Available online at http://arjournal.org

APPLIED RESEARCH JOURNAL

RESEARCH ARTICLE

ISSN: 2423-4796

Applied Research Journal *Vol.1, Issue, 2, pp.51-54, April, 2015*

MINING OF SIMPLE SEQUENCE REPEATS IN CHLOROPLAST GENOME SEQUENCE OF COCOS NUCIFERA

Deepika Srivastava and *Asheesh Shanker

Department of Bioscience and Biotechnology, Banasthali University, Rajasthan, India.

ARTICLE INFO

Article History:

Received: 17, March, 2015
Final Accepted: 17, April, 2015
Published Online: 25, April, 2015

Key words:

Chloroplast, Data Mining, Simple sequence repeats, Coconut.

ABSTRACT

Simple sequence repeats (SSRs), also known as microsatellites, are found in DNA sequences and consist of short repeating motifs of 1-6 nucleotides. These repeats are ubiquitous and play an important role in the development of molecular markers. Therefore, the present analysis was conducted to mine SSRs in chloroplast genome of Cocos nucifera. A total of 44 SSRs were identified in 154.731 kb sequence mined with an average length of 18.57 bp. Depending upon the repeat unit SSRs varied in length from 12 to 132 bp. The identified SSRs showed a density of 1SSR/3.52 kb. Mononucleotides (16, 36.36%) were found to be the most abundant repeat, followed by di and tetra nucleotide repeats with same frequency (9, 20.45%), tri and penta nucleotide repeats (4, 9.09%) and hexa nucleotide repeats (2, 4.55%). Compound SSRs were completely absent in the chloroplast genome of Cocos nucifera. Most of the perfect SSRs were confined in the non-coding region of the chloroplast genome.

1. INTRODUCTION

The Coconut (*Cocos nucifera*) is a member of the family Arecaceae (palm family) and is known for its great versatility in many traditional, culinary and commercial uses. Simple sequence repeats (SSRs) or microsatellites are tandem repeated motifs which consist of 1-6 nucleotides and are located in all prokaryotic and eukaryotic genomes [1]. Studies suggest that both coding and non-coding regions of DNA sequences contain SSRs [2] [3]. SSRs are considered to be one of the molecular markers of choice because of their high abundance within the genome [4]. Considering excessive time consumption and cost, generating SSR markers from genomic libraries have been replaced rapidly by *in silico* mining of SSRs from DNA sequences available in biological databases [3]. Recently, computational mining has been used to develop SSR specific databases including ChloroSSRdb [5] and MitoSatPlant [6]. Chloroplast is an organelle found in plant cells and performs photosynthesis [7].The availability of complete organelle genome sequences provides opportunity to analyze and compare these genomes [8]. Moreover, chloroplasts SSRs (cpSSRs) are more effective indicator of population genetic structure than nuclear SSR markers [9]. Despite being such important crop a detailed cpSSRs analysis of *Cocos nucifera* is not available unlike other monocots viz. *Oryza sativa* [10], *Saccharum* [11].

Therefore, this study was designed to mine SSRs and estimate their occurrence and distribution in the chloroplast genome sequences of *Cocos nucifera*.

2. MATERIAL AND METHODS

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

2.1. Chloroplast genome sequence retrieval and data mining

The complete organelle genome sequences of angiosperms are available at National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov). The chloroplast genome sequence of *Cocos nucifera* (NC_022417) was downloaded from NCBI in FASTA and GenBank format.

2.2. Mining of simple sequence repeats

MISA, a perl script (http://pgrc.ipk-gatersleben.de/misa), was used for the detection of SSRs. The minimum repeat size was set to be >=12-mono, >=6-di, >=4-tri, >=3-tetra, penta and hexa nucleotide repeats, respectively. Both perfect (where each repeat follows the next without interruptions) and compound (where two or more repeat units are adjacent to each other) SSRs were detected through MISA. The maximum difference between two SSRs was kept 0.

2.3. Analysis of mined cpSSRs

Data generated after SSR mining was analyzed for frequency & distribution of SSRs in coding and non-coding regions of cpDNA. The information about coding and non-coding regions was taken from GenBank files. SSRs were classified as coding and non-coding on the basis of their presence in coding and non-coding regions [5].

3. RESULT AND DISCUSSION

The present analysis deals with the identification of chloroplastic simple sequence repeats (cpSSRs) in *Cocos nucifera.* The distribution of mined cpSSRs is presented in figure 1.

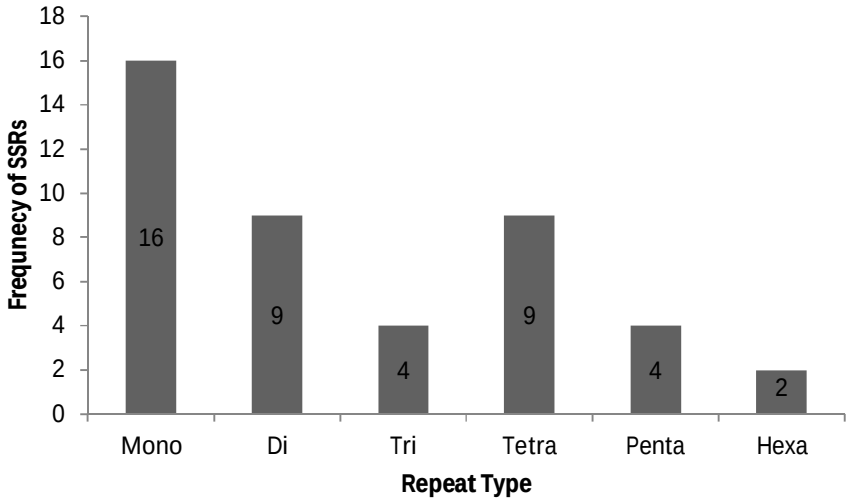

Figure 1 Frequency distribution of various repeat types

A total of 44 SSRs were identified in 154.731 kb sequence mined with an average length of 18.57 bp. Mononucleotides (16, 36.36%) were found to be the most abundant repeat, followed by di and tetra nucleotide repeats with same frequency (9, 20.45%), tri and penta nucleotide repeats (4, 9.09%) and hexa nucleotide repeats (2, 4.55%). Compound SSRs were completely absent in the chloroplast genome of *Cocos nucifera*. Among mononucleotide repeats presence of only A/T motifs showed consistency with SSR analysis of other organelle genomes [10] [11].

The density of SSR in chloroplast genome of *Cocos nucifera* was 1SSR/3.52kb. It found to be higher than the density of EST-SSRs in barley, maize, wheat, rye, sorghum and rice (1 SSR/6.0 kb) [12], cotton and poplar (1 SSR/20 kb and 1 SSR/14 kb respectively) [4] and Unigene sequences of *Citrus* (1 SSR/12.9 kb) [13]. Moreover, the density of SSRs in *Cocos* was higher when compared to the cpSSRs of rice (1SSR/6.5 kb) [10], however, lower than the cpSSRs density in family Solanaceae (1 SSR/1.26kb) [14].The variation in SSR density might be due to different parameters including minimum length of SSRs taken, the amount of data analyzed and genomic composition of the sequence mined. The identified 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 the majority of SSRs found in non-coding region of the chloroplast genome. This nonrandom distribution of cpSSRs towards non-coding regions showed consistency with earlier studies Solanaceae [15], Asteraceae [16], Fabaceae [17] and *Saccharum* [11].

Table 1 Identified SSRs motif, their length and start-end position in chloroplast genome of *Cocos nucifera*

S. No.	MOTIF	LENGTH	START	END	REGION
1.	(CTT)4	12	1701	1712	Coding
2.	(AAT)5	15	3807	3821	Non-coding
3.	(TCTA)4	16	5931	5946	Non-coding
4.	(AT)6	12	8480	8491	Non-coding
5.	(A)12	12	9066	9077	Non-coding
6.	(T)13	13	14267	14279	Non-coding
7.	(AT)6	12	14638	14649	Non-coding
8.	(AT)7	14	14663	14676	Non-coding
9.	(T)13	13	23073	23085	Non-coding
10.	(T)18	18	23402	23419	Non-coding
11.	(A)12	12	28277	28288	Non-coding
12.	(A)14	14	28787	28800	Non-coding
13.	(TATTT)3	15	40566	40580	Non-coding
14.	(TCTT)3	12	43979	43990	Non-coding
15.	(TA)6	12	45161	45172	Non-coding
16.	(TA)6	12	45324	45335	Non-coding
17.	(ATTT)3	12	45441	45452	Non-coding
18.	(A)13	13	45609	45621	Non-coding
19.	(AT)6	12	46272	46283	Non-coding
20.	(AT)6	12	47215	47226	Non-coding
21.	(A)12	12	49713	49724	Non-coding
22.	(A)13	13	60425	60437	Coding
23.	(AATG)3	12	61104	61115	Coding
24.	(TTTCA)3	15	64464	64478	Non-coding
25.	(A)12	12	66453	66464	Non-coding
26.	(T)14	14	70209	70222	Non-coding
27.	(T)13	13	70477	70489	Non-coding
28.	(ATAA)3	12	70678	70689	Coding
29.	(T)12	12	77655	77666	Non-coding
30.	(T)13	13	80409	80421	Non-coding
31.	(ATA)4	12	81541	81552	Non-coding
32.	(TTTA)3	12	82162	82173	Non-coding
33.	(T)15	15	82511	82525	Non-coding
34.	(TTTTA)3	15	83901	83915	Non-coding
35.	(TATTAG)22	132	99868	99999	Non-coding
36.	(AT)6	12	112945	112956	Non-coding
37.	(AATA)3	12	115456	115467	Coding
38.	(TA)9	18	117437	117454	Non-coding
39.	(A)15	15	117490	117504	Non-coding
40.	(TTTA)3	12	118211	118222	Non-coding
41.	(ATTC)3	12	120712	120723	Non-coding
42.	(TATAC)3	15	123848	123862	Non-coding
43.	(ATA)4	12	125833	125844	Coding
44.	(ACTAAT)22	132	138686	138817	Non-coding

4. CONCLUSION

The identified SSRs will be useful for the development of SSR markers, in genetic diversity studies and to reveal variation in genomes. Moreover, the study provides scientific base for phylogenetics, evolutionary genetics studies on different Arecaceae species in future.

5. ACKNOWLEDGEMENT

Financial support from DBT Center for Bioinformatics, Dept. of Bioscience and Biotechnology, Banasthali Vidyapith to DS and UGC-MRP (F.No. 42-138/2013) to AS is also acknowledged.

6. REFERENCES

[1] Zane. L., Bargelloni, L. And Patarnello, T. 2002. Strategies for microsatellite isolation: a review. *Mol. Ecol.* 11: 1-16.

[2] Katti, M.V., Ranjekar, P.K. and Gupta, V.S. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18: 1161-1167.
[3] 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.
[4] 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.
[5] Kapil. A., Rai, P.K. and Shanker, A. 2014. ChloroSSRdb: a repository of perfect and imperfect chloroplastic simple sequence repeats (cpSSRs) of green plants. Database: The Journal of Biological Databases and Curation doi:10.1093/database/bau107.
[6] Kumar, M., Kapil, A. and Shanker, A. 2014. MitoSatPlant: Mitochondrial microsatellites database of viridiplantae. Mitochondrion 19: 334-337.
[7] Campbell. N., Williamson, B. and Heyden, R.J. 2006. Biology: exploring Life. Boston, Massachusetts: Pearson Prentice Hall; ISBN 978-0-13-250882-7.
[8] Shanker. A., Sharma. V and Daniell, H. 2009. A novel index to identify unbiased conservation between proteomes. IJIB. 7:32-38.
[9] Birky, C.W. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proc. Nat. Acad. Sci. U.S.A. 92: 11331-11338.
[10] 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.
[11] Melotto-Passarin, D.M., Tambarussi, E.V., Dressano, K., de Martin, V.F. and Carrer, H. 2011. Characterization of chloroplast DNA microsatellites from *Saccharum* spp. and related species. Genet. Mol. Res. 10: 2024-2033.
[12] 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.
[13] 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. Hortic. 113:353-361.
[14] 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.
[15] Daniell, H., Lee, S.B., Grevich, J., Saski, C., Quesada-Vargas, T., Guda, C., Tomkins, J. and Jansen, R.K. 2006. Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. Theor. Appl. Genet. 112: 1503-1518.
[16] Timme, R., Kuehl, E.J., Boore, J.L. and Jansen, R.K. 2007. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. Am. J. Bot. 94: 302-312.
[17] Saski, C., Lee S.B., Daniell, H., Wood, T.C., Tomkins, J., Kim, H.G. and Jansen, R.K. 2005. Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Mol. Biol. 59: 309-322.