Full Length Research Article

Sequence analysis of rbcL and matK regions for a comparison study of Senna species

https://doi.org/10.62940/als.v13i1.3252

Issue: Volume 13, Issue 1

Received: 20-03-2024

Revised: 01-12-2025

Accepted: 24-01-2026

Published online: 31-03-2026

Keywords: DNA barcode, matK, rbcL, Senna species, Chaos Game Representation

Patcharawarin Ruanto[1],*, Somruthai Tunma[2], Natthiya Chaichana[3]

1. Biology program, Faculty of Education, Chiang Rai Rajabhat University, Thailand
2. Chemistry program, Faculty of Education, Chiang Rai Rajabhat University, Thailand
3. Science program, Faculty of Education, Chiang Rai Rajabhat University, Thailand

* p.ruanto@gmail.com

ABSTRACT

Background: DNA barcoding is an efficient molecular biology technique that utilizes a short genetic locus with sufficient variability to enable precise organism identification. Typically, regions such as ribulose- bisphosphate carboxylase (*rbcL*) and maturase K (*matK*) in plants are widely used due to their balance between interspecific divergence and intraspecific conservation. These standardized genomic fragments are amplified by Polymerase Chain Reaction (PCR) and subsequently sequenced for comparative analysis. Every species has its own characteristic DNA barcode, which can be matched against curated reference libraries, enabling accurate identification even when morphological traits are ambiguous, damaged, or insufficient for taxonomic classification. This approach has been particularly transformative in biodiversity monitoring, ecological studies, and the detection of cryptic or invasive species. Presenting DNA barcodes in a graphical form provides an alternative and powerful way to store and display sequence information as it facilitates cross- comparison. The generated graphical outputs can be incorporated into machine learning algorithms for species recognition in further large-scale ecological and conservation research.

Methods: The *matK* and *rbcL* genes of the collected five *senna* species were selected as DNA barcodes to confirm and distinguish these plant samples. An alternative method was also presented to extract the characteristics of the DNA sequences with graphical operation based on Chaos Game Representation (CGR).

Results: It was found that the power to discriminate between species was high enough when a two- locus barcode approach was applied with 90% successful amplification using the provided protocol optimization . The similarity/ dissimilarity comparison of the collected plant samples was also achieved by the obtained CGR.

Conclusion: Grouping of individuals based on genetic relationship was consistent with morphology and taxonomy, particularly when the primers for *matK* were used, whereas the *rbcL* barcode was less effective in distinguishing species.

INTRODUCTION

The uniqueness of plants is studied through a number of methods including examination of the morphological characteristics, which is the most common way. However, plants in the same or related family can appear to have similar characteristics and essential substances [1]. Therefore, molecular biology technique is used in this study for conducting DNA barcoding studies and providing guidelines of plant identification. DNA barcoding is an effective molecular biology technique that is gaining a lot of attention. The technique uses short standardized genomic sequences to identify species through PCR amplification. It can be applied in evolutionary biology field, biodiversity, as well as conservation genetics. A short genetic locus used as DNA barcode must be highly variable enough to be able to quickly identify each organism. To effectively identify the organism, more than one location (locus) of DNA barcode is used. At present, there are several DNA loci that are agreed to be used as plant DNA barcodes, such as the internal transcribed spacer (ITS) [2], *rbcL, trnL*-F intergenic spacer[3], *matK* [4], *ndhF, atpB* [5,6]. There are a number of research studies that confirm the ability of *matK* as a global plant DNA barcode; however, the success is still limited in gymnosperms [7]. Currently, we can say that more than one location of standard DNA barcode is required in order to effectively identify the organism. CBOL (the Consortium for the Barcode of Life) has proposed using a combination of *matK* and *rbcL,* gene loci found in chloroplasts, as a DNA barcode [8]. The *matK* gene (*maturaseK* gene) is located in the intron section II within the chloroplast DNA, encoding maturase enzyme which is related to RNA splicing process [9]. The *rbcL* gene (ribulose bisphosphate carboxylase gene) is an extremely important gene as the encoded enzyme is a subunit of ribulose bisphosphate carboxylase (RUBISCO), which is used in the process of CO_2 fixation during photosynthesis. It is also popularly used as a marker in studies of evolutionary relationships.

Presenting DNA barcode in a graphical form is a technique to collect information and show accomplished results in a more beneficial way. A graphic display can be easily understood and its numerical data can be further used for DNA sequence comparison analysis. Until now, there are many formats of multidimensional sequence of DNA, each of which has its advantages and limitations. Chaos Game Representation (CGR) is a two-dimensional graphical representation of DNA sequence. It uses a simple walk model to locate coordinates in the orthogonal x y system[10]. Different biological data represented by different nucleotide sequences are taken into account in the model to conduct the result that is relevant to the chemical structures and uncomplicated to understand. The particular numerical characterization of DNA sequences is also a proper data form for the further study of similarities and dissimilarities of gene structures.

The CGR-Walk applies a nonlinear dynamical system that is based on chaos theory to CGR mapping [11]. Representation in this graphical form can be used to display patterns of gene sequences. Moreover, the algorithm allows fractal structure pictures to be viewed as continuous reference systems of which all four possible nucleotides occupy unique positions corresponding to vertices of a binary square. The binary CGR vertices are assigned to four nucleotides; A = (0, 0), G= (1, 1), C= (0, 1), T = (1, 0). The CGR coordinates are calculated iteratively by moving the plot to half the distance between the previous and the current positions of base corner. The function can be given by

$$CGR_i = CGR_{i-1} - 0.5\ (CGR_{i-1} - g_i), \quad (1)$$

Where;

$$i = 1, \ldots, n_G;\ CGR_0 = (0.5, 0.5);\ g_i \in \{A, G, C, T\}. \quad (2)$$

The given formula represents a mathematical expression that maps DNA by starting in the center (0.5,0.5) and progressing halfway toward each nucleotide's designated corner (A, T, G, and C). DNA sequences are transformed in this manner to produce a fractal-like image that encodes information about the DNA sequence's structure.

METHODS

Plant samples were collected from the Chiang Saen Lake (20°15'N., 100°2'E.), a wetland located in Yonok district, Chiang Saen, Chiang Rai province, Thailand. Genomic DNA Extraction was accomplished using Viviantis GF-1 Plant DNA Extraction Kit. The DNA fragments of the targeted

gene of the collected plants were amplified by Polymerase Chain Reaction (PCR), using primers and protocols for the *matK* and *rbcL* barcodes [12] as shown in Table 1. Then, the purified PCR fragments were sent to Macrogen Inc. (Korea) for sequencing. The sequence analysis of each sample was further performed using finchTV 1.4.0 to remove all ambiguous sequences, and MAFFT 7.0 as well as NTSYS (2.02pc) were used for multiple sequence alignment and analysis of molecular phylogeny.

Then, the DNA sequence of each sample was converted into the pattern of CGR- RY walk, illustrated by virtual graph paper application for further DNA analysis.

RESULTS

DNA amplification and sequencing

The collected 5 species of plants belonged to Leguminosae family as shown in Table 2. DNA amplification for PCR products of *matK* was successful for 5 out of 5 species, whereas the *rbcL* amplification failed to give a good quality PCR band for *Senna siamea* as shown in Table 2.

The size of PCR products at the regions of *matK* and *rbcL* were approximately 900 and 600 base pairs, respectively. The successful purified PCR samples were sent to Macrogen Inc. (Korea) for sequencing. The obtained nucleotide sequences of *matK* and *rbcL* genes were further used for multiple sequence alignment and phylogenetic analysis.

Phylogeny analysis of *matK and rbcL barcodes*

To study genetic relationship and classification of plant species, the CLUSTAL format alignment and molecular phylogeny analysis based on Unweighted Pair- Group Method with Arithmetic Mean (UPGMA) of partial *matK and rbcL* barcode loci from 5 taxa were performed using NTSYS (2.02pc) to create the dendrogram as shown in Figure 1 and 2, respectively. For the *matK* barcode, *S. tora, S. alata* and *S. siamea* were closely related, as they were clustered together in one group, then joined together with *S. occidentalis* , while *S. hirsuta* was further isolated from the rest. The result indicated that nucleotide substitution rate in *matK* region was higher than *rbcL* . The successful amplified *rbcL* sequences of 4 species were used for constructing the molecular phylogenetic tree. The result was found to be compatible with the prior *matK* dendrogram, with *S. tora* and *S. alata* being more connected to one another than *S. occidentalis,* and *S. hirsuta* being the species most distant from the others."

Two-dimensional graphical representation

The previous branching diagrams showed evolutionary relationship based on DNA sequences and allowed highly related species to cluster together. DNA sequence differences of *S. alata, S. tora* and *S. siamea* were very small (3% and 2% differences among 3 species for *matK* and *rbcL* respectively), therefore, they were grouped together and separated from *S. occidentalis* and *S. hirsuta* . Apart from a dendrogram, we proposed another method to compare clear visual similarities/dissimilarities of these 5 models of *Senna* species. In this research, the sequences of *matK* and *rbcL* barcoding regions as indicated in table 3 were used for all 5 species to create a two-dimensional (2D) graph of CGR walk that represented partial *matK* and *rbcL* sequences.

The 2D graphical techniques were used to analyze and visualize overall characteristics of DNA sequences and even for thorough consideration of specific positions. In addition, all data were obtained in order to convert the base sequences to a numerical form which still contained the same data without any elimination. The sizes and directions of vectors representing nucleotide sequences of different species that made up the graph were unique. Different genes generated various different 2D patterns. The repetitive patterns with same bases among species were also represented at the same time. However, the graphs clearly showed different leaps at positions containing different bases. The relationship between the 2D graph and DNA sequence of any species or any gene was one to one. Since N represented the length of the studied DNA, any i = 1, 2,..., N would be considered a vector with 4 possible directions depending on the sequence of the next base. As a result, DNA dissimilarity could be clearly expressed in the 2D graph even if there was only one base difference of the compared DNA as shown in figure 3.

From the 2D graphs of partial *matK* sequences of 5 senna species, it was found that the CGR

patterns of *S. hirsuta* and *S. occidentalis* were clearly different from other species, while *S. alata, S. tora* and *S. siamea* had more similar patterns. However, these three closely related species still had some different degrees of vectors as all 3 graphs had their own characteristics. For *rbcL* sequences, it was clear that *S. hirsuta* diverged the most from the remaining three species of which similarities were very high so that it was harder to differentiate *S. alata, S. tora* and *S. occidentalis* from each other using CGR graphs of *rbcL* sequences.

Figures

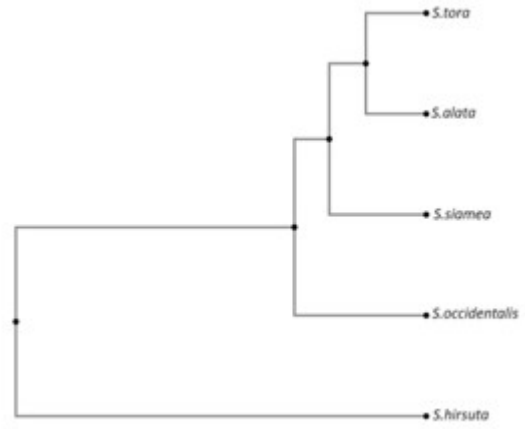

Figure 1: UPGMA tree constructed by using matK gene sequences from 5 Senna species

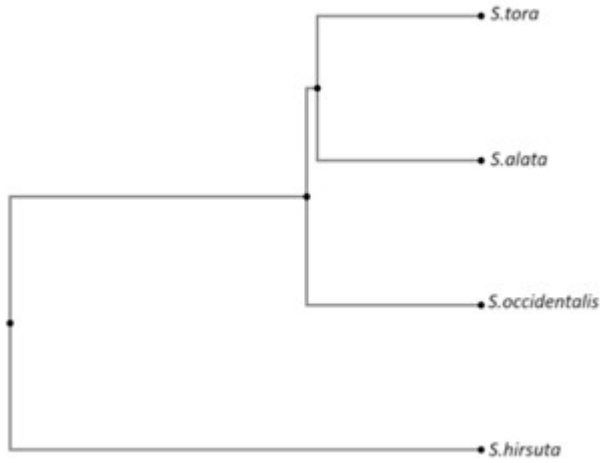

Figure 2: UPGMA tree constructed by using rbcL gene sequences from 4 Senna species

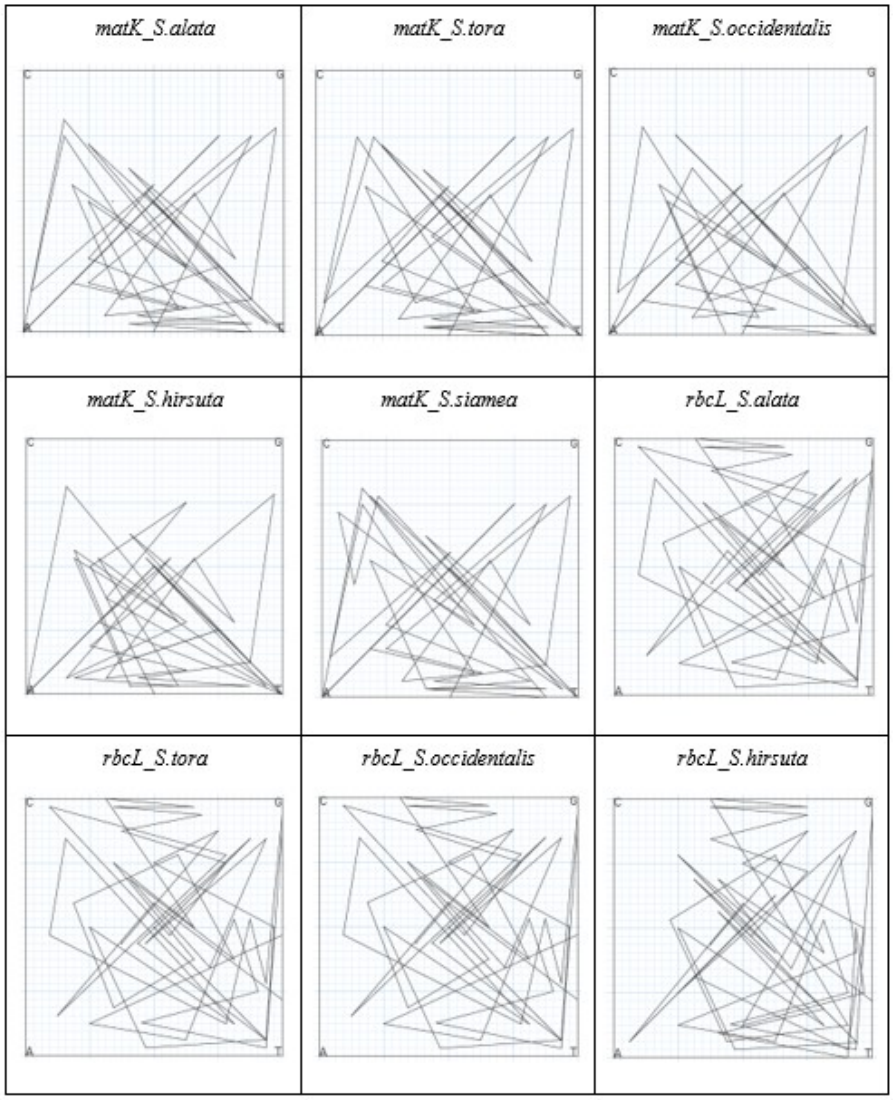

Figure 3: The 2D graphs of CGR walk representing partial matK and rbcL sequences of different Senna sp

Tables

the *matK barcoding* regions		the *rbcL barcoding* regions	
primers *matK*-f: ACCCAGTCCATCTGGAAATCTTGGTTC *matK*-r: CGTACAGTACTTTTGTGTTTACGAG		*rbcL*-f: ATGTCACCACAAACAGAGACTAAAGC *rbcL*-r: GTAAAATCAAGTCCACCRCG	
PCR reagent per 10 µl reaction			
5XHF buffer (with $MGCl_2$)	2.0 µl	5XHF buffer (with $MGCl_2$)	2.0 µl
100% DMSO	0.3 µl	100% DMSO	0.3 µl
10mM dNTPs	0.2 µl	10mM dNTPs	0.056 µl
10 µM Primer Forward	0.5 µl	10 µM Primer Forward	0.1 µl
10 µM Primer Reverse	0.5 µl	10 µM Primer Reverse	0.1 µl
ddH_2O	5.37 µl	ddH_2O	6.32 µl
Phusion HF Fisher DNA polymearse		Phusion HF Fisher DNA polymearse	
(Thermo Fisher Scientific, 5U/µl)	0.125 µl	(Thermo Fisher Scientific, 5U/µl)	0.125 µl
DNA template	1.0 µl	DNA template	1.0 µl
PCR Thermocycling Program 98°C for 45seconds 35 cycles of 98 °C for 10s, 52°C for 30s, 72°C for 40s Final extension 72°C for 10 minutes		98°C for 45seconds 35 cycles of 98 °C for 10s, 55°C for 30s, 72°C for 40s Final extension 72°C for 10 minutes	

Table 1: Primers and protocols for PCR amplification

	family	**Scientific names**	**Success of DNA amplification**	
			rbcL	***matK***
1	Leguminosae	*Senna tora*	/	/
2	Leguminosae	*Senna occidentalis*	/	/
3	Leguminosae	*Senna alata*	/	/
4	Leguminosae	*Senna hirsuta*	/	/
5	Leguminosae	*Senna siamea*		/

Table 2: List of plants and the results of DNA amplification at matK and rbcL locus (a '/' mark denotes successful DNA amplification in which a distinct PCR band of good quality was obtained. The absence of a '/' mark indicates amplification failure)

No	Scientific name	*matK*	*rbcL*
1	*S. alata*	TTGGAATAGTCTTATTACT CCAAAAAAATGGATTTCTA CTTTTTCAAAAAGGAATCC AAGATTATTCCT	TTTACTTCCATTGTGGGTA ATGTATTTGGATTCAAGGC CCTGCGCGCTCTACGTCTG GAGGATTTGCGA
2	*S. tora*	TTGGAATAGTCTTATTATTC CAAAAAAATGGATTTCTAC TTTTTCAAAAAGGAATCCA AGATTATTCCT	TTTACTTCCATTGTGGGTA ATGTATTTGGGTTCAAGGC CCTGCGCGCTCTACGTCTG GAGGATTTGCGA
3	*S. occidentalis*	TTGGAATAGTTTTATTACTC AAAAAAAATGGATTTCTAC TTTTTCAAAAAAGAATCCA AGATTTTTCCT	TTTACTTCCATTGTGGGTA ATGTATTTGGGTTCAAGGC CCTGCGCGCTCTACGTCTG GAGGATTTGCGA
4	*S. hirsuta*	TTGGAATAGTCTTATTACT CCAAAAAAATCGATTTCTA CTTTTTCAAAAAGTAATCT AAGATTTTTCTT	TTTACTTCTATTGTAGGTAA TGTATTTGGGTTCAAAGCT CTGCGCGCTTTACGTCTGG AAGATTTGCGA
5	*S. siamea*	TTGGAATAGTCTTATTACT CCAAAAAAATGGATTTCCA CTTTTTCAAAAAGGAATCC AAGATTATTCCT	

Table 3: List of plants and their partial DNA coding sequences used for CGR walk

DISCUSSION

All collected plant samples used in this study are medicinal species in the Leguminosae family commonly found in Thailand. A range of DNA sequence that is suitable to be used as a barcode must be sufficiently variable to discriminate among species, have appropriate length of DNA sequence for PCR and sequence analysis techniques and contain a conserved sequence located at both ends that are suitable for universal PCR primer design for a variety of species [6]. From the study, it was found that the prepared primers and conditions as mentioned in the procedure section showed good amplification of the *matK* and *rbcL* gene regions of all samples except the *rbcL* fragment of *Senna siamea* . The single DNA band of *matK* and *rbcL* during electrophoresis had the size as expected, approximately 900 and 600 base pairs, respectively.

DNA barcode could be used to group individuals based on genetic relationship, particularly when the primers for *matK* were used. The results were in agreement with the report of DNA barcoding in flowering plants that the *rbcL* genes appeared to have low effectiveness of evolutionary separation [5] and it was less effective in identification of some genera such as Dendrobium [13]. It was also found that a sequence of *rbcL* gene had lower rate of evolution when compared with the nucleotide sequences of other genes, therefore the information it provided alone to create a dendrogram of genetic relationship might not be enough [14]. Therefore, a two- locus global DNA barcode was more effective. Moreover, the altogether analysis of the *matK* gene and the *rbcL* gene showed that discrimination power between species was high enough for identification of 5 species of the collected *Senna* plants.

In addition, applying the Chaos Game Representation (CGR) to CGR-walk of the DNA sequences could be effectively used for comparison of similarities and dissimilarities between species. The 2D graph patterns of *matK* and *rbcL* barcoding regions, especially *matK*, of all five *Senna* species, were unique. It provided a more complete overview of biological information apart from a dendrogram. From the results, *S. alata, S. tora* and *S. siamea* were the closest in terms of DNA sequence. This indicated that 2D graphical representation was consistent with the evolutionary data as well. Although phylogenetic tree was sufficient to provide evolutionary relationship, CGR walk had more advantages as it was a graphical representation method that clearly displayed data of any length of DNA and easily detected differences between species, even with only slight sequence dissimilarities. Moreover, each base coordinate on the 2D graph could be converted into various digital forms for further downstream analysis.

While phylogenetic analysis requires the maximum amount of sequence information, CGR walk relies on characteristic patterns that can be representative of the sequence's overall features without a detailed analysis of the entire sequence. The selected representative regions for CGR should capture the essential sequence (e.g., coding regions, conserved motifs, or functionally important domains). Results from phylogenetic trees and CGR may differ due to the use of alternative data summaries and assumptions. CGR provides rapid screening, while phylogenetic analysis provides detailed evolutionary relationships and fine-scale resolution. Therefore, both methods should be used together. When results diverge, the differences themselves serve as an informative aspect requiring further investigation.

DNA sequences play an important role in modern biological research because all the information of the hereditary and species evolution is contained in these macromolecules. This research is capable of providing useful guidelines to optimize procedures for DNA amplification in combination with DNA barcoding which can be applied as a fundamental protocol to construct the DNA database of native plants based on a two- locus global DNA barcode consisting of *matK* and *rbcL*. The research also presents a novel method to extract the characteristic of the DNA sequence with the graphical operation which can effectively achieve the similarity/dissimilarity comparison of different species.

AUTHOR CONTRIBUTIONS

Patcharawarin Ruanto presented the idea, designed the methodology, planned and conducted the experiments, performed the computations and data analysis, and wrote the manuscript.

Somruthai Tunma conducted field sampling and assisted in performing the experiments.

Natthiya Chaichana investigated plant morphology and identification, and assisted with DNA experiments and data analysis.

ACKNOWLEDGMENT

We wish to thank for financial support from Chiang Rai Rajabhat University and we would like to thank staff from

Nong Bong Kai

restricted hunting area, Chiang Saen District, Chiang Rai and the Institute of biodiversity and environment for local and ASEAN development for the support.

REFERENCES

1. Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, et al. Cryptic species as a window on diversity and conservation. Trends in Ecology & Evolution, (2007); 22(3): 148-155.
2. Alvarez I, Wendel JF. Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution, (2003); 29(3): 417–434.
3. Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, et al. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American Journal of Botany, (2005); 92: 142–166.
4. Selvaraj D, Sarma RK, Sathishkumar R. Phylogenetic analysis of chloroplast
5. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the United States of
6. Shi LC, Zhang J, Han JP. Testing the potential of proposed DNA barcodes for species identification of Zingiberaceae. Journal of Systematics and Evolution, (2011); 49: 261–266.
7. Chase MW, Hollingsworth PM, Cowan RS, Berg CV. A proposal for a standardized protocol to barcode all land plants. The Journal of the International Association for Plant Taxonomy,

8. China Plant BOL Group, Li DZ, Gao AM, Li HT, Wang H, et al. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the
9. Lambowitz AM, Zimmerly S. Group II introns: mobile ribozymesthat invade DNA. Cold Spring Harbol Perspectives in Biology, (2011); 3(8): a003616.
10. Gao J. and Xu ZY. Chaos Game Representation (CGR)-walk model for DNA sequences. Chinese Physics B, (2009); 8(1): 370–376.
11. Almeida JS, Carrico JA, Maretzek A, Noble PA, Fletcher M. Analysis of genomic sequences by Chaos Game Representation. Bioinformatics, (2001); 17(5): 429–437.
12. Kress WJ, Erickson DL. DNA barcodes: methods and protocols. Methods of Molecular Biology, (2012); 858: 3-8.
13. Asahina H, Shinozaki J, Masuda K, Morimitsu Y, Satake M. Identification of medicinal Dendrobium species by phylogenetic analyses using
14. Freudenstein JV, Chase MW. Phylogenetic relationships in Epidendroideae (Orchidaceae), one of the great flowering plant radiations: Progressive specialization and diversification. Annals of Botany, (2014); 115(4): 665-681.