Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 15 Issue 3 Version 1.0 Year 2015
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Computational Analysis of Microsatellite Repeats in Chloroplast Genomes

By G. V. Padma Raju, P. Srinivasa Rao & V. Chandra Sekhar

SRKR Engineering College Andhra University, India

Abstract- Chloroplasts are the food producers of the cell. These organelles are found only in plant cells and algae. Chloroplasts work to convert light energy of the Sun into sugars that can be used by cells. Microsatellites are a special class of DNA repeats that are found to be helpful to understand evolution, diseases and are widely used in various applications including, DNA Fingerprinting, Paternity Studies, Linkage Analysis etc. These repeats are ubiquitously present in all genomes including chloroplasts and very little is known about their presence in organelle genomes. In this study, we have analyzed more than 370 chloroplast genomes and a brief report on the distribution and frequency of these repeats in chloroplast genomes has been presented.

Keywords: chloroplast; microsatellites; bioinformatics; genomes; repeats; distribution; computational analysis.

GJCST-C Classification : H.2.8

Strictly as per the compliance and regulations of:

Computational Analysis of Microsatellite Repeats in Chloroplast Genomes

G. V. Padma Raju [α], P. Srinivasa Rao [σ] & V. Chandra Sekhar [ρ]

Abstract- Chloroplasts are the food producers of the cell. These organelles are found only in plant cells and algae. Chloroplasts work to convert light energy of the Sun into sugars that can be used by cells. Microsatellites are a special class of DNA repeats that are found to be helpful to understand evolution, diseases and are widely used in various applications including, DNA Fingerprinting, Paternity Studies, Linkage Analysis etc. These repeats are ubiquitously present in all genomes including chloroplasts and very little is known about their presence in organelle genomes. In this study, we have analyzed more than 370 chloroplast genomes and a brief report on the distribution and frequency of these repeats in chloroplast genomes has been presented.

Keywords: chloroplast; microsatellites; bioinformatics; genomes; repeats; distribution; computational analysis;.

I. Introduction

Chloroplasts, the organelles responsible for photosynthesis, are in many respects similar to mitochondria. Both chloroplasts and mitochondria function to generate metabolic energy, evolved by endosymbiosis, contain their own genetic systems, and replicate by division. However, chloroplasts are larger and more complex than mitochondria, and they perform several critical tasks in addition to the generation of ATP. Most importantly, chloroplasts are responsible for the photosynthetic conversion of Carbon Di-oxide to carbohydrates. In addition, chloroplasts synthesize amino acids, fatty acids, and the lipid components of their own membranes. The reduction of nitrite to ammonia, an essential step in the incorporation of nitrogen into organic compounds, also occurs in chloroplasts. Moreover, chloroplasts are only one of several types of related organelles (plastids) that play a variety of roles in plant cells[1-7].

Microsatellites (sometimes referred to as a variable number of tandem repeats or VNTRs) are short segments of DNA that have a repeated sequence, and they tend to occur in DNA. In some microsatellites, the repeated unit may occur four times, in others it may be seven, or two, or three[8]. These repeats are ubiquitous in nature and are responsible for causing several diseases and cancers [9][10].

Author α: Professor & HOD, CSE, SRKR Engineering College, Bhimavaram, AP, India. e-mail: gvpadmaraju@gmail.com
Author σ: Professor & HOD, CS & SE, AU College of Engineering (A), Andhra University, Visakhapatnam, AP, India.
Author ρ: Associate Professor, CSE SRKR Engineering College, Bhimavaram, AP, India.

These are used in various applications like DNA Fingerprinting, DNA Forensics, Paternity Studies, and have been considered as potential markers for identifying species, for establishing phylogenetic relationships and also to study evolution [11]. Microsatellites are ubiquitously found in both coding and non-coding regions of all organisms and their distribution in coding regions (genes) is known to affect protein formation and gene regulation [12].

Next-generation sequencing enabled researchers to study biological systems at a level never before possible. Studying mutations in chloroplast microsatellite repeats can be very helpful to understand various biological questions and their usage in various other diverse applications. Few studies [13-16] earlier analyzed the distribution of microsatellites in chloroplast genomes but they are only confined to single or very low number of genomes. This paper describes the study performed to analyze microsatellite repeats in more than 370 chloroplasts genomes and details have been presented.

II. Materials & Methods

Imperfect microsatellites have been extracted from Chloro Mito SSRDB[17] version 2.0, an open-source microsatellite repository of sequenced organelle genomes. For this study, a total of 370 chloroplast genome sequences have been used that belong to various classes as shown in Table 1.

Table 1 : Category-wise chloroplast genomes used in this analysis and their numbers.

Category	Total No.
Alveolata	9
Cryptophyta	3
Euglenozoa	5
Glaucocystophyceae	1
Haptophyceae	4
Rhizaria	2
Rhodophyta	9
Stramenopiles	14
Viridiplantae	323
Total Genomes	370

Among the 370 genomes, 323 genomes belong to Viridiplantae (Green Plants), 47 genomes belongs to Non-Viridiplantae which include genomes of Alveolata, Cryptophyta, Euglenozoa, Glaucocystophyceae, Haptophyceae, Rhizaria, Rhodophyta and

Stramenopiles (Refer Table 1). A total of 78,536 microsatellites from these 370 genomes have been analyzed by querying the database Chloro MitoSSRDB2.0 using in-house C and Java programs. The current study focuses on microsatellite distribution and their frequency of occurrence in the two major categories namely Viridiplantae, and Non-Viridiplantae.

III. Discussion

a) Genome Size Analysis

We did a preliminary study to analyze the genome sizes of all chloroplasts. The chloroplast genome sizes vary from few kbs to a maximum of 1 Mb. The smallest chloroplast genome reported is of size 29529bp that belongs to plant named Plasmodium falciparum HB3 apicoplast (ID: NC_017928) belongs to Non- Viridiplantae category. The largest chloroplast genome spans about 1021616 bp of length that belongs to Paulinella chromatophora chromatophore (ID: NC_011087) belongs to Rhizaria.

In Viridiplantae, the smallest chloroplast genome is Helicosporidium sp. ex Simulium jonesii plastid(ID: NC_008100) of length 37454 bp where as the largest chloroplast genome is Floydiella terrestris(ID: NC_014346) chloroplast of length 521168 bp.

In Non- Viridiplantae, the smallest chloroplast genome is found as Plasmodium falciparum HB3 apicoplast (ID: NC_017928) of length 29529 bp where as the largest chloroplast genome is Paulinella chromatophora chromatophore (ID: NC_011087) chloroplast of length 1021616 bp. It is observed that this non-Virdiplantae category genome size is greater than the Viridiplantae genomes.

When the average genome sizes of chloroplast are considered category wise, it has been observed that the average lengths of Viridiplantae chloroplast genomes are little bit higher when compared to those of other non Virdiplantae(Refer Fig 1).

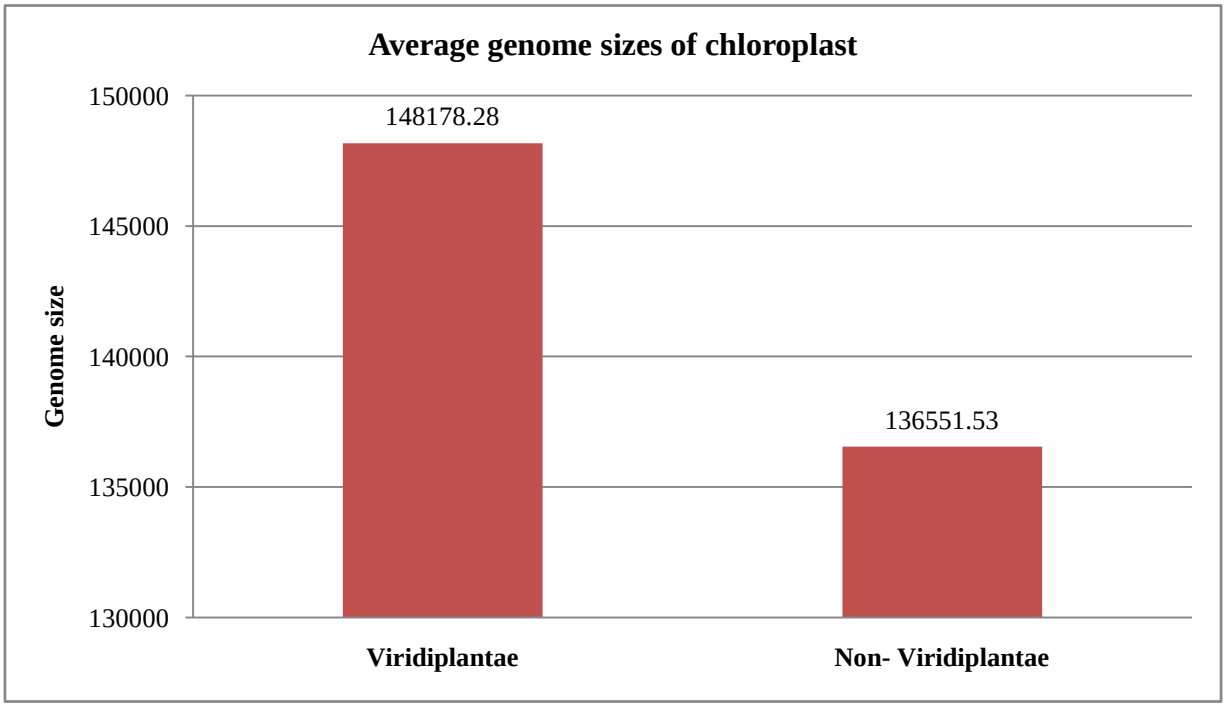

Figure 1 : Bar Graph representing the average genome sizes of Viridiplantae and Non-Viridiplantae

Table 2 gives a summary of the total number of genomes categorized based on genome sizes of the two classes of chloroplast. It has been observed that majority of the genome sizes lie between 10kb to 500kb, only two genomes namely *Floydiella terrestris* chloroplast (NC_014346) and *Paulinella chromatophora* chromatophore (NC_011087) are found to be greater than 500kb. On the other hand, 311 plants of Viridiplantae show genome sizes between 100kb and 500kb.

Table 2 : Chloroplast Genome Sizes and their classification based on different size ranges

Size Range	No. of plants
>= 10 Kb and <50 Kb	
Non- Viridiplantae	5
Viridiplantae	2
>= 50 Kb and <100 Kb	
Non- Viridiplantae	10
Viridiplantae	9
>= 100 Kb and <500 Kb	
Non- Viridiplantae	31

Viridiplantae	311
>= 500 Kb and < 1Mb	
Viridiplantae	1
>1Mb	
Non- Viridiplantae	1

b) Distribution of Microsatellites

Microsatellites in or near genes (coding regions) are found to impact protein formation and gene regulation. When the distribution of microsatellites has been analyzed overall, it is found that around 57% of microsatellite repeats fall in coding regions of all chloroplast genomes. Out of the total 78536 chloroplast microsatellites, 45518 microsatellites fall in gene regions where as the rest 33018 repeats fall in non-coding regions. However, it is surprising to see that the distribution differs when the two classes have been compared separately (Refer Fig.2).

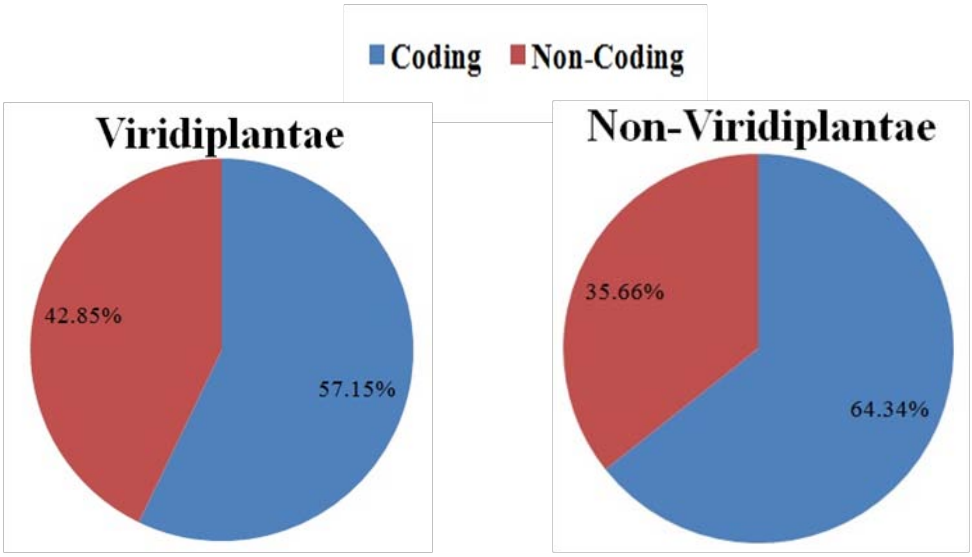

Figure 2 : Distribution of Microsatellite Repeats in Coding and Non-coding regions of Viridiplantae, Non-Viridiplantae

Genomes of Non-Viridiplantae are found to be having majority of its microsatellites in coding regions (64%). On the other hand, green plants (Viridiplantae) show that around 57% of their microsatellites to be distributed in coding regions. When two chloroplast categories are compared (Refer Fig. 3), these two categories exhibit a similar distribution of its microsatellites in coding and non coding regions.

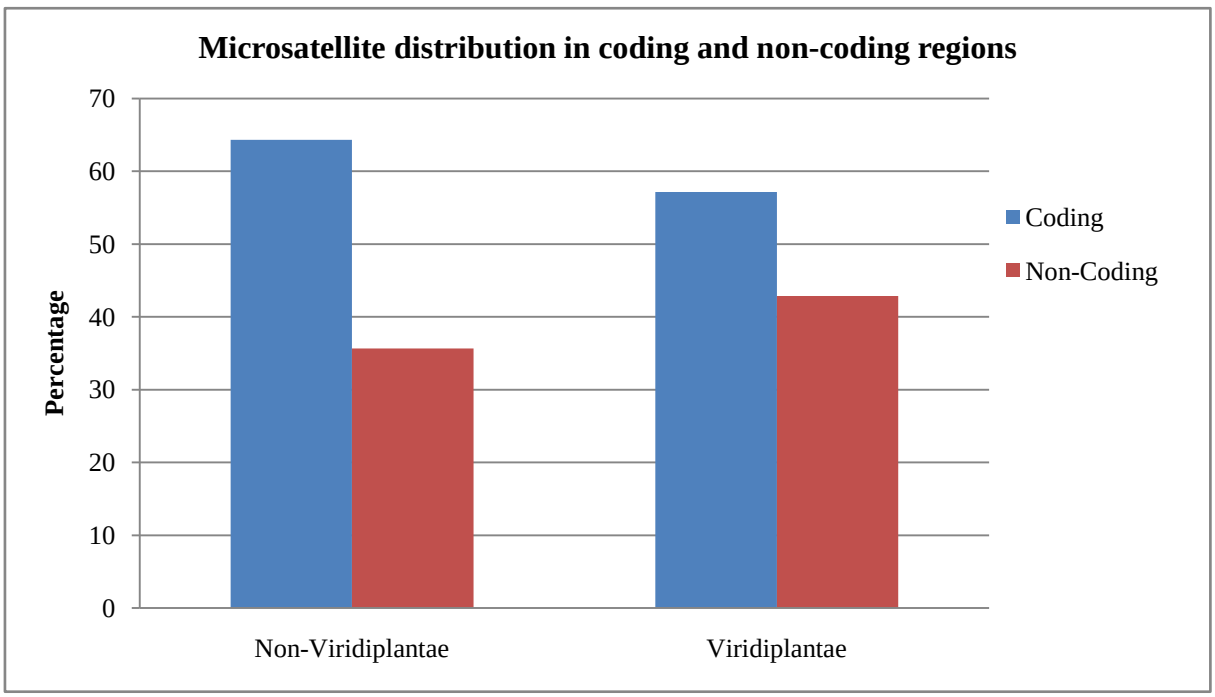

Figure 3 : Distribution of Microsatellite Repeats in Coding and Non-coding for all chloroplast Categories

It would be interesting to study the reason behind the major number of microsatellite repeats in Viridiplantae.

c) Motif-size wise Analysis

We have further analyzed the distribution of chloroplast microsatellites based on their motif sizes. Table 3 lists the proportionate distribution of chloroplast microsatellites motif-size wise. It has been observed that chloroplast genomes are rich in tri and tetra nucleotide repeats which tohether account for more than 77% in Non-virdiplantae, and around 62% in Virdiplantae. Mono, Penta and Hexa-nucleotide repeats are found to be very low in number.

Table 3 : Motif-size wise distribution of Microsatellites in chloroplast Genomes of Non-Viridiplantae and Viridiplantae

Motif Size	Non-Viridiplantae	Viridiplantae
Mono	159(1.80%)	8602(12.33%)
Di	840(9.55%)	7909(11.34%)
Tri	3506(39.87%)	17055(24.45%)
Tetra	3300(37.52)	26796(38.42%)
Penta	623(7.08%)	5680(8.14%)
Hexa	365(4.15%)	3701(5.31%)
Total	8793	69743

When the microsatellite tract lengths have been analyzed, the genomes reported few interesting tract lengths for almost all motif sizes. The average microsatellite tract lengths are usually observed to be not more than 19 bp. But, it is surprising to note that some of the tetra and tri repeats have shown exceptional tract lengths as large as 276bp have been observed.

Table 4 : Motif-size wise report Microsatellite Tract Lengths (High, Low and Average) in chloroplast Genomes of Non-Virdiplantae and Viridiplantae

	Non-Viridiplantae			Viridiplantae		
Motif Size	High	Low	Avg	High	Low	Avg
MONO	25	12	13.93	46	12	14.49
DI	54	11	12.90	83	11	13.24
TRI	51	11	12.19	276	11	12.38
TETRA	29	11	11.91	203	11	12.13
PENTA	65	14	15.27	100	14	15.41
HEXA	42	17	18.74	145	17	19.70

Based on the results in Table 4, we have further tried to find repeats in chloroplast genomes that have exceptional tract lengths. Interestingly, we found 10 repeats in chloroplast with tract lengths 100bp or more; out of those, two repeats have tract lengths 200bp or more. Two significant tract lengths of 276 and 203 have been reported for genomes with IDs NC_020321, NC_008117 respectively.

IV. Conclusion

In this paper, we have presented a brief description about the distribution of microsatellite repeats in all sequenced chloroplast genomes of Plants. This study forms the first comprehensive analysis of microsatellite repeats in chloroplast genomes and the statistics of this study can be a useful resource for biologists.

References Références Referencias

1. Palmer, Jeffrey D., and William F. Thompson. "Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost." *Cell* 29, no. 2 (1982): 537-550.
2. Bryan, G. J., J. McNicoll, G. Ramsay, R. C. Meyer, and W. S. De Jong. "Polymorphic simple sequence repeat markers in chloroplast genomes of Solanaceous plants." *Theoretical and Applied Genetics* 99, no. 5 (1999): 859-867.
3. Powell, W., M. Morgante, R. McDevitt, G. G. Vendramin, and J. A. Rafalski. "Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines." *Proceedings of the National Academy of Sciences* 92, no. 17 (1995): 7759-7763.
4. Reith, Michael, and Janet Munholland. "Complete nucleotide sequence of thePorphyra purpurea chloroplast genome." *Plant Molecular Biology Reporter*13, no. 4 (1995): 333-335.
5. Shaw, Joey, Edgar B. Lickey, Edward E. Schilling, and Randall L. Small. "Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in

angiosperms: the tortoise and the hare III."*American journal of botany* 94, no. 3 (2007): 275-288.
6. Cosner, Mary E., Robert K. Jansen, Jeffrey D. Palmer, and Stephen R. Downie. "The highly rearranged chloroplast genome of Trachelium caeruleum (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families."*Current genetics* 31, no. 5 (1997): 419-429.
7. Cronn, Richard, Aaron Liston, Matthew Parks, David S. Gernandt, Rongkun Shen, and Todd Mockler. "Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology." *Nucleic acids research* 36, no. 19 (2008): e122-e122.
8. Schlotterer,C 2000 *Evolutionary dynamics of microsatellite DNA. Chromosoma;109:365-371.*
9. *Tautz, D. and* Schlotterer, C. 1994 *Simple sequences. Curr. Opin. Genet. Dev. 4:832-837.*
10. Thibodeau, S. N., Bren, G., and Schaid, D. 1993. Microsatellite instability in cancer of the proximal colon.*Science*,*260* (5109), 816-819.
11. Goldstein, D. B., and Schlotterer, C. 2001. Microsatellites: evolution and applications. Oxford: Oxford University Press;
12. Li, Y. C., Korol, A. B., Fahima, T., and Nevo, E. 2004. Microsatellites within genes: structure, function, and evolution. *Molecular biology and evolution*, *21*(6), 991-1007.
13. 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.
14. Rajendrakumar, P., Biswal, A.K., Balachandran, S.M. and Sundaram, R.M. 2008, In silico analysis of microsatellites in organellar genomes of major cereals for understanding their phylogenetic relationships. *In Silico Biol*., 8, 87-104.
15. Kuntal, H., Sharma, V. and Daniell, H., 2012, Microsatellite analysis in organelle genomes of Chlorophyta. *Bioinformation*, 8, 255-59.
16. Kassai-Jáger, E., Ortutay, C., Tóth, G., Vellai, T., and Gáspári, Z. 2008. Distribution and evolution of short tandem repeats in closely related bacterial genomes. *Gene*, *410* (1), 18-25.
17. Sablok, G., Mudunuri, S. B., Patnana, S., Popova, M., Fares, M. A., and La Porta, N. 2013. ChloroMitoSSRDB: open source repository of perfect and imperfect repeats in organelle genomes for evolutionary genomics. *DNA research*, *20*(2), 127-133.