International Journal of Chemical Studies

P-ISSN: 2349–8528
E-ISSN: 2321–4902
IJCS 2018; 6(6): 2446-2451

Received: 09-09-2018
Accepted: 13-10-2018

Sumana Sikdar
Biotechnology Centre, Jawaharlal Nehru Agriculture University, Jabalpur, Madhya Pradesh, India

Sharad Tiwari
Department of Plant breeding and Genetics, College of Agriculture, Jawaharlal Nehru Agricultural University, Jabalpur, Madhya Pradesh, India

Vishwa Vijay Thakur
ICAR-Indian Institute of Natural Resins and Gums, Namkum, Ranchi, Jharkhand, India

Swapnil Sapre
Biotechnology Centre, Jawaharlal Nehru Agriculture University, Jabalpur, Madhya Pradesh, India

Correspondence
Sumana Sikdar
Biotechnology Centre, Jawaharlal Nehru Agriculture University, Jabalpur, Madhya Pradesh, India

An *in silico* approach for evaluation of *rbcL* and *matK* loci for DNA barcoding of fabaceae family

Sumana Sikdar, Sharad Tiwari, Vishwa Vijay Thakur and Swapnil Sapre

Abstract
The main aim of present work is to demonstrate effectiveness of *in silico* approach to gain a prior insight into the efficacy of the *rbcL* and *matK* loci as DNA barcodes for a group of plants belonging to fabaceae family. For this *in silico* analysis, we retrieved all available sequences of selected plants species from NCBI and examined for their species resolution ability. The species resolution ability of *matK* and *rbcL* loci was varied from genus to genus. In genus *Vigna*, *Cassia* and *Crotolaria*, *matK* is an ideal locus for DNA barcoding whilst in genus *Sesbania*, *rbcL* has proved their efficacy. This *in silico* approach makes the process of validation of novel barcodes much easier by decreasing the number of candidate primer pairs which are required to be tested *in vitro*.

Keywords: DNA barcoding, *matK*, *rbcL*, fabaceae

1. Introduction
Fabaceae is the third largest family of angiosperms worldwide, with approximately 800 genera and more than 20000 species. A numbers of species belongs to this family which has huge medicinal importance. However some species of this family have been reported to be toxic. For instance, *Cassia occidentalis* has been shown to contain considerable amount of various toxic alkaloids, anthraquinones (Vashishtha *et al*., 2009) [13], and numerous species under genus Crotalaria contains a toxic alkaloids pyrrolizidine, which are lethal to mammals and birds (Williams and Molyneux, 1987) [16]. There are many plant species of this family are at the point of extinction due to the biological consequence of climate change, ascending habitat destruction and over exploitation. In addition, for successful commercialization of herbal formulations the quality of phytopharmaceuticals must be as high as that of its other counterparts (Butt *et al*., 2018) [1]. It is very difficult to discover potential adulterants in the processed herbal formulation which directly diminish the quality of phytopharmaceuticals (Thakur *et al*., 2016) [11]. Hence, advancement in DNA-based markers is critical for identification of plants as well as confirmation of herbal ingredients in products to diminish the chance of adulteration and illegal trade of plant species belonging to this family.
In 2003 a Canadian scientist Dr. Paul D.N. Hebert proposed a novel approach for species identification termed as DNA barcoding. In this approach a short stretch of DNA sequence has been used as a signature sequence for identifying any particular species. Since its first introduction in 2003 a single region 5′ end of mitochondrial cytochrome c oxidase 1 (CO1) has been proved to successfully identify animal species (Herbert *et al*. 2003; Shneer, 2009) [4, 10]. However, in case of plants due to low nucleotide substitution rate this region was failed to identify plants species. Hence, Universal regions for recognition of plants species are also being searched from nuclear and plastid genomes. However, until now, the single and universal locus for DNA barcoding of plants remains debatable because some loci have just been efficient for some specific taxonomic groups. Thus, to fulfill the demand of plant identification, separate sequences for identifying particular taxonomic groups have been done. There were a number of recent papers have reviewed DNA barcoding in plants (Vijayan and Tsou, 2010; Hollingsworth *et al*., 2011) [5, 14]. After that, a number of loci in the chloroplast genome have been tested for barcoding and ribulose-1, 5-biphosphate carboxylase (*rbcL*) and maturase K (*matK*) have been proposed as preferred plant barcoding loci by consortium for the barcodes of life (CBOL, 2009) [2].
The main aims of this *in silico* study to evaluate the species resolution ability of *matK* and *rbcL* loci in fabaceae family available on the National Center for Biotechnology Information (NCBI). This study may contribute to gain a prior insight in to the efficacy of the

Above-mentioned locus/loci as DNA barcodes for identification of species of fabaceae family, serving for conservation and diversity study of the plant.

2. Material and Method

2.1 Collected sequences of Fabaceae from GenBank

The sequences of two most potential barcode loci, *matK* and *rbcL*, belonged to fabaceae family were selected and downloaded from GenBank. All the downloaded sequences were grouped into four sets according to their genus. The species details of each locus with their accession number are shown in Table 1.

Table 1: List of analyzed plants species with GenBank accession number divided in different groups according to their genus

Group	Species name	GenBank accession no.	
		rbcL	***matK***
Group I	*Vigna aconitifolia*	KX087386.1	JN008217.1
	Vigna aconitifolia	KX087385.1	JN008216.1
	Vigna aconitifolia	KX087384.1	JN008214.1
	Vigna angularis	EU288936.1	JN008218.1
	Vigna angularis	EU288935.1	JN008269.1
	Vigna angularis	EU288934.1	-
	Vigna mungo	KX087448.1	JN008223.1
	Vigna mungo	KX087447.1	AY582994.1
	Vigna mungo	KX087446.1	-
	Vigna radiata	KU519327.1	AP014691.1
	Vigna radiata	KU519326.1	JN008226.1
	Vigna radiata	KU519325.1	DQ445950.1
	Vigna umbellata	KX087507.1	JN008237.1
	Vigna umbellata	KX087506.1	JN008238.1
	Vigna umbellata	Z95543.1	AY582995.1
	Vigna unguiculata	KX119333.1	EU717407.1
	Vigna unguiculata	EU717266.1	JN008198.1
	Vigna unguiculata	-	AY582999.1
Group II	*Senna occidentalis*	KU551086.1	JQ301880.1
	Senna occidentalis	KP095069.1	KJ638443.1
	Senna occidentalis	JQ301860.1	KJ638444.1
	Senna tora	JF949969.2	JQ301877.1
	Senna tora	JQ301857.1	KY549329.1
	Senna tora	KX119325.1	KJ638441.1
	Senna uniflora	JQ301867.1	-
	Senna uniflora	KY464122.1	-
	Senna uniflora	KY464121.1	-
	Senna alexandrina	JQ301866.1	JQ301886.1
	Senna alexandrina	KY464110.1	KY513093.1
	Senna alexandrina	KY464109.1	KY513092.1
	Cassia fistula	JQ301850.1	KJ638430.1
	Cassia fistula	KX385947.1	KJ638429.1
	Cassia fistula	KY368649.1	JQ301870.1
	Senna siamea	-	KJ012767.1
	Senna siamea	-	JQ301882.1
	Senna siamea	-	KJ638438.1
Group III	*Crotalaria pallid*	KJ773413.1	KJ772687.1
	Crotalaria pallid	KX119281.1	KX119367.1
	Crotalaria lanceolata	KJ773412.1	-
	Crotalaria lanceolata	EU348040.1	-
	Crotalaria humilis	EU348042.1	-
	Crotalaria humilis	EU348041.1	-
	Crotalaria virgultaris	EU348044.1	-
	Crotalaria virgultaris	EU348037.1	-
	Crotalaria spectabilis	-	KJ772689.1
	Crotalaria spectabilis	-	AB649973.1
Group IV	*Sesbania sesban*	MF135445.1	-
	Sesbania sesban	MF135431.1	-
	Sesbania sesban	Z95541.1	-
	Sesbania emerus	JQ592035.1	-
	Sesbania emerus	JQ592034.1	-
	Sesbania emerus	JQ592033.1	-
	Cicer arietinum	-	AY386897.1
	Cicer arietinum	-	AB198874.1
	Cicer arietinum	-	AB198873.1
	Cicer pinnatifidum	-	AB198890.1
	Cicer pinnatifidum	-	AF522081.1

	Cicer reticulatum	-	AB198933.1
	Cicer reticulatum	-	AB198899.1

2.2 Multiple sequence alignment and Assessment of sequence divergence

Sequences at each locus were aligned using clustal W program in MEGA X software (Kumar *et al*., 2018) [6]. Alignments were then manually optimized and the parsimony informative sites, variable sites, nucleotide composition, GC% and inter-specific and intra-specific distances calculations for each data set were performed by same software. In the process of alignment, sequences that were too short and too divergent were removed from data sets.

2.3 Assessment of species resolution

A number of methods have been used for the investigation of species resolution ability of barcoding loci. Among others, phylogenetic analysis and barcoding gap approaches are the most frequently used for DNA barcode data analysis.

For phylogenetic analysis we used NJ tree method with 1000 bootstrap because of its robustness and accuracy. In order to estimate species resolution for a given barcode locus, we considered the species were resolved if conspecific individual grouped into one monophyletic branch in the phylogenetic tree with well bootstrap support. Second, in converse, if conspecific individuals were separated in paraphyletic branches, then the species was considered as identification failure.

In another approach, the minimum inter-specific p distance and maximum intra-specific p distance were determined using MEGA X. In this approach, a species is considered to be resolved if its minimum inter-specific distance is greater than maximum intra-specific distance.

3. Results

3.1 Estimation of sequence divergence

In order to estimate sequence divergence of *rbcL* and *matK* loci, mean sequence length, parsimony informative sites, variable sites, GC percent, average inter-specific and intra-specific distance were calculated and presented in Table 2.

In the first group, percentage of parsimony informative sites and variable sites of *rbcL (20.08 and* 21.50) is much higher than *matK* (2.58 and 3.37). The *rbcL* also showed higher average GC percent (42.925) than *matK* (28.192). However mean sequence length of *matK* (1802.125) is larger than *rbcL* (631.941). The average inter-specific and intra-specific distance revealed by *rbcL* was 0.512 and 0.007 respectively whilst, *matK* showed 0.012 and 0 respectively.

Analyses carried out on samples belonging to Group II showed that the sequence divergences such as mean sequence length, parsimony informative sites, variable sites and average GC percent of marker *rbcL* was 645.6, 2.52%, 4.79% and 43.35% respectively. On the other hand except GC% (32.06%), mean sequence length (805.66), parsimony informative sites (5.65%), and variable sites (6.76%) revealed by *matK* were higher than *rbcL*. The average inter-specific and intra-specific distance revealed by *rbcL* was 0.013 and 0.012 respectively whilst, *matK* showed 0.029 and 0.003 respectively.

Analyses carried out with *matK* and *rbcL* on Group III samples revealed that the mean sequence length, parsimony informative sites, variable sites and average GC percent of marker *matK* was 895.5, 0.59%, 1.38% and 31.79% respectively. In contrast, except variable sites (43.46%) *rbcL* had the highest mean sequence length (1258.375), parsimony informative sites (0.7%), and average GC percent (43.46%). The average inter-specific and intra-specific distance revealed by *rbcL* was lower than *matK* (Table 2).

In group IV the mean sequence length and the average GC% of *matK* was 1650.142 and 30.733 respectively. In case of rbcL the mean sequence length and the average GC% was 927.8 and 42.6 respectively. 0.787% parsimony informative sites and 1.36% variable sites was found in *matK* while *rbcL* showed 0.07% parsimony informative sites and 0.7% variable sites. The average inter-specific and intra-specific distance revealed by *matK* was 0.005 and 0 respectively whilst, *rbcL* showed 0.002 and 0 respectively.

Table 2: Comparative performances and variability of *rbcL* and *matK* loci

	Group I		**Group II**		**Group III**		**Group IV**	
	rbcL	***matK***	***rbcL***	***matK***	***rbcL***	***matK***	***rbcL***	***matK***
No. of species (no. of individuals)	6(17)	6 (16)	5 (15)	5(15)	4 (8)	2 (4)	2 (6)	3 (7)
Mean sequence length	631.941	1802.125	645.6	805.666	1258.375	895.5	927.8	1650.142
Average GC%	42.925	28.192	43.35	32.064	43.468	31.798	42.6	30.733
% Parsimony informative characters	20.08	2.58	2.52	5.65	0.7	0.59	0.07	0.787
% variable sites	21.50	3.37	4.79	6.76	0.98	1.38	0.70	1.36
Average inter-specific distance	0.512	0.012	0.013	0.029	0.005	0.022	0.002	0.005
Average intra-specific distance	0.007	0	0.012	0.003	0.002	0.008	0	0

3.2 Estimation of species resolution

To evaluate the species resolution efficiency of the *rbcL* and *matK*, we followed two approaches, namely, Distance based approaches and Neighbor-joining (NJ) tree based approaches.

In the Distance based approaches, *matK* performed well in all group except group IV (Figure 1). The species resolution ability of *matK* was 100% in group I, group II and group III whereas, *rbcL* showed 82.35% for group I, 20% in group II and 50% in group III. In contrast, in group IV where *matK* and *rbcL* sequences represented by two different genus, namely, cicer and sesbania respectively, and the species identification ability of *matK* and *rbcL* were 28.7% and 100% respectively.

In Neighbor-joining (NJ) tree based approaches, where we assumed the species is resolved if they construct monophyletic clad with their conspecific individuals (Figure 2 and Figure 3). In group I 47.05% species was resolved by *rbcL* which is less than *matK* where 81.25% species was resolved. In group II the NJ tree of *matK* sequences showed 100% species resolution with presence of 5 distinct well supported clad, whereas *rbcL* failed to identify any species. Analysis carried out in group III revealed that *rbcL* showed comparatively better species resolution i.e. 75% than *matK* (50%). Species resolution ability of *rbcL* and *matK* were 100% and 28.57% in group IV.

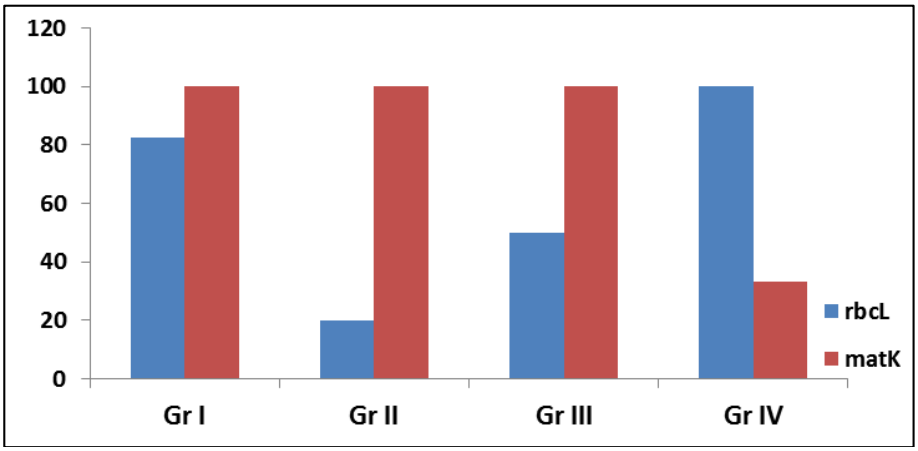

Fig 1: Percent species resolution ability of *rbcL* and *matK* loci using distance based approach

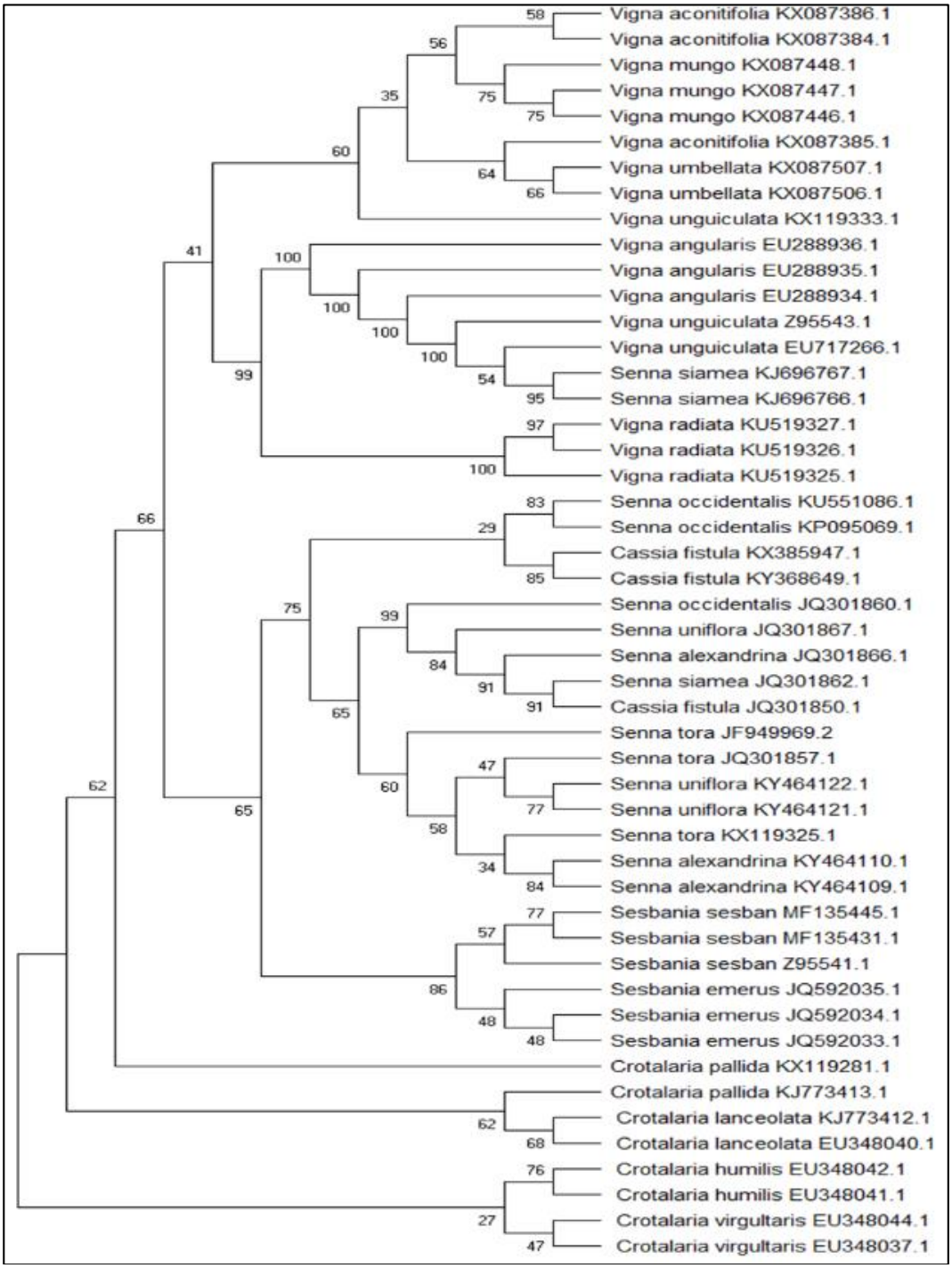

Fig 2: NJ tree with 1000 bootstrap replicates based on *rbcL* gene sequence data

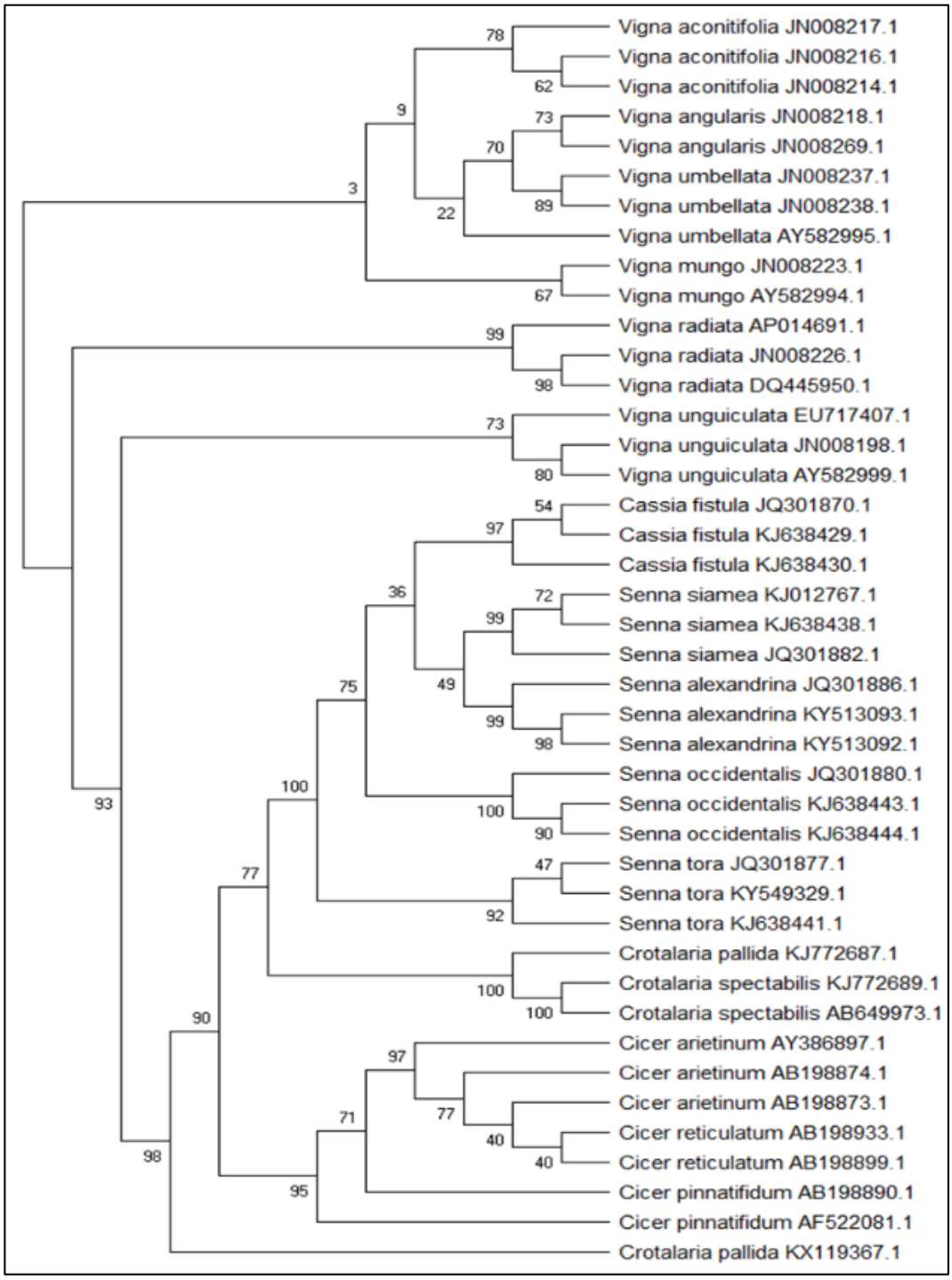

Fig 3: NJ tree with 1000 bootstrap replicates based on *matK* gene sequence data

4. Discussion

The main goal of this study was to resolve the sequence divergence and species identification ability of *rbcL* and *matK* sequences of fabaceae family available on GenBank. The *matK* and *rbcL* loci are recommended as primary barcode loci for plants by consortium for the barcode of life (CBOL, 2009) [2]. To have a good species resolution; first, the preferred barcode sequences should have a suitable length for trouble-free amplification, sequencing and alignment. Second, for better species resolution maximum intra-specific variability should be less than minimum inter-specific variability, means sequences should have enough variability at species level but not too variable at within species level. In our study, we did not investigate the PCR amplifications efficiency and sequencing success rats because of the sequences were retrieve from GenBank. However, other parameter such as the sequence length, the sequence variability and the species identification ability were all taken into account.

In this study, we provide the first *in silico* evaluation of *matK* and *rbcL* region as land plant barcode for fabaceae family. Our results showed that the variability of these two regions is consistently high across fabaceae species. Even within the species complexes represented by four different genera of fabaceae family, *matK* and *rbcL* revealed the enough sequence variability, advocating their use as a core DNA barcode. In order to estimate species resolution ability on the basis of pair wise distance, *matK* showed extremely promising results in group I, group II and group III except in genus *cicer* (group IV). In NJ tree approach, highest species resolution was observed in group II, followed by group I, group III and group IV. The high species resolution ability of *matK* region was has already been revealed by many research groups and proposed

as a core marker for DNA barcoding in plants (Lahaye *et al*., 2008; Ragupathy *et al*., 2009) [7, 9]. Low species resolution in genus *cicer* (group IV) was also revealed in previous study conducted by Gao *et al*., (2010) [3] where they found only 50% species resolution. The low species resolution percentage in *cicer* genus might be due to small sample size, so there are need to examine the species identification ability of this region with large number of samples. Our results reveal that *matK* is an ideal locus for species identification within fabaceae family.

In order to investigate species resolution ability of *rbcL* locus on the basis of pairwise distance approach were highest in group IV followed by group I, group III and group II. On the other hand on the basis of NJ tree based approach, species discrimination ability of this locus is highest in group IV followed by group III and group I. In group II this locus is completely failed to resolve any species. However, this locus contains very small sample size in group IV henceforth the results revealed by *rbcL* were not reliable to evaluate species resolution in group IV. The species resolution ability of *rbcL* markers was also questionable in previous studies (Liu *et al*., 2011; Yuan *et al*., 2015; Tripathi *et al*. 2013; Vu *et al*., 2018) [8, 12, 15, 17]. It should be noted that availability of *matK* and *rbcL* sequences on NCBI database particularly for some genus belongs to group III and group IV are very few in number and for more reliable information about species resolution ability of these reasons higher number of samples are required.

5. Conclusions

Our *in silico* analyses revealed that species resolution ability of *matK* and *rbcL* loci were varied from genus to genus. In genus vigna, cassia and crotolaria *matK* is an ideal locus for DNA barcoding whilst in genus sesbania *rbcL* proved their species identification ability. Besides, the parsimony informative sites and variable sites could be used as effective supporting data for molecular discrimination of species. However, due to paucity of available sequence information for some species on NCBI and the differences of sequence number among different loci our evaluation still limited when comparing directly among loci. Therefore, we suggest a further in silico study should be done on a higher number of samples to enhance identification ability for many useful applications in conservation and biodiversity foundation for this medicinally important family. In addition, this *in silico* method has proved to be helpful to identify the most apposite barcode for different purposes. Numerous bioinformatics tools and large databases available make such *in silico* approach possible to be conducted on any particularly targeted DNA regions. Such approach makes the process of validation of novel barcodes much easier by decreasing the number of candidate primer pairs which are required to be tested in vitro.

6. Acknowledgements

The author is thankful to DST-INSPIRE Fellowship, Department of Science and Technology, Government of India, India for financial support to carry out research work.

7. Conflict of interest

We declare that we have no conflict of interest.

8. References

1. Butt J, Ishtiaq S, Ijaz B, Ali Mir Z, Arshad S, Awais S. Authentication of polyherbal formulations using PCR technique. Annals of phytomedicine. 2018; 7:131-139.
2. CBOL Plant Working Group. A DNA barcode for land plants. Proceedings of the National Academy of Sciences. USA. 2009; 106:12794-12797.
3. Gao T, Yao H, Song J, Liu C. Identification of medicinal plants in the family Fabaceae using a potential DNA barcode *ITS2*. Journal of Ethno pharmacology. 2010; 130:116-121.
4. Hebert PDN, Ratnasingham S, deWaard JR. Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings Biological sciences. 2003; 270:96-99.
5. Hollingsworth PM, Graham SW, Little DP. Choosing and Using a Plant DNA Barcode. PLoS One, 2011, 6(5). doi: ARTN e19254
6. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X. Molecular Evolutionary Genetics Analysis across computing platform. Molecular Biology and Evolution. 2018; 35:1547-1549.
7. Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G. DNA barcoding the floras of biodiversity hotspots. Proceedings of the National Academy of Sciences. 2008; 105:2923-2928.
8. Liu J, Moller M, Gao L, Zhang D, Li D. DNA barcoding for the discrimination of Eurasian yews (Taxus L., Taxaceae) and the discovery of cryptic species. Molecular Ecology Resources. 2011; 11(1):89-100.
9. Ragupathy S, Newmaster SG, Murugesan M, Balasubramaniam V. DNA barcoding discriminates a new cryptic grass species revealed in an ethnobotany study by the hill tribes of the Western Ghats in southern India. Molecular Ecology Resources. 2009; 9:164-171.
10. Shneer VS. DNA barcoding is a new approach in comparative genomics of plants. Genetika. 2009; 45:1436-1448.
11. Thakur VV, Tiwari S, Tripathi N, Tiwari G, Sapre S. DNA barcoding and phylogenetic analyses of mentha species using *rbcL* sequences. Annals of phytomedicine. 2016; 5:59-62.
12. Tripathi AM, Tyagi A, Kumar A, Singh A, Singh S. The Internal Transcribed Spacer (ITS) Region and trnhH-psbA Are Suitable Candidate Loci for DNA Barcoding of Tropical Tree Species of India. PLoS ONE. 2013; 8(2):e57934.
13. Vashishtha VM, John TJ, Kumar A. Clinical and pathological features of acute toxicity due to *Cassia occidentalis* in vertebrates. The Indian Journal of Medical Research. 2009; 130:23-30.
14. Vijayan K, Tsou CH. DNA barcoding in plants: taxonomy in a new perspective. Current Science. India. 2010; 99:1530-1541.
15. Vu HT, Huynh P, Tran HD, Le L. In Silico Study on Molecular Sequences for Identification of Paphiopedilum Species. Evolutionary Bioinformatics. 2018; 14:1-9.
16. Williams MC, Molyneux RJ. Occurrence, concentration, and toxicity of pyrrolizidine alkaloids in Crotalaria seeds. Weed Research. 1987; 35:476-481.
17. Yuan Q, Zhang B, Jiang D, Wang NH, *et al*. Identification of species and materia medica within *Angelica* L. (*Umbelliferae*) based on phylogeny inferred from DNA barcodes. Molecular Ecology Resources. 2015; 15(2):358-371.