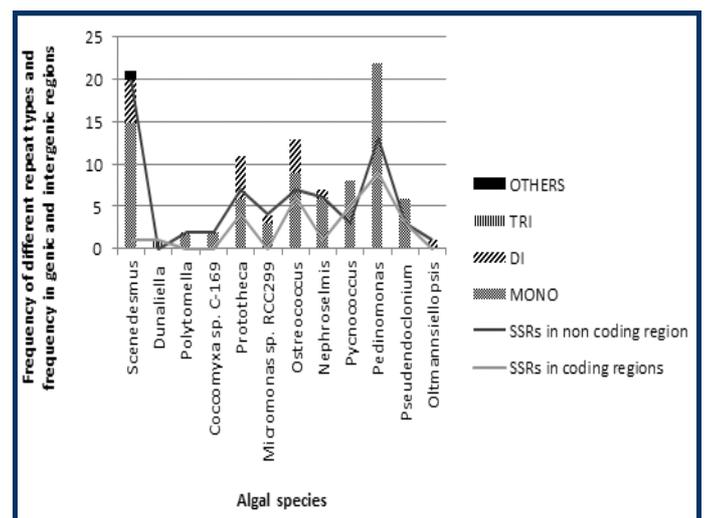

Figure 2: Frequency of different type of microsatellite repeats and abundance of SSRs in genic and intergenic regions of mitochondrial genomes.

Distribution of SSRs in genic and intergenic regions
In both chloroplast and mitochondrial genomes the major proportion of SSRs were detected in intergenic regions as

compared to genic and intragenic regions **(Figure 1 & 2).** In chloroplast genome sequences maximum microsatellites in coding regions were represented by *Chlorella vulgaris* 1.48% of gross genic region. In other chloroplast genome sequences the microsatellite content in genic regions ranged from 0.01% to 0.57% of total coding content of genomes. In mitochondrial genomes the number of SSRs in genic regions was less which ranged from 0.02% to 0.68% and in almost 50% of available sequenced mitochondrial genomes no microsatellites in genic regions were observed. Interestingly the overall content of microsatellites in genic region of mitochondrial genomes was more than the chloroplast genomes, despite the fact that the former has overall less SSR content.

The number of SSRs detected in intergenic regions of chloroplast and mitochondrial genomes is higher than SSRs in genic regions, which represented a similar pattern with SSR distribution in organelle genomes of Streptophytes **[12, 13].** The differences in frequency of microsatellites in genic and intergenic regions suggest that the polymorphism associated with coding regions is lower than non-coding region. The density of microsatellites in mitochondrial genomes of Streptophyta ranged from 1 SSR/2.06 to SSR/75.27 kb **[13]** whereas in Chlorophyta this density ranged from 1 SSR/1.25 to 1 SSR/56.76 kb **Table 2 (see supplementary material)** Polymorphisms associated with a specific locus are due to the variation in length of the microsatellite, which in turn depends on the number of repetitions of the basic motif **[14].** The lower substitution rate of chloroplast DNA compared to the nuclear genome has been documented **[15]** and the mutational processes at simple repeat loci in the chloroplast genome also occur less frequently than those in the nuclear genome **[11].** The microsatellite patterns observed on various loci in chloroplast and mitochondrial genomes of Chlorophyta differed from each other making it impossible to measure associated relatedness and polymorphism present within alleles.

SSR Patterns

Most prominent repeat patterns observed in organelle genomes of members of Chlorophyta were mono-nucleotide repeats, poly A/T. The length distribution of SSRs indicated that the frequency of repeats decreases with repeat length **(Figure 1 & 2).**

Mono-Di-Trinucleotide and other repeats

In both chloroplast and mitochondrial genomes the most prominent pattern of poly A/T repeats observed was A/T_{10} and others were A/T_{11-20}, in chloroplast genomes and A/T_{11-14}, in mitochondrial genomes. Poly C/G was either absent in most chloroplast and mitochondrial genomes sequences or in very less frequency in few genomes. Dinucleotide repeat patterns were less frequent as compared to mononucleotide repeats in both organelle genome sequences. In both genome sequences the maximum dinucleotide repeats observed was poly AT/TA and most frequent was AT/TA_6. In chloroplast genomes other observations were AT/TA_{7-12} while in mitochondria, they were AT/TA_8, AT/TA_{11-13}. Another rare pattern, AC/GT_6, was also detected in mitochondrial genome of *Ostreococcus tauri*. The major di-nucleotide repeat pattern observed in Chlorophyta organellar genomes had shown the similarity with organelle and nuclear genomes of higher taxa **[12, 13, 16];** but in nuclear genome of *Chlamydomonas* it is AC/GT **[16].**

Trinucleotide repeats of AAT/TTA_5 type were observed in chloroplast genomes in family members of Chlorophyceae and mitochondrial genomes of *Dunaliella salina*. In very minute amounts other repeat pattern tetra-, penta-, hexa- were also observed in chloroplast genomes of Chlorophycean members but mitochondrial genomes were devoid of more than tri-nucleotide repeats, except a single panta-nucleotide repeat observed in *Scenedesmus obliquus*.

Organelle genomes of streptophytes also revealed a similar pattern representing abundance of mononucleotide A/T in both the organellar genomes in comparison to C/G repeats **[12, 13, 17]** whereas SSRs in nuclear ESTs of other plant species and some cereal species tri-nucleotide repeats were the most abundant class followed by di-nucleotide repeats **[8, 18].**

Functional categorization of genes having microsatellites

In all chloroplast genomes, maximum numbers of SSRs were detected in the genes encoding for beta subunit of RNA polymerase, followed by genes encoding for cell division proteins, photosystem II proteins, hypothetical proteins, photosystem I proteins, group II intron, endonuclease, DNA polymerase and transport proteins. In mitochondrial genomes, maximum number of genes containing SSRs encode for NADH dehydrogenase followed by genes encoding for ribosomal proteins, ATP synthase genes, hypothetical, cox2 and transporter proteins. Among the chloroplast and mitochondrial genomes of major cereals maximum number of SSRs was found in *rpo* and *ndh* gene clusters **[12].**

Conclusion:

To studying the distribution pattern of SSRs in organellar genomes of green algae, the mitochondrial and chloroplast genomes of Chlorophyta have been analysed and it is concluded that distribution pattern varies significantly in mitochondrial and chloroplast genomes. The distribution of SSR pattern diversed with the genomic regions, characterized by species and organelle examined. The number of SSRs in chloroplast genomes is higher in comparison to mitochondrial genomes and maximum numbers of SSRs are found in intergenic regions than genic regions in both chloroplast and mitochondrial genomes. Mononucleotides are found to be the most abundant repeat type in both organelles and the repeat motifs are not evenly distributed. The overall representation of SSRs in mitochondrial and chloroplast genomes of Chlorophyta demonstrated that distribution pattern of SSRs in organellar genomes is not uniform and two organelles showed different patterns and arrangement of microsatellites. The study is important in term of revealing the simple repeat patterns in organellar genomes of lower plant species and to measure their abundance and polymorphism. This can be used to suggest the evolution of chloroplast and mitochondrial genomes independent of nuclear genomes. Significant differences in SSR patterns among members of same family also suggest that organellar genomes can be affected by evolutionary factors with different rate and randomness in contrast to nuclear genomes.

BIOINFORMATION

References:
[1] Dhar M *et al. Plant Sci.* 1988 **55**: 43
[2] Subramanian S *et al. Genome Biol.* 2003 **4**: R13 [PMID: 12620123]
[3] Toth G *et al. Genome Res.* 2000 **10**: 967 [PMID: 10899146].
[4] Jarne P & Lagoda PJ, *Trends Ecol Evol.* 1996 **11**: 424 [PMID:21237902]
[5] Legendre M *et al. Genome Res.* 2007 **17**: 1787 [PMID: 17978285]
[6] Grellet F *et al. J Mol Bio.* 1986 **187**: 495 [PMID: 3012090]
[7] Shakyawar SK *et al. Bioinformation* 2009 **4**: 158 [PMID: 20198193]
[8] Varshney RK *et al. Cell Mol Biol Lett.* 2002 **7**: 537 [PMID: 12378259]
[9] Lawson & Zhang, *Genome Biol.* 2006 **7**: R14. [PMID: 16507170]
[10] Zapiola ML *et al. Mol Ecol Resour.* 2010 **10**: 738 [PMID: 21565082].
[11] Provan J *et al. Genetics.* 1999 **153**: 943 [PMID: 10511569]
[12] Rajendrakumar P *et al. In Silico Biol.* 2008 **8**: 87 [PMID: 18928198]
[13] Kuntal H & Sharma V, *OMICS.* 2011 **15**: 783 [PMID: 22011339]
[14] Pagnotta MA *et al. Diversity.* 2009 **1**: 19
[15] Wolfe KH *et al. Proc Natl Acad Sci USA.* 1987 **84**: 9054 [PMID: 3480529]
[16] von Stackelberg M *et al. BMC Plant Biol.* 2006 **6**: 9 [PMID: 16734891]
[17] Melotto-Passarin DM *et al. Genet Mol Res.* 2011 **10** : 2024 [PMID: 21948764]
[18] Jayashree B *et al. In Silico Biol.* 2006 **6**: 607 [PMID: 17518768]

BIOINFORMATION

Supplementary material:

Table 1: List of algae species used for SSR analysis in Chloroplast genomes

Algae Species	Accession no.	Genome size	1SSR/kb
Bryopsis hypnoides	NC_013359	153429 nt	2.79
Pseudendoclonium akinetum / strain:UTEX 1912	NC_008114	195867 nt	2.96
Chlamydomonas reinhardtii	NC_005353	203828 nt	10.19
Scenedesmus obliquus / strain:UTEX 393	NC_008101	161452 nt	1.5
Schizomeris leibleinii	NC_015645	182759 nt	1.8
Stigeoclonium helveticum / strain:UTEX 441	NC_008372	223902 nt	1.58
Floydiella terrestris	NC_014346	521168 nt	2.18
Oedogonium cardiacum / strain:SAG 575-1b	NC_011031	196547 nt	3.38
Chlorella variabilis plastid / isolate:NC64A	NC_015359	124579 nt	4.61
Chlorella vulgaris	NC_001865	150613 nt	0.51
Coccomyxa sp. C-169 plastid / isolate:C-169	NC_015084	175731 nt	43.93
Helicosporidium sp. ex Simulium jonesii	NC_008100	37454 nt	2.08
Leptosira terrestris	NC_009681	195081 nt	1.93
Parachlorella kessleri	NC_012978	123994 nt	3.64
Micromonas pusilla CCMP1545 / strain:CCMP1545	NC_012568	41811 nt	2.9
Ostreococcus tauri / strain:OTTH0595	NC_008289	71666 nt	no SSRs
Micromonas sp. RCC299 / strain:RCC299	NC_012575	72585 nt	3.29
Monomastix sp. OKE-1	NC_012101	114528 nt	12.72
Nephroselmis olivacea	NC_000927	200799 nt	no SSRs
Pycnococcus provasolii / strain:CCMP 1203	NC_012097	80211 nt	80.21
Pyramimonas parkeae / strain:CCMP 726	NC_012099	101605 nt	10.16
Oltmannsiellopsis viridis	NC_008099	151933 nt	6.07

Table 2: List of algae species used for SSR analysis in mitochondrial genomes

Algae Species	Accession no.	Genome size	1SSR/kb
Chlamydomonas eugametos	NC_001872	22897 nt	0
Chlamydomonas reinhardtii	NC_001638	15758 nt	0
Coccomyxa sp. C-169	NC_015316	65497 nt	32.73
Dunaliella salina	NC_012930	28331 nt	28.33
Micromonas sp. RCC299	NC_012643	47425 nt	11.86
Nephroselmis olivacea	NC_008239	45223 nt	6.46
Oltmannsiellopsis viridis	NC_008256	56761 nt	56.76
Ostreococcus tauri	NC_008290	44237 nt	3.40
Pedinomonas minor	NC_000892	25137 nt	1.25
Polytomella capuana	NC_010357	12998 nt	6.50
Prototheca wickerhamii	NC_001613	55328 nt	5.03
Pseudendoclonium akinetum	NC_005926	95880 nt	15.98
Pycnococcus provasolii	NC_013935	24321 nt	3.04
Scenedesmus obliquus	NC_002254	42781 nt	2.04