Sequencing and Bioinformatics Analysis of an Intergenic Region of Chloroplast Genome: A Step towards Development of Potential DNA Barcodes for Angiosperms

Tammineni Rajitha[1], Tammineni Ramya[2], Aparajita Mohanty[3] and Jasmeet Kaur Abat[4]

[1,2,3,4]Department of Botany, Gargi College, University of Delhi
E-mail: [1]tammineni.rajita@gmail.com, [2]tramya030@gmail.com,
[3]aparajita.gargi@gmail.com, [4]jasmeet29@yahoo.com

Abstract—*DNA barcoding is a molecular technique that uses variations in DNA sequences to identify biological specimens. Unlike animals, in plants DNA barcoding is not simple. The plastid loci like rbcL, matK and trnH, are used for identification purposes but sequences of these loci are unable to provide 100% species identification across the plant kingdom. Also, difficulty in amplification and sequencing restricts their suitability as universal barcodes. Therefore, search for additional and universal barcodes for plant kingdom is imperative to further facilitate identification across taxa. Keeping this in view, the objective of the present investigation was to assess chloroplast genomes for DNA barcodes that can supplement and/or complement the existing barcodes. Based on preliminary in silico sequence analysis of chloroplast genomes of angiosperms, trnD–trnT non-coding spacer region (DT) of chloroplast genome was chosen to experimentally validate its suitability as DNA barcode. For the present study, five individuals from 18 angiosperm species belonging to seven different families were selected. PCR products were sequenced and analysed for variations by multiple sequence alignment tool (CLUSTALW). All genera and/or species belonging to seven angiosperm families could be discriminated at the DNA sequence level. The present study proposes that besides matK and rbcL sequences (which are genic regions) intergenic region (i.e. trnD-trnT) can also be used as DNA barcode in plants. Our results instigate further investigation on other intergenic regions of chloroplast genome, for their potential as barcodes in plants.*

1. INTRODUCTION

Identifying organisms has become important as we monitor the biological effects of global climate change and attempt to preserve species diversity in spite of accelerating habitat destruction. DNA barcoding is a genomic approach to taxon diagnosis which exploits diversity among DNA sequences to identify organisms. It involves using short standardized stretches of DNA as a tool for species identification [1]. DNA sequence is the order of nucleotides in a DNA molecule. This order changes due to point mutations which accumulate over the course of evolution. When DNA sequences of a particular region from different species are aligned together they show certain similarities and differences in the arrangement of base pairs. These similarities and differences (created due different order of base pairs) help in determining the phylogenetic relationship between the given species and species identification.

DNA barcoding as a uniform and practical method for species identification has broad scientific applications. From conservation biology, biodiversity surveys to identification of juveniles and cryptic species [2]. It has a very legitimate use in forensics as well as in taxonomy. It has even been referred to as a potential method for DNA taxonomy. DNA barcodes of species level are routinely obtained using mitochondrial COX I (Cytochrome c oxidase subunit 1) gene for almost all phyla of animal kingdom [3]. COX I shows high rate of sequence change at species level in animal kingdom which has enabled it to become the core of global bio-identification system for animals [3]. It has also shown remarkable promise as a barcode in various other groups such as red algae [4] and fungal communities like Pezizomycotina and Agaricomycetes [5].

In case of plant kingdom although COX I gene proved less than desirable as a potential barcode sequence due to slow nucleotide substitution rate in it and in other mitochondrial genes of plant cell. Also no such analogous sequence has been standardized for land plants. The ideal gene target for DNA barcoding needs to be sufficiently conserved to be amplified with universal primers yet divergent enough to resolve closely related species [2]. The Consortium for the Barcode of Life (CBOL) has proposed that locus selected should have Universality (i.e. the marker should be routinely sequenced across the land plants), Sequence quality (good quality sequences should be obtained) and Resolving power (the new

marker should have the capacity to show interspecific divergences between species). For creating DNA barcodes for plant kingdom chloroplast genome sequence data was brought into application. Chloroplast genome is as required highly conserved and stable in structure but shows enough diversification in the DNA sequence from one species to another i.e. it consists of regions with very high evolutionary rates which can be utilized for DNA barcoding. Universal primers are available for a number of loci and intergenic spacers that are evolving at a variety of rates. The plastid locus most commonly sequenced by most plant taxonomists for phylogenetic purposes is rbcL, followed by the trnL-F intergenic spacer, matK, ndhF, and atpB. Also a few multiple loci combinations were suggested like rpoc1+matK+trnH-psbA, matK+atpF-H+trnH-psbA, rbcl+trnH-psbA [6] and rbcL+matK. Out of all the proposed sequences, based on assessment of recoverability, sequence quality, and levels of species discrimination, the two-loci combination of rbcL+matK was recommended as the standard plant barcode. But any of the recommended sequences are unable to provide 100% species identification across the plant kingdom and thus are unlikely to act as a universal barcode [7]. To solve phylogenetic problems at the species level, or to identify species using DNA sequences, we need to identify regions with very high evolutionary rates. Greater availability of such regions will increase our ability to resolve such identification problems.

Therefore, the objective of our current study is to assess the use of trnD-trnT (DT), a non-coding spacer region of chloroplast genome to test whether it can be used as DNA barcode in angiosperms. This intergenic region of chloroplast genome was chosen as universal primers of this region are already known [8]. Also, the size of the amplicon is not too large (800 bp to 1600 bp) as shown by our preliminary in silico analysis using sequence information of data deposited in NCBI database.

2. MATERIALS AND METHODS

2.1 Plant Sample Selection

A taxon based approach of selecting plants belonging to same taxa and a species-pair approach of selecting different species but same genus was taken to select the plants for barcoding. The taxon based approach will help in determining the universality of the sequence while the species-pair approach will be helpful in establishing the uniqueness of the sequence. For the present study, 18 angiosperm species growing in the Botanical garden of Gargi College were selected (Table 1). These belonged to seven different families Lamiaceae, Amaranthaceae, Euphorbiaceae, Cannabaceae, Solanaceae, Fabaceae and Asparagaceae. Fresh and young leaf samples of different species of angiosperms were collected and used for DNA extraction.

2.2 Total DNA extraction

CTAB (Hexadecyl trimethyl-ammonium bromide) method was used for extraction of total DNA. Fresh plant tissue (0.5 g) was grounded in 2 ml CTAB buffer (2% CTAB, 0.5M EDTA, 5M NaCl, 1% PVP, 1M Tris HCl pH-8) containing 20µl of β-mercaptoethanol. About 0.5ml of the slurry was transferred into a micro-centrifuge tube (1.5 ml) containing 100 µl of chloroform: isoamyl alcohol (v/v 24:1). The extract was then incubated at 65°C for 1 hour in water bath. After incubation, the micro-centrifuge tube was filled with chloroform and isoamyl alcohol mixture (v/v 24:1) and was centrifuged for 10 min at 7,000 rpm. The supernatant was transferred to a clean micro-centrifuge and double the amount of chilled ethanol was added to precipitate the DNA. The DNA was pelleted by centrifugation at 5000 rpm at 4°C for 5min. The DNA pellet was washed in 75µl of 70% ethanol + 0.2 M sodium acetate mix, 70% ethanol + 10mM ammonium acetate and 70% ethanol. The pellet was then dried and dissolved in 200 µl of TE buffer and stored at -20°C until use. For qualitative check, extracted DNA was resolved on agarose gel electrophoresis (AGE).

Qualitative estimation of DNA by Agarose Gel Electrophoresis

To check the quality of the extracted genomic DNA, it was electrophoreses on 1% agarose gel. The cleaned, dried gel tray was sealed with cello-tape. The comb was positioned 1mm away from the edge of the tray to form wells for sample loading. Gel was prepared by melting 1g of agarose in 100ml of TBE buffer in a microwave. The mixture was allowed to cool down to 50°C before adding ethidium bromide. The mixture was poured onto gel tray and was left undisturbed to solidify. The cellotape was removed and the gel was placed in running buffer (1X TBE). DNA (1µl) was mixed with sample buffer (2µl) and was loaded along with 1 Kb ladder. The gel was run at 40 volts. DNA was visualized by placing the gel on a UV transilluminator. The gel was examined under UV light using gel documentation system.

PCR reaction

The trnD-trnT region of the DNA molecule was amplified using PCR reaction. Primers were synthesized according to of Demesure et al. [8]. PCR reaction (30 µl) was set up with 28 µl of master mix (23 µl water, 3.0 µl 10x Taq Buffer, 0.5 µl dNTP mix, 0.5 µl trnD primer, 0.5 µl trnT primer, 0.5 µl Taq polymerase enzyme) and 2 µl of DNA dissolved in 1XTE in each PCR tube. The components were mixed and centrifuged. The temperature was set at 94°C for 2 min for initial denaturation then again 94°C for 45 sec for denaturation. Annealing temperature was varied between 50-55°C for different species. For extension the temperature was set at 72°C for 3 min. The reaction was run for 30 cycles. The

temperature for final extension was set at 72 °C for 10 min. The PCR product was then run on a gel to check size and concentration.

Sequencing and Bioinformatics analysis

The PCR products were sequenced (XCELRICS LAB) and the obtained sequences were further analyzed using Pair wise or Multiple Sequence Alignment (ClustalW) and Phylogenetic tree diagrams to establish variability.

3. RESULTS

Total DNA was extracted from all the 18 plant species collected from Botanical Garden of Gargi College. The extracted DNA was used for PCR amplification of the DT region of chloroplast genome. PCR amplifications succeeded in all the plant samples (Fig. 1). Size of the amplicon ranged from 600 bp to 1200 bp (Table 1). The amplicons were sequenced bi-directionally and then the sequences obtained were subjected to bioinformatics analysis. For some genera like Ocimum differences were seen on the gel in the size of the amplicon. For *Ocimum basilicum* amplicon size of 1Kb was obtained while for *Ocimum tenuiflorum* amplicon size of 1.2 Kb was obtained using same set of primers.

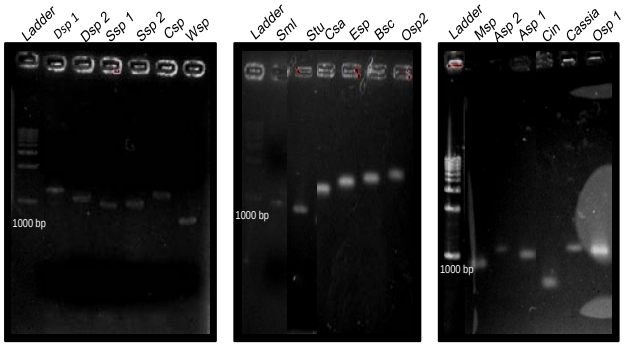

Fig. 1: PCR amplification of the trnD-trnT region from different plant samples

Table 1: PCR amplifications (Fig. 1) were sequenced from both the directions.

S. no	Specimen	size	Sequence length (bp)	
			F' primer	R' primer
1.	Ocimum basilicum	1000	928	957
2.	Ocimum tenuiflorum	1200	944	964
3.	Mentha sp	900	838	835
4.	Clerodendrum inerme	600	563	581
5.	Bassia scoporia	1200	990	1000
6.	Spinacia oleracea	1000	925	997
7.	Spinacia sp.	1000	947	1000
8.	Euphorbia cyathophora	1200	1000	1000
9.	Cannabis sativa	1100	1000	1000
10.	Datura inoxia	1100	1000	1000
11.	Datura metel	1000	1000	990
12.	Solanum tuberosum	1000	967	1000
13.	Solanum melongena	1000	909	993
14.	Withania somnifera	900	850	800
15.	Capsicum sp	1100	1000	1000
16.	Cassia senna	1200	772	791
17.	Asparagus setaceus	1000	976	988
18.	Asparagus racemosus	1100	1000	934

To analyze potential of DT region as barcode, phylogenetic tree was constructed using sequences obtained from the plant species. Phylogenetic analysis revealed that the amplified sequence of trnD-trnT region was able to discriminate different families which were used in the present study (Fig. 2). The phylogenetic tree constructed clearly grouped members belonging to same families, depicting the power of trnD-trnT to discriminate between different families. To find out differences at the species level, pair wise sequence alignment was performed using MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle).

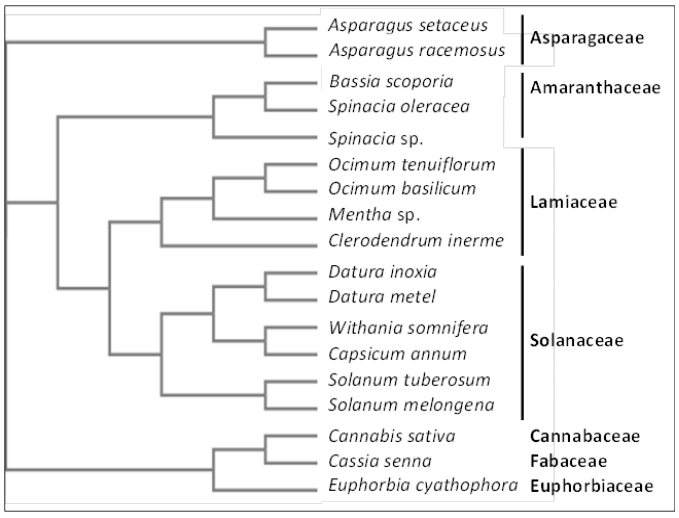

Fig. 2: Phylogenetic tree constructed using sequences obtained after sequencing the amplicons of all the plant samples

Ocimum basilicum and *Ocimum tenuiflorum* sequences showed 96% similarity with some differences which account for species level differences. Alignment clearly depicts that differences in the DT region in these species is due to indels (i.e. insertion or the deletion of bases in the DNA of an organism) and point mutations at certain places responsible for creating Single Nucleotide Polymorphism (SNP's). Similar percentage of similarity >94% was obtained during pair wise sequence alignments between same species belonging to other genera like *Datura*, *Asparagus* and *Spinacia*.

The DT sequences obtained were subjected to similarity analysis BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) against non-redundant database to find similar sequences. Our sequences showed similarity to other trnD-trnT sequences of plants belonging to same genera and of same family. Of these,

sequences of same genera were subjected to multiple sequence alignment (msa) using clustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/) to find out regions of similarity and differences. Multiple sequence alignment of *Ocimum* is shown in Fig. 3. In BLAST analysis trnD-trnT sequence of *O. tenuiflorum* showed similarity with sequences of *O. gratissimum*, *O. basilicum* and *O. selloi* along with other genera of same family. Out of these all the four sequences of *Ocimum* genera were used for msa, which revealed > 92% similarity but had differences also in terms of insertions and deletions. Similar analysis was done with other sequences also.

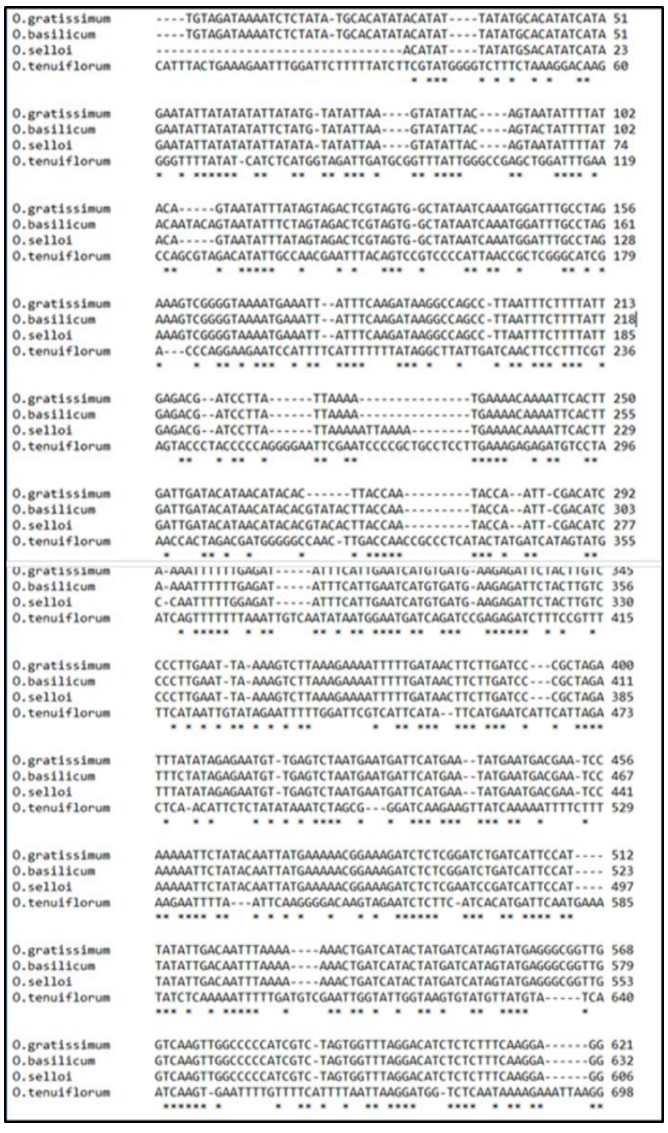

Fig. 3: Multiple sequence alignment. The nucleotide sequence of trnD-trnT of *O. tenuiflorum* was aligned with trnD-trnT sequence of *O. gratissimum*, *O. basilicum* and *O. selloi*. Multiple alignment was performed using CLUSTALW. Symbols at the bottom of each column denote degree of conservation observed in each column. '*' means residues in each column are identical.

4. DISCUSSION

An ideal DNA barcode should be retrievable with a single primer pair, be responsive to bidirectional sequencing and provide maximal discrimination among species. From the results obtained in the present study it is evident that primers of trnD-trnT can be used as universal primers as PCR amplification was attained in all the plant specimen belonging to seven different families Lamiaceae, Amaranthaceae, Euphorbiaceae, Cannabaceae, Solanaceae, Fabaceae and Asparagaceae. Our results agree with those of Demesure et al. [8].

For genera (like *Ocimum*) differences were seen on the gel in the size of the amplicon obtained from different species. These results suggest that for such species a quick PCR gel based approach can be used instead of sequencing for species identification. However further detailed study is required involving more number of species for a clear picture. In addition, we also propose based on the present results that trnD-trnT can be used for DNA barcoding upto the species level because of its ability to discriminate between species of same genus. This is proposed not only on the basis of our amplification and sequencing results, but also by comparing our sequences with the other trnD-trnT sequences available in the NCBI database. The variations (indels and SNP's) are sufficient in case of Ocimum (Lamiaceae), Solanum (Solanaceae), Datura (Solanaceae), Asparagus (Asparagaceae) and Spinacia (Amaranthaceae) to show interspecific variations and thus discriminate the species.

Further, we infer that for specificity of amplification, both forward and reverse primers must be used. However, sequencing information available by using only forward primer is sufficient to be used as bar code in case of trnD-trnT. The novelty of the present study is proposition that besides matK and rbcL sequences (which are genic regions), intergenic region (i.e. trnD-trnT) can also be used as DNA barcode in plants. Our results instigate further investigation on similar intergenic regions of chloroplast genome to be tested as potential barcodes in plants.

5. ACKNOWLEDGEMENTS

The research work was supported under Star College Scheme and Bioinformatics Infrastructure Facility of Department of Biotechnology.

REFERENCES

[1] Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R., "Biological identifications through DNA barcodes", *Proceedings of the Royal Society B: Biological Sciences*, 270, 2003, pp. 313-21.

[2] Stoeckle, M., "Taxonomy, DNA and Bar Code of Life". *Bioscience*, 53, 2003, pp. 2-3.

[3] Hebert, P.D.N., Ratnasingham, S. and deWaard, J.R., "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species", *Proceedings of the Royal Society B: Biological Sciences*, August 2003, 270 Suppl 1:S96-9.

[4] Robba, L., Russell, S.J., Barker, G.L. and Brodie, J., "Assessing the use of the mitochondrial cox1 marker for use in DNA barcoding of red algae (Rhodophyta)", *American Journal of Botany*, August 2006, 93, 8, pp.1101-8.

[5] Damon, C., Barroso, G., Ferandon, C., Ranger, J., Fraissinet-Tachet, L. and Marmeisse, R., "Performance of the COX1 gene as a marker for the study of metabolically active Pezizomycotina and Agaricomycetes fungal communities from the analysis of soil RNA", 74, December 2010, pp. 693-705.

[6] Hollingsworth, M.L., Andra Clark, A., Forrest, L.L., Richardson, J., Pennington, R.T., Long, D.G., Cowan, R., Chase, M.W., Gaudeul, M. and Hollingsworth, P.M., "Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants" *Molecular Ecology resources*, 9, March 2009, 2, pp. 439-57.

[7] Singh, H.K., Parveen, I., Raghuvanshi, S. and Babbar, S.B., "The loci recommended as universal barcodes for plants on the basis of floristic studies may not work with congeneric species as exemplified by DNA barcoding of Dendrobium species", *BMC Research Notes*. January 2012, 19, 5, pp. 42.

[8] Demesure, B., Sodzi, N. and Petit, R.J., "A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants", *Molecular Ecology*, 4, 1995, pp. 129-131.