CHAPTER 4
RESULTS

CHAPTER 4

RESULTS

The search for a species specific DNA barcode for higher plants is hampered by problems in finding a suitable locus due to the inadequate sequence variability to allow species level discrimination. The key requirement of such loci is that they should amplify using universal primer from a wide a range of taxa and possess sufficient sequence variability to enable species diagnosis. Since *matK* (Coding) and *trnH-psbA* (non coding) are the most variable region in chloroplast genome and usually has only moderate level of intra- species variation, so it very much useful to study the variation in inter-species level.

Thus our aim is to compare the performance of core DNA barcode loci (*matK* and *trnH-psbA*) for medicinal plants and were assessed based on recoverability and discrimination in **CHAPTER 4.1**.

Study the species level phylogeny of medicinal plants with special reference to the family Apocynaceae in **CHAPTER 4.2**.

We also developed species diagnostic characters based on single nucleotide polymorphism or insertion and deletion in **CHAPTER 4.3**.

Chapter 4.1 : Chloroplast *matK* as well as *trnH-psbA* intergenic spacer as possible DNA passport for common medicinal plants

4.1.1 Primer efficiency and recovery:

All the DNA samples were of high molecular weight and there were no degradation observed during the isolation period. The amplified *matK* products of ~900 bp and *trnH-psbA* products of ~450 bp were resolved in agarose gel (Figure- 4.1 and Figure- 4.2). The forward primer of *matK* (*matK* X-F) was annealed at 301-322 positions, 5'- 3' directions and amplified ~900 bp which is exactly complementary to that of another strand of Cp DNA. The reverse primer (*matK* 5r) was annealed at 1303-1322 positions of amplified ~950 bp that was exactly complimentary of another strand. Similarly in *trnH-psbA*, forward primer was amplified ~400 bp and the reverse primer was amplified ~450 bp. PCR was run using approximately 100-200 ng of DNA from each sample using non-coding (*trnH-psbA*) and coding (*matK*) primers. In case of *trnH-psbA* region, single band PCR product of approximately 450bp was obtained for all the samples. While in the case of *matK* coding region multiple bands were obtained from *C. limon* of approximately 600-800bp and *Citrus. maxima, Citrus limetta, Citrus paradisi* did not amplify (Figure- 4.3). The highly variable *matK* region has lower PCR amplification success than the more conserved *trnH-psbA* spacer. We developed family-specific primers of *matK* in case of Rutaceae family member because the publish primer were problematic in C*itrus* species. Our newly designed universal primer for Rutaceae family (RUT-*matK*-F and RUT-*matK*-R) was tested on *Citrus* species. Primers were shown 100% efficiency to amplify the ~900bp matK sequences (Figure- 4.4). The forward primer of *matK* was annealed and amplified ~800 bp which is exactly complementary to that of another strand of Cp DNA. The reverse primer was amplified ~900 bp and exactly complimentary of another strand. So the targeted *matK* and *trnH-psbA* sequences were estimated as combination of forward primer sequence.

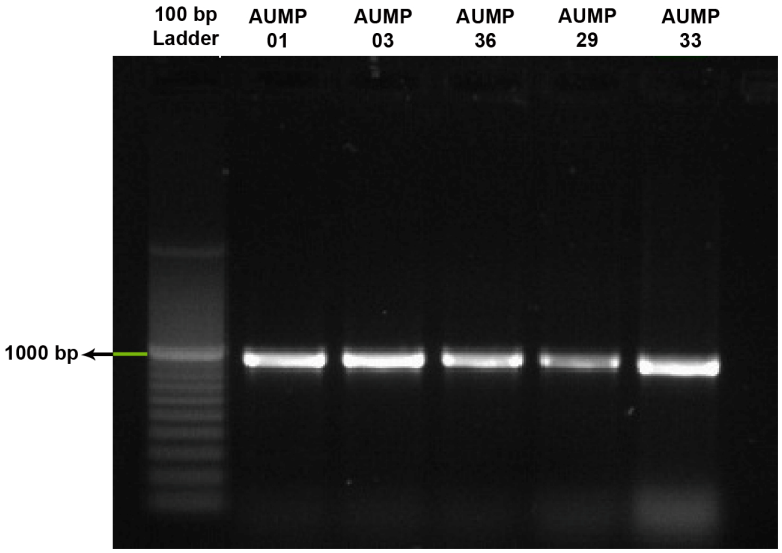


Figure - 4.1 Agarose gel electrophoresis of 1000 bp of amplification products of *matK* of Apocynaceae family members. Left to right sample numbers are AUMP 01, 03, 36, 29, 33. Amplification products size are ~850 bp. DNA ladder is 100 bp in extreme left.

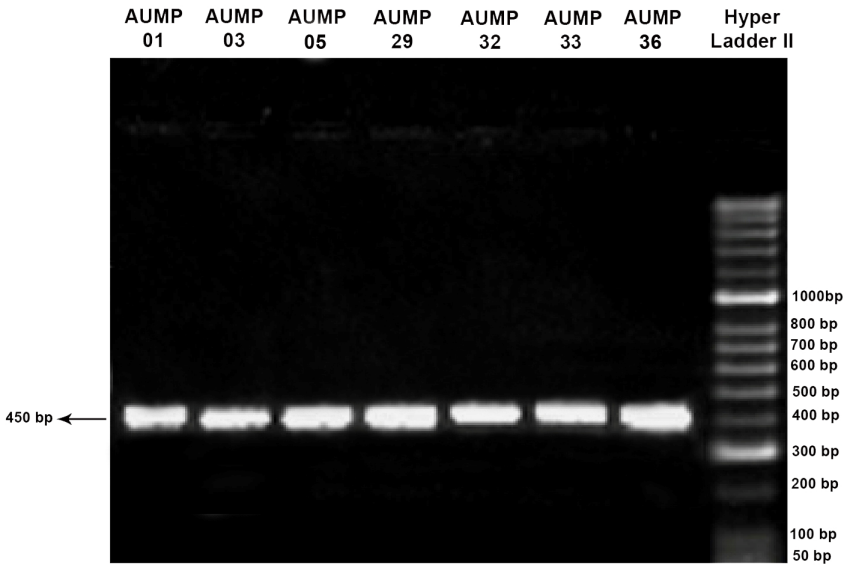


Figure - 4.2 Agarose gel electrophoresis of ~450 bp PCR amplification products of *trnH-psbA* of Apocynaceae plants members. Left to right sample numbers are AUMP01, 03, 05, 29, 32, 33, 36. DNA ladder is Hyper variable II.

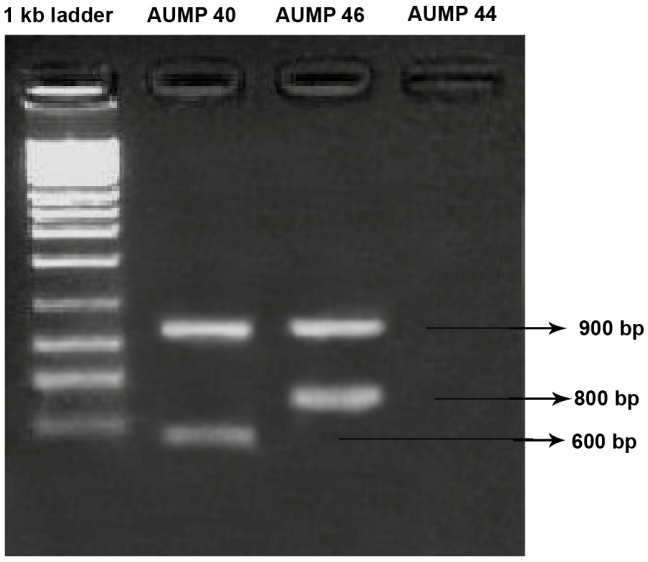


Figure - 4.3 Agarose gel electrophoresis of *matK* amplification products of *Citrus* species by using *matK-X F and matK 5r* primer and produce unwanted band or multiband in sample AUMP40 and AUMP46.There is no amplification in sample AUMP 44. DNA ladder is Hyper variable II.

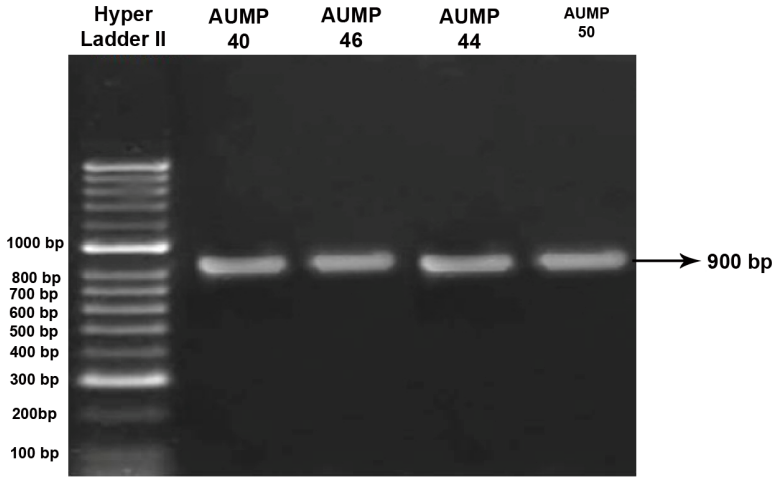


Figure - 4.4 Agarose gel electrophoresis of 900 bp of *matK* amplification products of *Citrus* species by using RUT *matK-F* and RUT *matK-R* primer and produce single band. DNA ladder is Hyper variable II.

4.1.1.1 Selection of primer :

PCR is capable of amplifying a single target DNA fragment out of a complex mixture of DNA. This ability depends on the specificity of the primers. Primers are short single-stranded oligo-nucleotides which anneal to template DNA and serve as a "primer" for DNA synthesis. In order to achieve the geometric amplification of a DNA fragment, used two primers, one of each primer flanked each end of the target DNA. The primers were complementary to the target DNA. The forward primer was complementary to the positive strand and amplified 5' to 3' direction. The reverse primer was complementary to the negative strand and amplified 3'-5' direction. However, we always write DNA sequences in the 5'-3' direction so the reverse primer would be written as a reverse complementary form (Figure- 4.5). The consensus sequences were developed using bioinformatics tool Clustal X which covers all the available complete *matK* sequence of the family Rutaceae. The designed consensus sequences are shown in Figure-4.6. All the primer pairs amplify approximately the same fragment. We take forward primer sequences from consensus sequences from 5' end and reverse primer from 3' end of multiple sequence alignment and slightly modified some nucleotide of the primer sequences on the basis of criteria of primers.

Selected forward primer:

RUT- *matK*-F 5'- **TCA GAG GTA TTT GCT GCT GTG GTG**-3' (24nt)

Selected Reverse complementary sequence (Reverse primer):

RUT- *matK***-** R 5'- **GAC CAA GTC GAC CTA CTG ATA GG** -3' (23nt)

4.1.1.2 *Insilico* validation of primers:

Primer development to improve the overall success rate of *matK* amplification is essential for its application as part of the universal plant barcode. Before starting a barcode project on any new taxonomic group, it is essential to test the efficiency

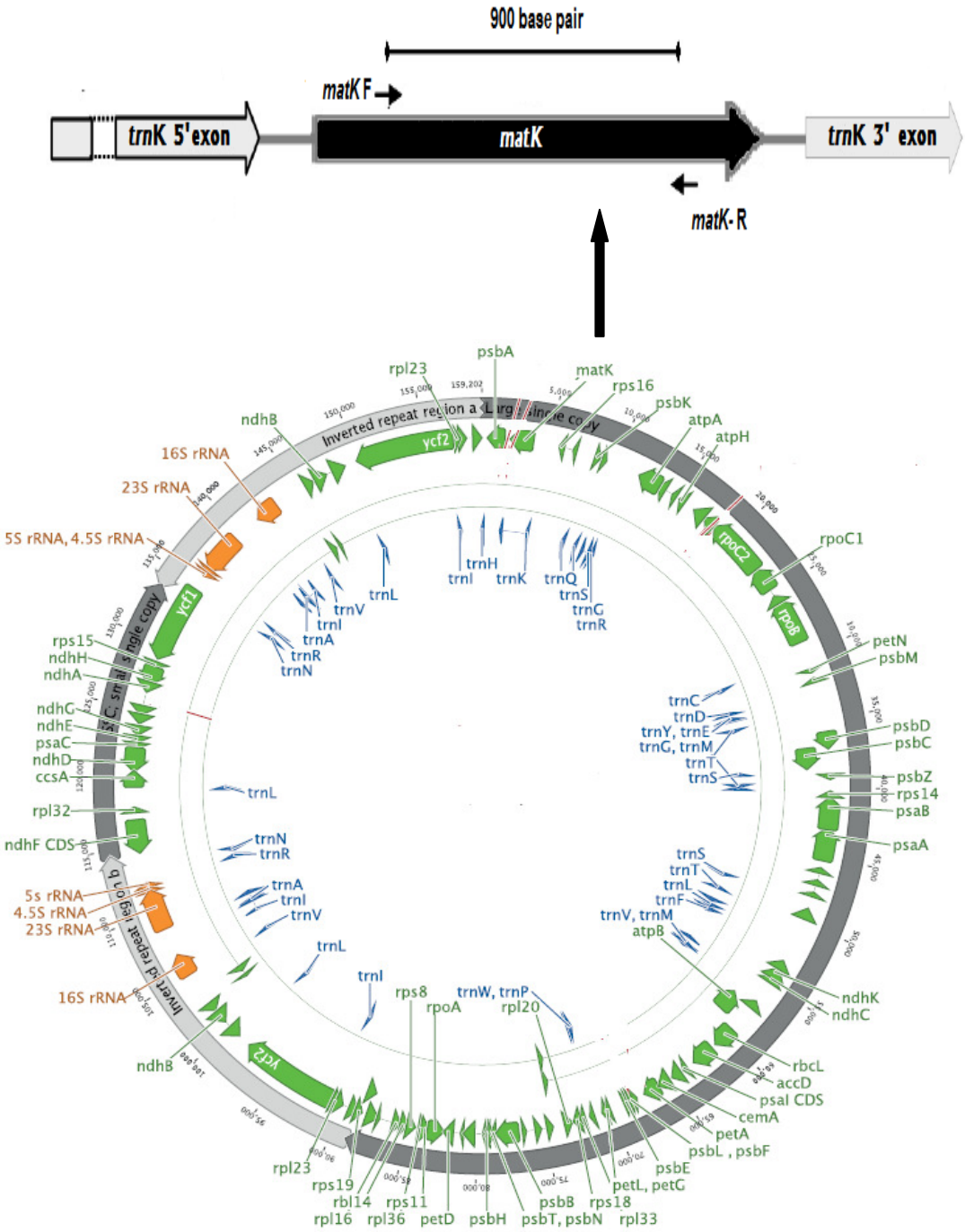


Figure - 4.5 Diagrammatic representation of *matK* primer design for Rutaceae. Bottom: Diagrammatic representation of chloroplast genome.

Multiple sequence alignment (MSA) for Forward sequences:

```
Citrus trifoliate                TGGGCACAACAATAAGTTGTATTCTCAAATTATATCAGAGGGATTTGCTGTCGTGGTGGAAATC
Citrus reticulata                TGGGCACAACAATAAGTTATATTCTCAAATTATATCAGAGGTATTTGCTGCTGTGGTGGAAATT
Zanthoxylum  americanum          TGGGCACAACAATAAGTTATATTCTCAAATTATATCAGAGGTATTTGCTGCTGTGGTGGAAATT
Zanthoxylum simulans             TGGGCACAACAATAAGTTGTATTCTCAAATTATATCAGAGGGATTTGCTGTCGTGGTGGAAATT
Casimiroa edulis                 TGGGCACAACAATAAGTTGTATTCTCAAATTATATCAGAGGGATTTGCTGTCGTGGTGGAAATT
Ruta angustifolia                TGGGCACAACAATCTGTTGTATTCTCAAATTCTATCAGAAGGATTTGCTGTCGTGGTGGAAATT
Severinia buxifolia              TGGGCACAACAATCTGTTGTATTCTCAAATTCTATCAGAAGGATTTGCTGTCGTGGTGGAAATT
Ruta microcarpa                  TGGGCACAACAATAGCTTGGATTCTCAAATTATATCAGAGGTATTTGTTGCCGTGGCGGAAATT
Haplophyllum laeviusculum        TGGGCACAACAATCTGTTGTATTCTCAAATTCTATCAGAAGGATTTGCTGTCGTGGTGGAAATT
Ruta oreojasme                   TGGGCACAACAATAAGTTGTATTCTCAAATTATATCAGAGGTATTTGCTGCTGTGGTGGAAATT
                                 *************   **  *********** ***************  **  **** ******
```

Multiple sequence alignments (MSA) for reverse primer:

```
Citrus trifoliate       TTTTGTAACGCATTAGGGCATCCCATCAGTAAGGCGACCTGGGCCGATTTCTCAGATTCTCATCTTATCGA
Citrus reticulata       TTTTGTAACGCATTAGGGCATCCTATCAGTAAGTCGACTTGGTCCGATTTCTCTGATTCTCATCTTATCGA
Z. americanum           TTTTGTAACGCATTAGGGCATCCTATCAGTAAGTCGACTTGGTCCGATTTCTCTGATTCTCATCTTATCGA
Z. simulans             TTTTGTAACGCATTAGGGCATCCCATCAGTAAGGCGGCCTGGGCCGATTTCTCTGATTCTCATCTTATCGA
Casimiroa edulis        TTTTGTAACGCATTAGGGCATCCCATCAGTAAGGCGGCCTGGGCCGATTTCTCTGATTCTCATCTTATCGA
Ruta angustifolia       TTTTGTAATGCATTAGGGCATCCCATCAGTAAGTCGTCCTGGGCCGATTTCTCTGATTCTCATCTTATCGA
Severinia buxifolia     TTTTGTAATGCATTAGGGCATCCCATTAGTAAGTCGTCCTGGGCCGATTTCTCTGATTCTCATCTTATCGA
Ruta microcarpa         TTTTGTAACGCATTAGGGCATCCCATCAGTAAGTCGACTTGGGCCGATTTTTCTGATTCTAATCTTATCGA
H. laeviusculum         TTTTGTAATGCATTAGGGCATCCCATCAGTAAGTCGTCCTGGGCCGATTTCTCTGATTCTCATCTTATCGA
Ruta oreojasme          TTTTGTAACGCATTAGGGCATCCTATCAGTAAGTCGACTTGGGCCGATTTCTCTGATTCTCATCTTATCGA
                        ********  *************  ********  *  * *** ******  ******************
```

Figure - 4.6 Show Multiple Sequence Alignment of Forward and Reverse primer region and developed clade (family) specific forward and reveres primers.

of existing primers *insilico* by using PCR test, Primer blast and Oligocal. Properties of published primers were depicted in Table-4.1. Our designed primers set used for PCR amplification and have a major effect on the specificity and sensitivity of the reaction. There are many possible regions of the entire *matK* gene that may be used as a plant barcode. We chose the region from 400 to 1300 bp to be the best representative of the whole-gene sequence. Considering the conservative regions for primer design and the mononucleotide repeats that cause sequencing problems. The final region of *matK* was adjusted to contain the sequence from 432 to 1348 bp. When we were choosing two PCR amplification primers, the following guidelines are considered.

	***matK*-X F**	***matK*-5r**	**RUT-*matK*-F**	**RUT-*matK*-R**	***trnH*-F**	***psbA*-R**
Tm	50° C	56° C	58^0 C	59^0 C	66°C	58°C
GC Content	24%	45%	54%	52%	58%	41%
Length	21 bp	20 bp	24 bp	23 bp	23 bp	22 bp
Primer Dimer	NO	NO	NO	NO	NO	NO
Hair Pins	NO	NO	NO	NO	NO	NO
Insilico PCR product length	945 bp		902 bp		560 bp	

Table- 4.1 Properties of primers (Melting Temperature (Tm) , GC content, Length, primer dimer, Hair pins, Insilico PCR product length) of *matK*-X F, *matK*-5r, RUT-*matK*-F, RUT-*matK*-R, *trnH*-F, *psbA*-R

Primer length - In broad-spectrum studies, primers of typically 18-30 nucleotides in length are the best. Length of the published primer *matK*-X F , *matK*-5 r are 21 bp and 20 bp and *trnH*-F, *psbA*-R are 23bp and 22 bp. Our newly designed, RUT-*matK*-F (Froward) primers have 24 bp and RUT-*matK*-R (Reverse) primer have 23 bp (Table- 4.1).

Melting Temperature (Tm) - The optimal melting temperatures for primers in the range 50-58^{0}C, generally produce better results than primers with lower melting temperatures. All the primer set maintain the optimal melting temperature except *trnH-F* (66^{0}C) (Table- 4.1)

GC Content (Tm and Ta are interrelated) - GC% is an important characteristic of DNA and provides information about the strength of annealing. Primers should have GC content between 45 and 60 percent. GC content of *matK* X-F and *psbA*-R were below the thresold value (Table- 4.1).

3'-End Sequence - Stable base pairing of the 3' end of a primer and the target DNA is necessary for efficient DNA synthesis. To ensure the stability of this interaction, primers are often designed to end in either a G or a C. This terminal G or C is called a G/C clamp. We found the G/C clamp in each primer used in this study.

Dimers and false priming cause misleading results - Primers should not contain complementary (palindromes) within themselves. Hairpins that form below 50^o C generally are not such a problem. We didn't find any primer dimer in our designed primer because all forward and reverse primers were not contained any repeated sequence for which it binds to its specific template DNA. Published and design primers also have no primer dimer formation (Table- 4.1).

4.1.2. *matK* barcode identification

4.1.2.1 Identification of important medicinal plants:

For each *matK* barcode, BLAST searches of GenBank were performed. The closest matches in same species and nearest neighbor (NN) of same or different genus were recorded (Table- 4.2). *matK* sequence of sample AUMP16 showed 100% identical with *Aloe vera* and 99% identical with *Aloe compressa*, *Aloe capitata*, etc. 99.9% similarity with *Cynodon dactylon* and 99% similarity with *Cynodon transvaalensis* and *Brachyachne ciliaris* were found against barcode of

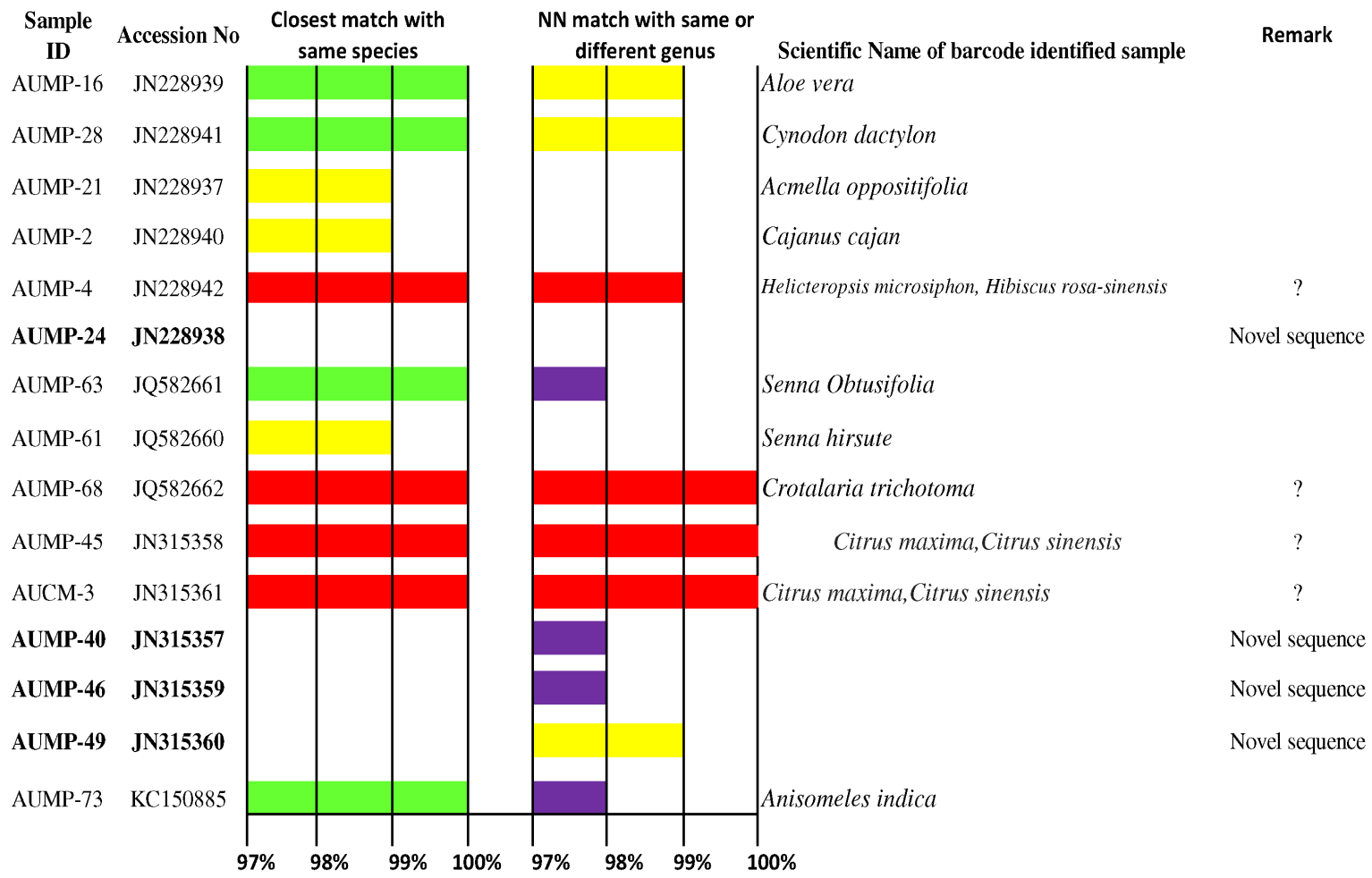


Figure - 4.7 *matK* barcode identification of important medicinal plants. For each sample, sample ID, Accession Number graphical representation of match result and identification are shown. Colour bars depict percentage identity to close match with same species and Nearest Neighbor (NN) in the same or different genus, with scale at bottom

AUMP28 sample. In case of AUMP61, AUMP21 and AUMP2, respectively each sample 99% identical with *Senna hirsuta*, *Acmella oppositifolia* and *Cajanus cajan* and NN of same or different genus was below 97%. From AUMP63 showed 100% identical with *Senna obtusifolia* and NN of same or different genus was 98% with *Senna alata*. Based on BLAST result, we were able to identify all samples except few samples. The *matK* barcode sequence of AUMP4 is 100% identical with *Helicteropsis microsiphon* and *Hibiscus rosa-sinensis*. In case of AUMP45, AUCM3, respectively each sample showed 100% identical with *Citrus maxima, Citrus sinensis.* 100% similarity with *Crotalaria trichotoma and Crotalaria saltiana* was found against barcode of AUMP68. *matK* sequence of sample AUMP73 showed 100% identical with *Anisomeles indica.* The *matK* sequences of *Jasticia adhatoda* (JN228938), *Citrus limon* (JN315357, JN315359) and *Citrus paradisi* (JN315360) are the novel sequences contributed from the study (Figure- 4.7).

4.1.2.2 Identification of Apocynaceae family members:

matK sequence of sample AUMP03, AUMP36, showed 100% identical with *Catharanthus roseus* (Table- 4.2). 98% similarity with *Allamanda schottii* was found against barcode of AUMP29, AUMP33. AUMP05 sample showed 100% similarity with *Alstonia scholaris* and 99% similarity with *Alstonia microphylla.* AUMP01, AUMP92 showed 98% and 99% identical with *Thevetia peruviana.* Barcode sequence from AUMP32 showed 99% identical with *Tabernaemontana divericata.* In case of AUMP34, 100% identical with *Nerium oleander* and 99% identical with *Adenum swazicumw* was found in Genbank.

Sample ID	Accession No	Species Name	Barcode identification	Closest match with same species (%)	NN match with same or different genus (%)
AUMP-03	JN228930	*Catharanthus roseus*	*Catharanthus roseus*	100	97
AUMP-36	JN228936	*Catharanthus roseus*	*Catharanthus roseus*	100	97
AUMP-05	JN228931	*Alstonia scholaris*	*Alstonia scholaris*	100	99
AUMP-01	JN228929	*Thevetia peruviana*	*Thevetia peruviana*	100	96
AUMP-92	JN416982	*Thevetia peruviana*	*Thevetia peruviana*	100	96
AUMP-29	**JN228933**	***Allamanda cathartica***	*Allamanda schottii*	-	98
AUMP-33	**JN228935**	***Allamanda cathartica***	*Allamanda schottii*	-	98
AUMP-32	JN228934	*Tabernaemontana divaricata*	*Tabernaemontana divaricata*	99	99
AUMP-22	**JN228932**	***Calotropis gigantea***	*Asclepias subulata, Asclepias masonii*	-	98
AUMP-34	JN416981	*Nerium oleander*	*Nerium oleander*	100	99
AUMP-16	JN228939	*Aloe vera*	*Aloe vera*	100	99
AUMP-28	JN228941	*Cynodon dactylon*	*Cynodon dactylon*	100	99
AUMP-21	JN228937	*Acmella oppositifolia*	*Acmella oppositifolia*	99	97
AUMP-2	JN228940	*Cajanus cajan*	*Cajanus cajan*	99	97
AUMP-4	JN228942	*Hibiscus rosa-sinensis*	*Hibiscus rosa-sinensis, Helicteropsis microsiphon,*	100	100
AUMP-24	**JN228938**	***Justicia adhatoda***	*Schlegelia parviflora Sesamum indicum*	-	92
AUMP-61	JQ582660	*Senna hirsuta*	*Senna hirsute*	99	97
AUMP-63	JQ582661	*Senns obtusifolia*	*Senna Obtusifolia*	100	98
AUMP-68	JQ582662	*Crotalaria trichotoma*	*Crotalaria trichotoma*	100	100
AUMP-45	JN315358	*Citrus maxima*	*Citrus maxima, Citrus sinensis*	100	100
AUCM-3	JN315361	*Citrus maxima*	*Citrus maxima, Citrus sinensis*	100	100
AUMP-40	**JN315357**	***Citrus limon***	*Citrus sinensis*	-	98
AUMP-46	**JN315359**	***Citrus limon***	*Citrus sinensis*	-	98
AUMP-49	**JN315360**	***Citrus paradisi***	*Citrus maxima*	-	99
AUMP-73	KC150885	*Anisomeles indica*	*Anisomeles indica*	100	98

Table- 4.2 *matK* barcode identification of medicinal plants. For each sample, sample ID, Accession Number, closet match with same species and nearest neighbor (NN) of same or different genus result and identification are given.

The *matK* sequences of *Allamanda cathartica* (JN228933, JN228935) and *Calotropis gigantea* (JN228932) are the novel sequences contributed from the study (Figure- 4.8).

4.1.3 *trnH-psbA* barcode identification:

For each *trnH-psbA* sequence of barcode, BLAST searches of GenBank were performed. The closest matches in same species and nearest neighbor (NN) of same or different genus were recorded (Table-4.3). The *trnH-psbA* sequences of *Catharanthus roseus* (JN245988, JN245984), *Allamanda cathartica* (JN245987), *Thevetia peruviana* (JN245983), *Calotropis gigantea* (JN245986), *Citrus limon* (JN315362, JN315363) *Citrus paradisi* (JN315366, JN315367) and *Citrus limetta* (JN315368, JN315369) are the novel sequences contributed to GenBank. *trnH-psbA* sequence of sample AUMP-03, AUMP-36 showed 88% similarity with *Alstonia macrophylla.* Sample AUMP-05 showed 100% identical with *Alstonia scholaris* and NN of same or different genus was 97% with *Alstonia macrophylla.* In case of AUMP01, 81% identical with *Thevetia ahouai.* 98% similarity with *Allamanda schottii* was found against sample AUMP29. In case of AUMP61, AUMP62, respectively each sample was 99% identical with *Senna hirsuta*, and *Senna occidentalis* and NN of same or different genus was below 97% and 98%. Sample of AUMP66 showed 99% identical to *Senna siamea* and *Senna montana*. *trnH-psbA* sequence of sample AUMP64 showed 100% identical with *Senna occidentalis*. In case of AUMP45, AUCM03, AUMP40, AUMP46, AUMP49, AUMP55, AUMP50, AUMP54, 100% identical sequence of same and different species was found (Figure- 4.9).

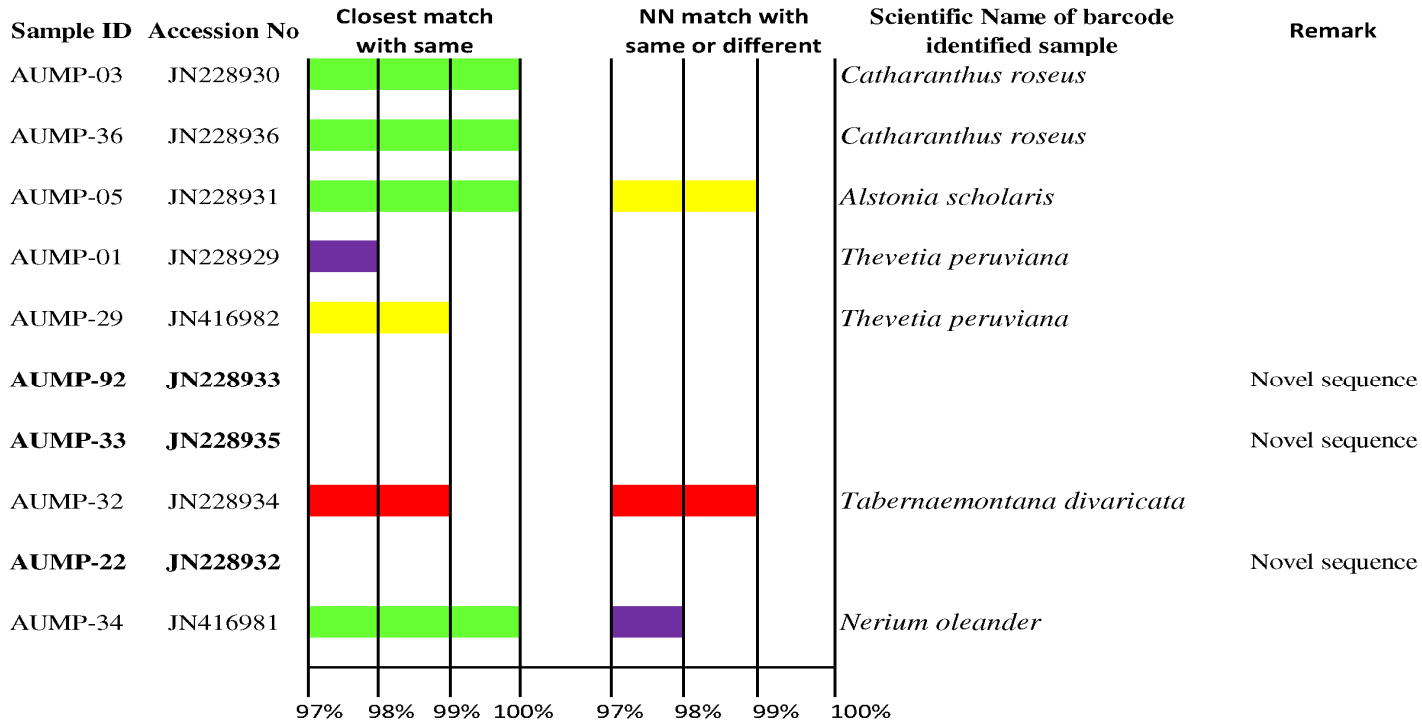


Figure - 4.8 *matK* barcode identification of Apocynaceae family members. For each sample, sample ID, Accession Number graphical representation of match result and identification are shown. Colour bars depict percentage identity to close match with same species and Nearest Neighbor (NN) in the same or different genus, with scale at bottom

Sample ID	Accession No	Species Name	Barcode identification	Closest match with same species (%)	NN match with same or different genus (%)
AUMP-03	**JN245988**	***Catharanthus roseus***	*Alstonia macrophylla*	-	88
AUMP-36	**JN245984**	***Catharanthus roseus***	*Alstonia macrophylla*	-	88
AUMP-05	JN245985	*Alstonia scholaris*	*Alstonia scholaris*	100	
AUMP-01	**JN245983**	***Thevetia peruviana***	*Thevetia ahouai*	-	81
AUMP-29	**JN245987**	***Allamanda cathartica***	*Allamanda schottii*	-	98
AUMP-22	**JN245986**	***Calotropis gigantea***	*Asclepias syriaca*	-	98
AUMP-61	KC150886	*Senna hirsuta*	*Senna hirsuta*	99	98
AUMP-66	KC150889	*Senna siamea*	*Senna siamea, Senna montana*	99	99
AUMP-64	KC150888	*Senns obtusifolia*	*Senna tora*	100	98
AUMP-62	KC150887	*Senna occidentalis*	*Senna occidentalis*	99	
AUMP-45	JN315365	*Citrus maxima*	*Citrus sinensis, Citrus maxima*	100	100
AUCM-3	JN315364	*Citrus maxima*	*Citrus sinensis, Citrus maxima*	100	100
AUMP-40	**JN315362**	***Citrus limon***	*Citrus medica, Citrus sinensis, Citrus maxima*	-	100
AUMP-46	**JN315363**	***Citrus limon***	*Citrus medica, Citrus sinensis, Citrus maxima*	-	100
AUMP-49	**JN315366**	***Citrus paradisi***	*Citrus aurantium, Citrus reticulata*	-	100
AUMP-55	**JN315367**	***Citrus paradisi***	*Citrus aurantium, Citrus reticulata*	-	100
AUMP-50	**JN315368**	***Citrus limetta***	*Citrus sinensis, Citrus maxima*	-	100
AUMP-54	**JN315369**	***Citrus limetta***	*Citrus sinensis, Citrus maxima*	-	100

Table- 4.3 *trnH-psbA* barcode identification of medicinal plants. For each sample, sample ID, Accession Number, closet match with same species and nearest neighbor (NN) of same or different genus result and identification are given.

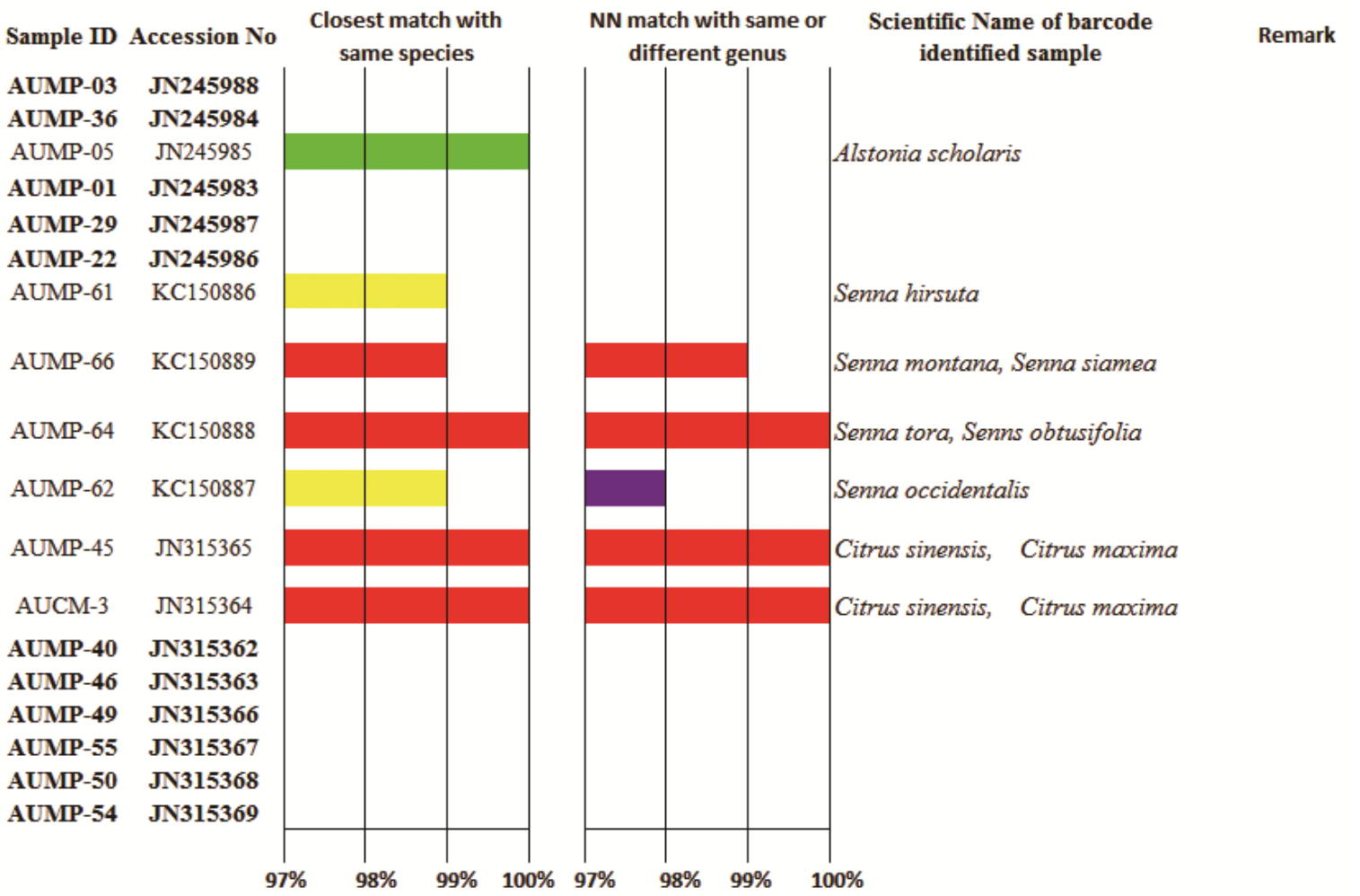


Figure - 4.9 *trnH-psbA* barcode identification. For each sample, sample ID, Accession Number graphical representation of match result and identification are shown. Colour bars depict percentage identity to close match with same species and Nearest Neighbor (NN) in the same or different genus, with scale at bottom.

Chapter-4.2: Evalution of phylogenetic relationship of common medicinal plants with special reference to Apocynaceae family

4.2.1 Phylogenetic relationship of important medicinal plants:

Successful amplification was achieved using a single set of primer for the *matK* barcode sequences (585 bp **-** 872 bp) from selected ethnomedicinal juices and leaves and were used in phylogenetic study in different dataset. We determined the sequences of the *matK* region and identified the studied sample being true to first hand identification based on common names as well as cross checking the common vernacular name according to the Flora of Assam. The species are *Hibiscus rosa-sinensis, Jasticia. adhatoda, Aloe vera, Cynodon dactylon, Acmella oppositifolia, Cajanus cajan, Anisomeles indica, Senna hirsuta, Senna obtusifolia*. Among the 10 determined sequences, the *matK* sequences of *J. adhatoda* were determined for the first time and submitted (Accession No- JN228938). Databases were less well populated with closely related species of *J. adhatoda*. So, we did not include for subsequent analysis because below 92% similar sequence were found in NCBI.

Seven data sets were constructed on the basis of genus and closely related sequences. *Hibiscus, Senna* and *Aloe* are genus dataset. *Acmella*, *Cajanus*, *Cynodon, Anisomeles* are closely related sequences (similarity more than 96%) data set. For *Acmella, Cajanus, Cynodon, Anisomeles* data set contained 8, 7, 9 and 10 sequences respectively. The phylogenetic tree of *Aloe, Senna* and *Hibiscus* dataset constructed on the basis of 15, 11 and 13 sequences. The number of variable site and parsimony informative site were highest (89 and 50) in *Cajanus* dataset and lowest (14 and 13) in *Hibiscus* genus data set (Table- 4.5). The variable and parsimony-informative positions contribute a considerable number of characters for resolving the phylogeny of the medicinal plants.

The different species of *Aloe* genera have formed distinctive clusters. *Aloe* data included, *Aloe vera*, *A. conifera, A. boylei, A. ciliaris, A.aristata, A. plicatilis.* 7 sequences of *Aloe vera* from Genbank and our determined sequence of Sample AUMP16 (JN228939) formed a different cluster in the phylogeny. Others member of *Aloe* also formed different cluster in phylogeny (Figure- 4.10).

Sample ID	Sample type	Common Name in Bengali	Scientific Name (according to Flora of Assam)	Accession No	Sequence Length (bp)
AUMP16	Leaf juice	Ghrita kumari	*Aloe vera*	JN228939	781
AUMP28	Juice	Durba ghash	*Cynodon dactylon*	JN228941	771
AUMP21	Leaf juice	Akarkara	*Acmella oppositifolia*	JN228937	823
AUMP2	Leaf juice	Arhar	*Cajanus cajan*	JN228940	872
AUMP4	Young Leaf	Jaba	*Hibiscus rosa-sinensis*	JN228942	585
AUMP73	Young Leaf	Apang	*Anisomeles indica*	KC150885	741
AUMP61	Young Leaf	Swarnapatri	*Senna hirsuta*	JQ582660	639
AUMP63	Young Leaf	Chakunda	*Senna obtusifolia*	JQ582661	634
AUMP68	Young Leaf	-	*Crotalaria trichotoma*	JQ582662	645
AUMP16	Young Leaf	Basak	*Jasticia adhatoda*	JN228938	836

Table- 4.4 List of selected ethnomedicine with their sample ID, sample type, common and scientific name (according to Flora of Assam), Accession Number of sequences submitted to NCBI also given.

Similarly, AUMP4 formed a different cluster of *Hibiscus rosa-sinensis* in the *Hibiscus* data set phylogeny. *H. Mandrarensis* and *H. bojerianus* also formed a different cluster. However, sequences of *Hibiscus etatus, H. tiliaceus* and *Hibiscus pernambucensis* formed one cluster irrespective though they are different species (Figure 4.11). *Senna* data set included *Senna hirsuta*, *Senna pedula, Senna bauhinioides*, *Senna alata*, *Senna obtusifolia* and *Crotalaria*

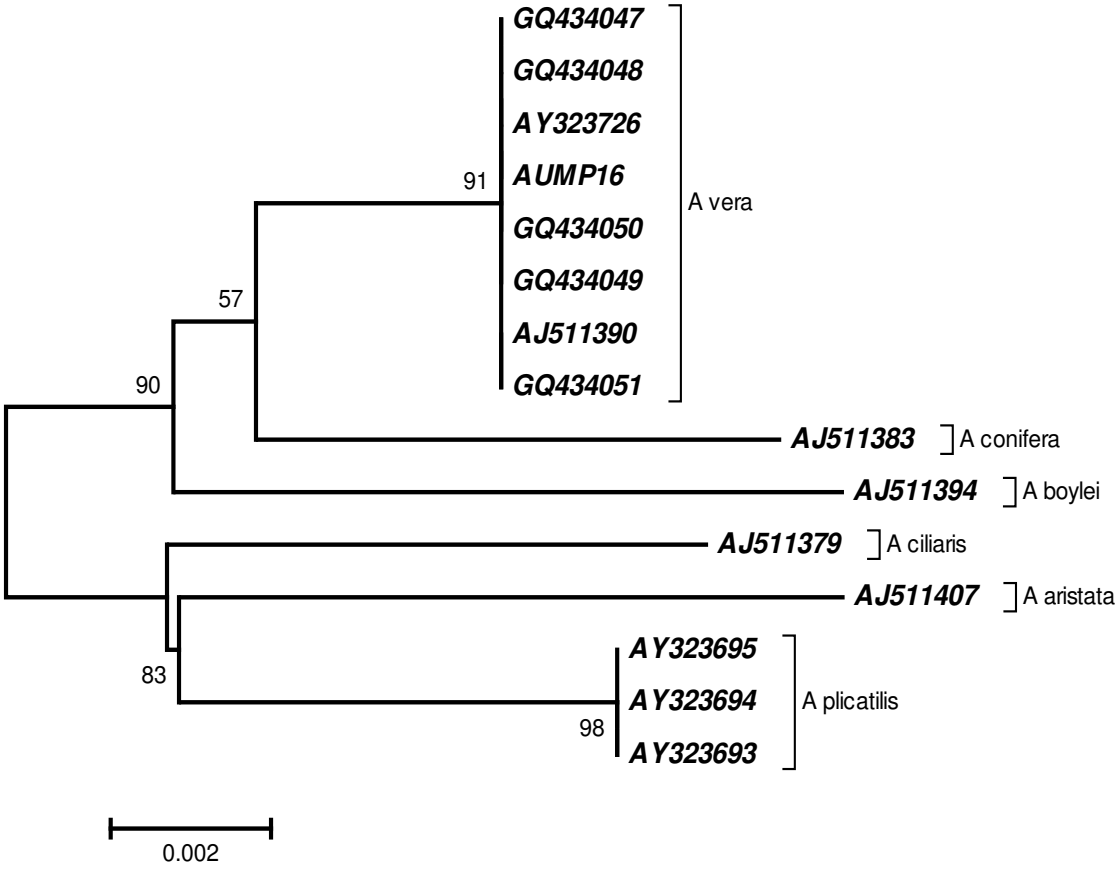


Figure - 4.10 Phylogenetic tree from partial *matK* (767bp) of six *Aloe* species. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.002 nucleotide substitution per site.

trichotoma. We generated *matK* sequence of *S. hirsuta* (JQ582660), S. *obtusifolia* (JQ582661) and *Crotalaria trichotoma* (JQ582662). The different species of *Senna* dataset were formed distinctive clusters. Evidently all the database sequences and the conspecific generated sequences of *S. hirsuta*, *S. obtusifolia* with Genbank accession numbers were clustered cohesively. However the *Crotalaria trichotom*a was located at the basal position hence used as an out group of the phylogenetic tree (Figure- 4.12).

Data Set	**No of Sequence**	**Analyzed Sequence length (bp)**	**No of Variable site**	**No of parsimony Informative site**
Hibiscus	10	585	14	13
Aloe	15	767	47	13
Acmella	9	823	57	16
Cajanus	7	771	89	50
Cynodon	9	872	45	15
Anisomeles	13	740	85	72
Senna	11	645	54	35

Table- 4.5 Seven data sets constructed on the basis of genus and closely related sequences. *Hibiscus* and *Aloe* are genus dataset. *Acmella*, *Cajanus*, *Cynodon, Anisomeles* are closely related sequence data set. Number of sequences, analyzed, sequence length, Number of variable site(s) and parsimony informative site of each data set are in table.

In case of *Acmella, Cajanus, Cynodon, Anisomeles* dataset, very few numbers of highly similar sequences or close member of such genus sequences were available in Genbank. *C. dactylon, A. oppositifolia, C. cajan, A. indica*

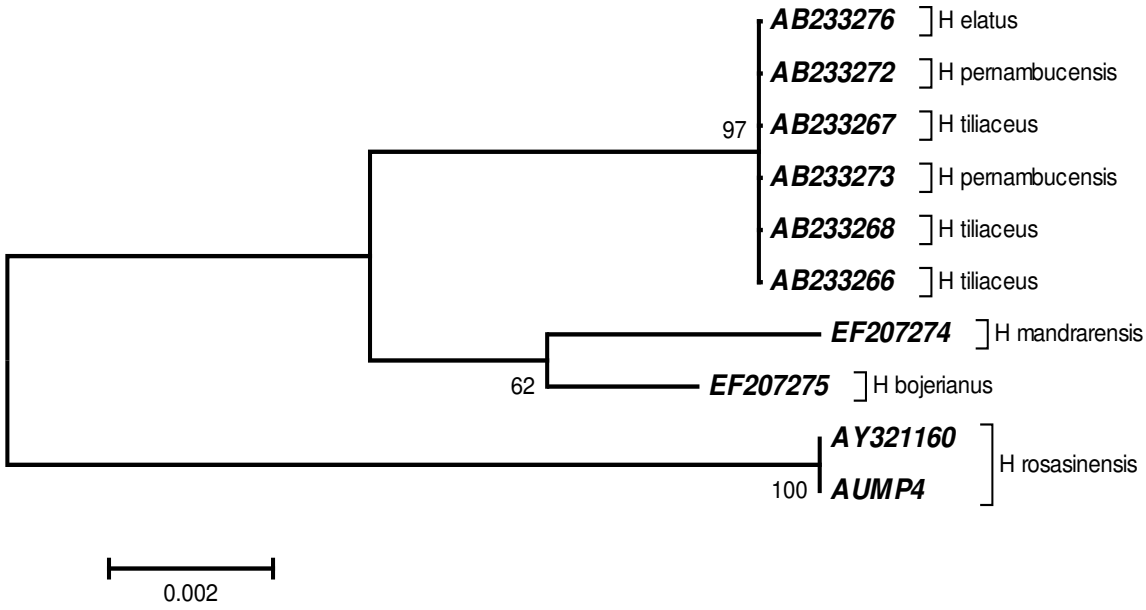


Figure -4.11 Phylogenetic tree from partial *matK* (585 bp) of seven *Hibiscus* species. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.002 nucleotide substitution per site.

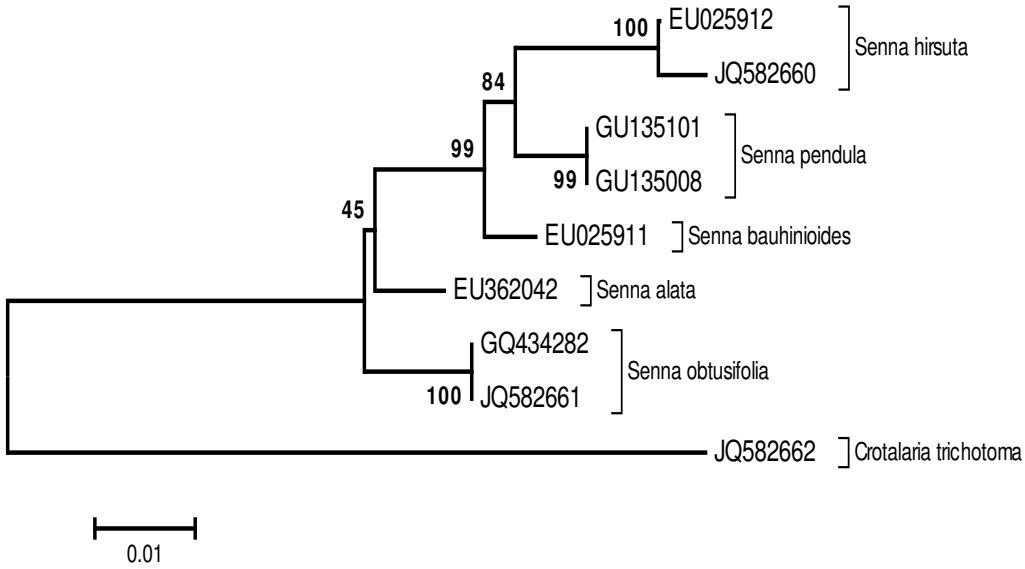


Figure -4.12 Phylogenetic tree from partial *matK* (645 bp) of five *Senna* species. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.002 nucleotide substitution per site.

formed a different cluster in respective the phylogeny. *Acmella* dataset included *A. oppositifolia, Viguiera tomentoda, Otopappus epaleaceus, Blainvillea rhomboidea, Heterosperm pinnatum, Perityle lindheimeri, Eutrochium maculatum, Critonia sexangulari.* GenBank sequences of *A. oppositifolia* (FJ697074) and our determined sequence of Sample AUMP21 (JN228937) formed a different cluster in the phylogeny (Figure- 4.13).

In Cajanus data set, *Cajanus cajan, Bolusafra bituminosa, Apios amencana, Kennedia nigricans, Pueraria candollei* were included based on similarity search in GenBank. *matK* sequence from sample AUMP 2 (JN228940) formed distinctive clusters with conspecific sequence (EU717414) of database (Figure- 4.14).

Cynodon data set included *Cynodon dactylon, Brachyachhe cillaris, Cynodon transvaalensis, Chloris canterai, Tetrapogon tenellus, Triodia scariosa, Oxychloris scariosa, Lepturus repens.* We generated *matK* sequence of *C. dactylon* (JN228941) and the conspecific Genbank sequences (AF144584) were clustered cohesively. The different species of *Cynodon* dataset were formed distinctive clusters (Figure- 4.15).

Anisomeles indica and its different closely related species have formed distinctive clusters. Evidently, the database sequences (FJ513162) and the conspecific generated sequences of *Anisomeles indica* (KC150885) are clustered cohesively. Similarly, all the database sequences of *Betonica Officinales, Pogostemen cablin, Leucosceptrum canum* are clustered cohesively (Figure- 4.16). Our study showed that species identification of the entire sample is possible using phylogenetic analyses constructed from *matK* sequences.

4.2.1.1 Phylogenetics of *Citrus* based on *matK* and *trnH-psbA*:

In this study, We determined the 5 sequences of the *matK* region and 8 sequence of *trnH-psbA* intergenic spacer of *Citrus* species (Table- 4.6) The species are

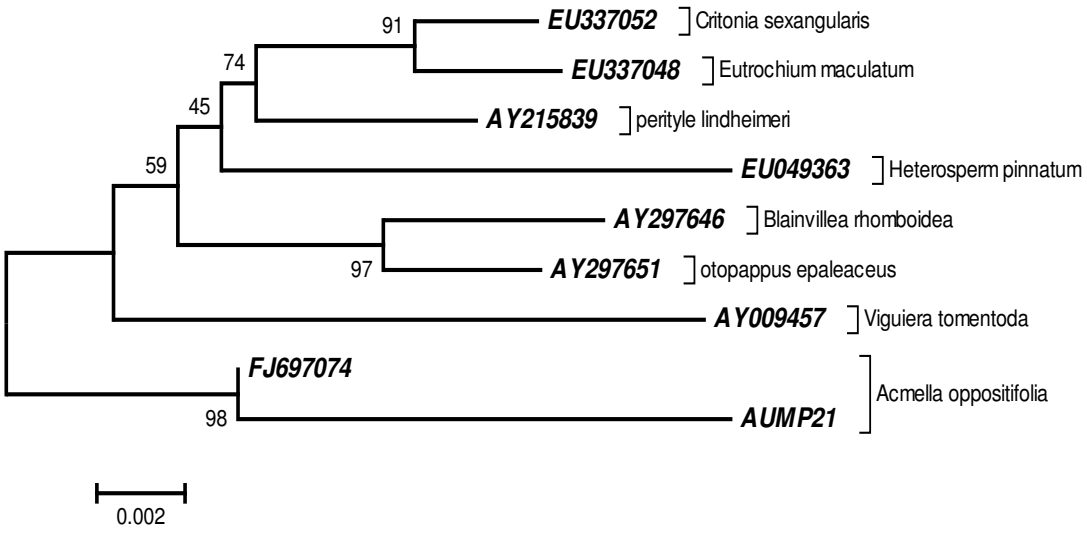


Figure -4.13 Phylogenetic tree from partial *matK* (823 bp) of eight closely related species of *Acmella oppositifolia*. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.002 nucleotide substitution per site.

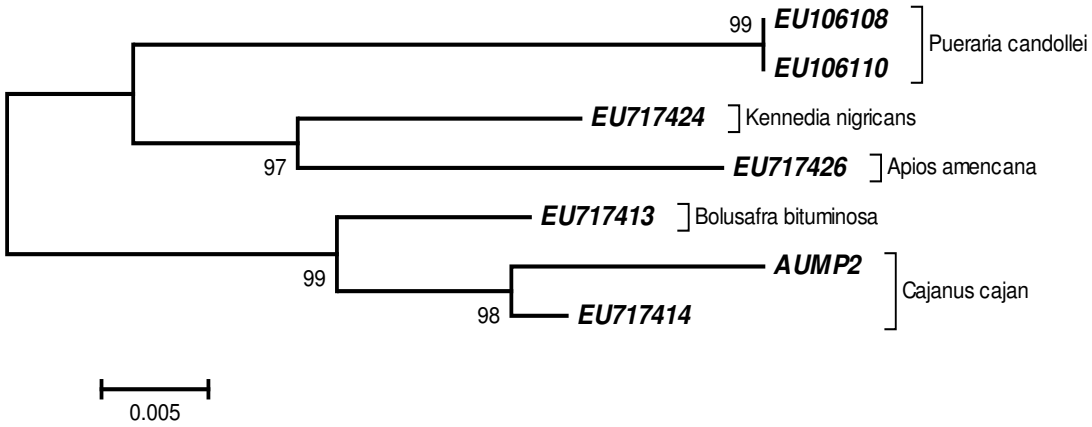


Figure- 4.14 Phylogenetic tree from partial *matK* (771 bp) of five closely related species of *Cajanus cajan*. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.005 nucleotide substitution per site.

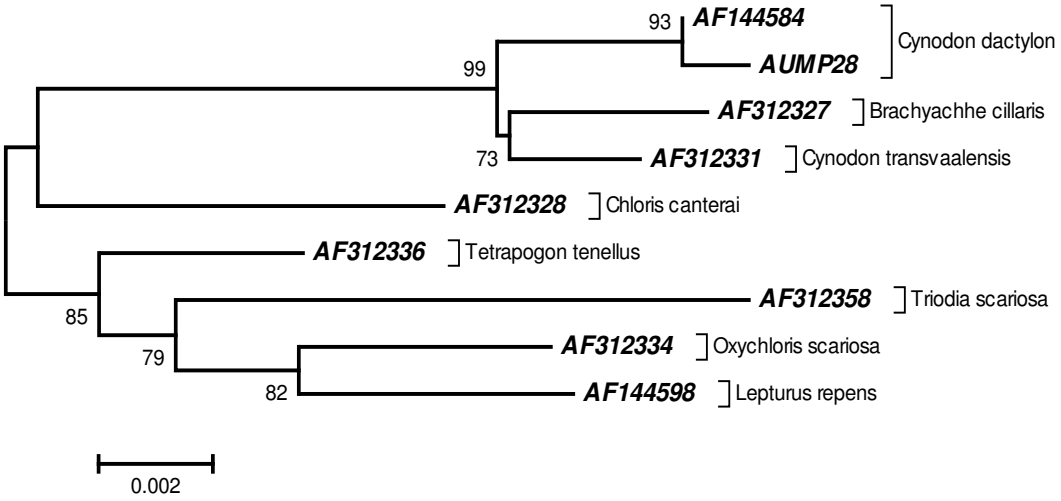


Figure - 4.15 Phylogenetic tree from partial *matK* (872 bp) of eight closely related species of *Cynodon dactylon*. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.002 nucleotide substitution per site.

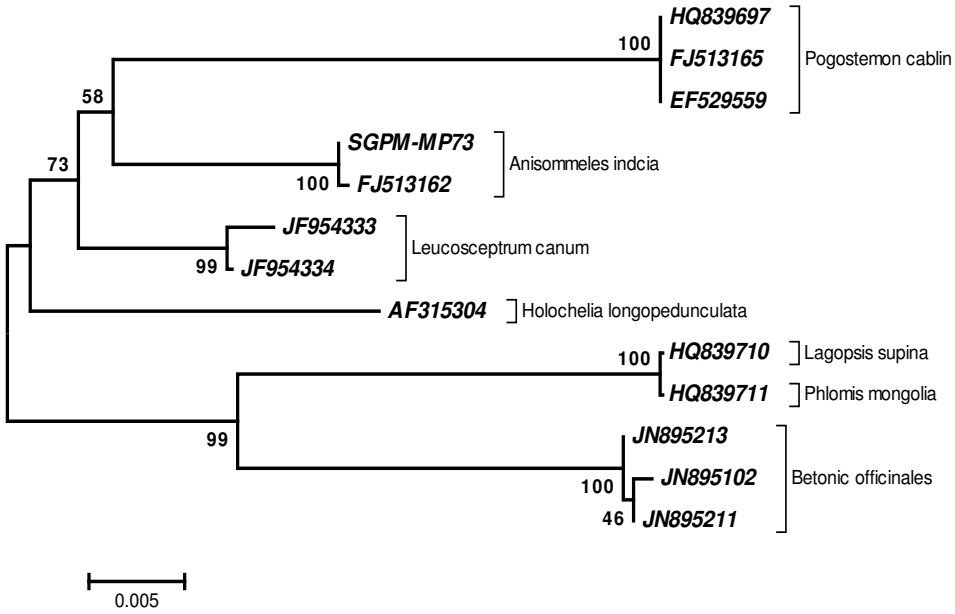


Figure -4.16 Phylogenetic tree from partial *matK* (741 bp) of seven closely related species of *Anisomeles indica*. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.002 nucleotide substitution per site.

Citrus maxima, Citrus limon, Citrus paradisi ,*Citrus limetta*. Among the 8 determined sequences of *trnH-psbA* sequences, *Citrus limetta* were determined for the first time and submitted (JN315368, JN315369). In case of *matK* sequences *Citrus limon* (JN315357, JN315359), *Citrus paradisi* (JN315359) were determined for the first time and submitted to GenBank (Table- 4.6). Very few no of *matK* sequence of *Citrus* is available in Genbank. Due to length variations, 791 bp of *matK* and ~452 bp of *trnH-psbA,* aligned sequences were used in phylogenetic analysis and low variations were found in both the *matK* and *trnH-psbA* of *Citrus* species.

Voucher Number	**Common Name**	**Scientific Name**	**Accession No *matK***	**Accession No *trnH-psbA***
AUMP45	Robab tenga	*Citrus maxima*	JN315358	JN315365
AUCM3	Robab tenga	*Citrus maxima*	JN315361	JN315364
AUMP49	Nebu tenga	*Citrus paradisi*	JN315360	JN315366
AUMP55	Nebu tenga	*Citrus paradisi*	-	JN315367
AUMP40	Jora tenga	*Citrus limon*	JN315357	JN315362
AUMP46	Jora tenga	*Citrus limon*	JN315359	JN315363
AUMP50	Mitha nimbu	*Citrus limetta*	-	JN315368
AUMP54	Mitha nimbu	*Citrus limetta*	-	JN315369

Table- 4.6 List of *Citrus* species with voucher number, common name in local language (Assamess) and scientific name and accession number of sequences in GenBank also given.

Mean pair wise distance was computed using K2P to check the interspecies divergence between all species among the members of the genus

Citrus based on *matK*. A record was made to discriminate the intra and interspecies variation in *Citrus* genera. The maximum mean intra-specific distance was found in *Citrus limon* i.e. 0.023 and minimum in *Citrus maxima* and *Citrus trifoliata* (0.00). The maximum inter-specific distance was found between *Citrus limon and Murraya exotica* (0.048) and other *Citrus* species also showed high distance with *Aegle marmelos* and *Murraya exotica*. Minimum inter-specific distance was found between *Citrus reticulate and Citrus maxima* (0.004). Inter specific divergence of being low to discriminate the *Citrus* species (Table- 4.7).

	Species Name	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	***Citrus reticulata***	**0.001**						
2	***Citrus paradisi***	0.005	**n/c**					
3	***Citrus trifoliata***	0.004	0.008	**0.000**				
4	***Citrus limon***	0.022	0.026	0.023	**0.023**			
5	***Murraya exotica***	0.027	0.032	0.029	0.048	**n/c**		
6	***Aegle marmelos***	0.021	0.026	0.023	0.040	0.030	**n/c**	
7	***Citrus maxima***	0.001	0.004	0.004	0.022	0.027	0.020	**0.000**

Table- 4.7 Mean divergence (K2P) within (bold number on diagonal) and among (below diagonal) the six species of *Citrus* species based on *matK* (n/c indicates comparable due to only one accession number).

The phylogenetic tree was constructed on the basis of 15 partial *matK* sequences belonging to 5 taxa of the genus *Citrus* (*C. trifoliata*, *C. limon*, *C. paradisi, C. maxima* and *C. reticulate*). *C. trifoliata* and *C. limon* were formed distinctive clusters. All the database sequences and conspecific generated sequences of *C. paradisi, C. maxima* and *C. reticulate* were clustered into one clade. The member of Ruatceae family, *Aegle marmelos* and *Murraya exotica*

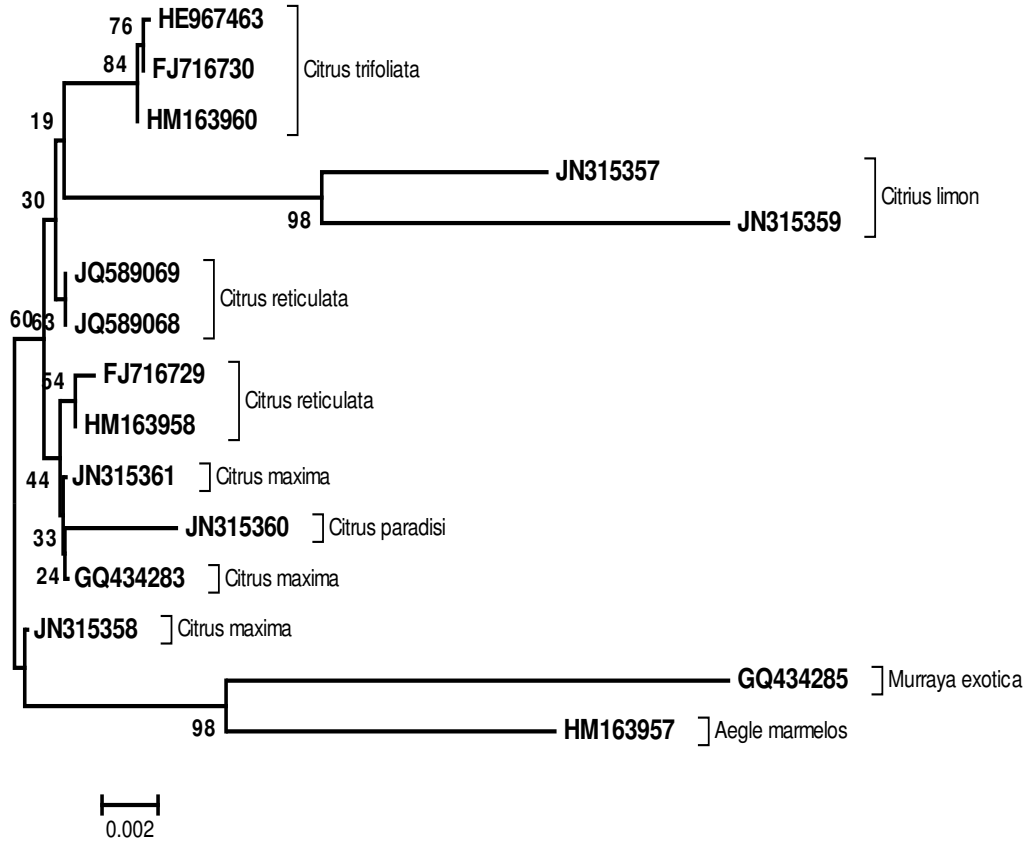


Figure - 4.17 Phylogenetic tree from partial *matK* (764 bp) of seven *Citrus* species. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.002 nucleotide substitution per site.

located at basal position and used as the out group of the phylogentic tree (Figure-4.17).

To evaluate the degree of DNA polymorphism, sequence divergence based on *trnH-psbA* sequence between and within 8 species of the *Citrus* were calculated by Kimura 2 parameters that revealed maximum intra specific distance in *Citrus maxima* (0.0065) and minimum in *C. medica*, *C. limetta* and *C. trifoliate*. In some cases, we found intra specific distances were higher than inter specific distance. The highest inter specific distance was 0.0262 between *C. maxima* and *C. trifoliata*. Other species also showed high distance with *C. trifoliata* (Table- 4.8). The accuracy of DNA barcoding depends on the barcode gap between intra and inter specific variation. Here, sequence variation between species has not to be enough to tell them apart.

	Species Name	1	2	3	4	5	6	7	8
1	***C. maxima***	**0.0065**							
2	***C. limetta***	0.0036	**0.0000**						
3	***C. aurantium***	0.0123	0.0086	**0.0024**					
4	***C. reticulata***	0.0123	0.0086	0.0012	**0.0024**				
5	***C. paradisi***	0.0206	0.0172	0.0146	0.0146	**0.0023**			
6	***C. limon***	0.0183	0.0144	0.0104	0.0104	0.0205	**0.0017**		
7	***C. medica***	0.0190	0.0152	0.0111	0.0111	0.0212	0.0008	**0.0000**	
8	***C. trifoliata***	0.0262	0.0222	0.0160	0.0160	0.0259	0.0183	0.0199	**0.0000**

Table- 4.8 Intraspecific and intraspecific variation in K2P distances of *Citrus* species based on *trnH-psbA*. Mean divergence (K2P) within (bold number on diagonal) and among (below diagonal). (n/c indicates not comparable due to only one accession number).

In this study, we generated 8 sequences of *trnH-psbA* intergenic spacer of *Citrus* species. Total 35 partial *trnH-psbA* sequences belonging to 8 taxa from the

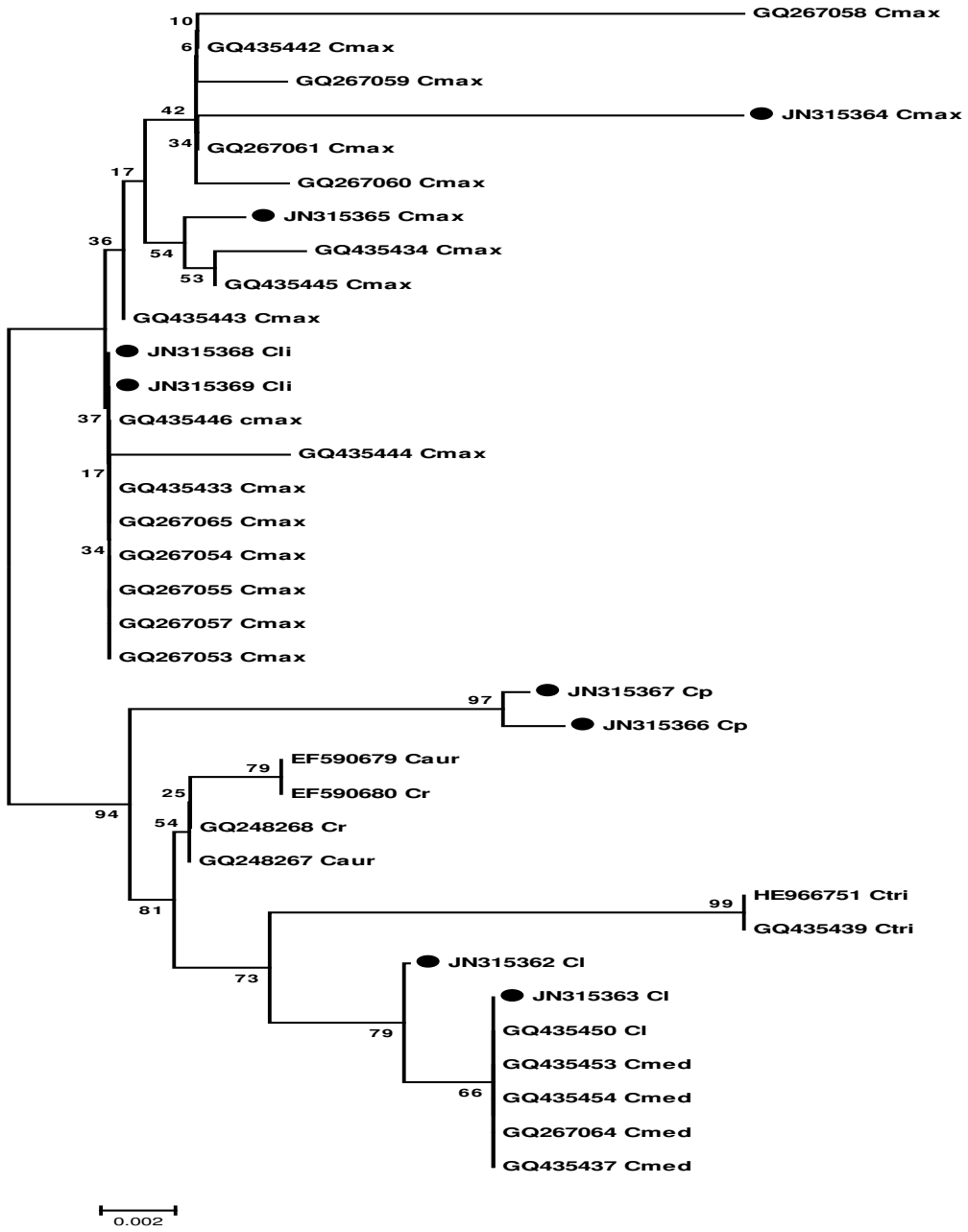


Figure - 4.18 Phylogenetic tree from *trnH-psbA intergenic spacer* (415 bp) of seven *Citrus* species. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.002 nucleotide substitution per site. Abbreviations of *Citrus* species are Cmax- *Citrus maxima*, Cli-*Citrus limetta*, Cl- *Citrus limon*, Cp-*Citrus paradisi*, Cr-*Citrus reticulata*, Cmed- *Citrus medica*, Ctri-*Citrus trifoliate.*

- Generated sequence from this study

genus *Citrus* (*C. paradisi, C. trifoliat, C. medica, C. limon*, *C. aurantium, C. reticulata, C. maxima* and *C. limetta*) were included for phylogenetic analysis and distance calculation. However, GenBank sequences previously submitted by

several authors were also included in the phylogenetic analysis as reference sequences. An elaborated phylogenetic tree of members of genus *Citrus* was constructed using Mega 4.1. All the database sequences and conspeific generated sequences of *C. paradisi* and *C. trifoliate* formed a distant cluster. But *C. maxima* and *C. limetta* were shared a common cluster. Similarly, *C. medica* and *C. limon*, *C. aurantium and C. reticulate* were also clustered into a clade (Figure- 4.18).

4.2.2 Sequence and Phylogenetics of medicinal plant from family Apocynaceae based on *matK* and *trnH-psbA*:

In this study, we unraveled 8 sequences of the *matK* region and 6 sequences of *trnH-psbA intergenic spacer* of important medicinal plants of the Apocynaceae. However, GenBank sequences previously submitted by several authors were also included in the phylogenetic analysis as reference sequences. Total 28 partial *matK* sequences were belonging to 14 taxa from subfamily Rauvolfioideae and 2 taxa from subfamily Asclepiadaceae of Apocynaceae (Table- 4.9). Among the 8 determined sequences, the *matK* sequences of *Allamanda cathartica* and *Calotropis gigantea* were determined for the first time and submitted (JN228933, JN228935 and JN228932). Very few numbers of highly similar sequences of *trnH-psbA* intergenic spacer of this family were available in Genbank. Among the 6 determined sequences, *trnH-psbA* sequences of *Allamanda cathartica* (JN245987), *Catharanthus roseus* (JN245988, JN24584), *Thevetia peruviana* were novel sequences in Genbank.

4.2.2.1 Nucleotide and amino acid composition:

Nucleotide composition of *matK* sequences of Apocynaceae was shown in Table-4.10. It showed that the *matK* sequences had a strong A+T bias (average 65.6% of

codon position, highest A+T bias (68.6 %) was found and *C. roseus* had maximum (70.5 %) and minimum (66.2%) in *A. curassavica.* The almost equal rates of substitution among the three codon positions (Figure- 4.19). Partial *matK* amino acid sequences (~204 aa) of different species of Apocynaceae were analyzed and the standard deviation (SD) of this composition was calculated as a measure of variation. Amino acids were categorized as nonpolar (A, G, V, L, I, P, F, M, W, C), uncharged polar at pH 7 (N, Q, S, T, Y), basic acid (K, R, H), acidic (D, E), aromatic (F, W, Y) or special (C, G, P). On the basis of this amino acids categories, the partial *matK* amino acid sequences were highly nonpolar (more than 50%) and low acidic amino acid more than 5% (Table- 4.11 and Figure 4.20).

Species	**Subfamily**	**Sample ID**	**Accession No. of *matK***	**Accession No. of *trnH-psbA***
Catharanthus roseus	Rauvolfioideae	AUMP-03 AUMP-36	JN228930, JN228936	JN245988*, JN245984*
Thevetia peruviana	Rauvolfioideae	AUMP-01	JN228929,	JN245983*
Allamanda cathartica	Rauvolfioideae	AUMP-29 AUMP-33	JN228933*, JN228935*	JN245987*
Tabernaemontana divaricata	Rauvolfioideae	AUMP-32	JN228934	-
Calotropis gigantea	Asclepiadeae	AUMP-22	JN228932*	JN245986*
Alstonia scholaris	Rauvolfioideae	AUMP-05	JN228931	JN245985

Table- 4.9 List of Plant samples of Apocynaceae examined in this study scientific name, subfamily, Voucher, Accession Number of sequences of *matK* and *trnH-psbA* also given.

* Sequence submitted first time in GenBanK.

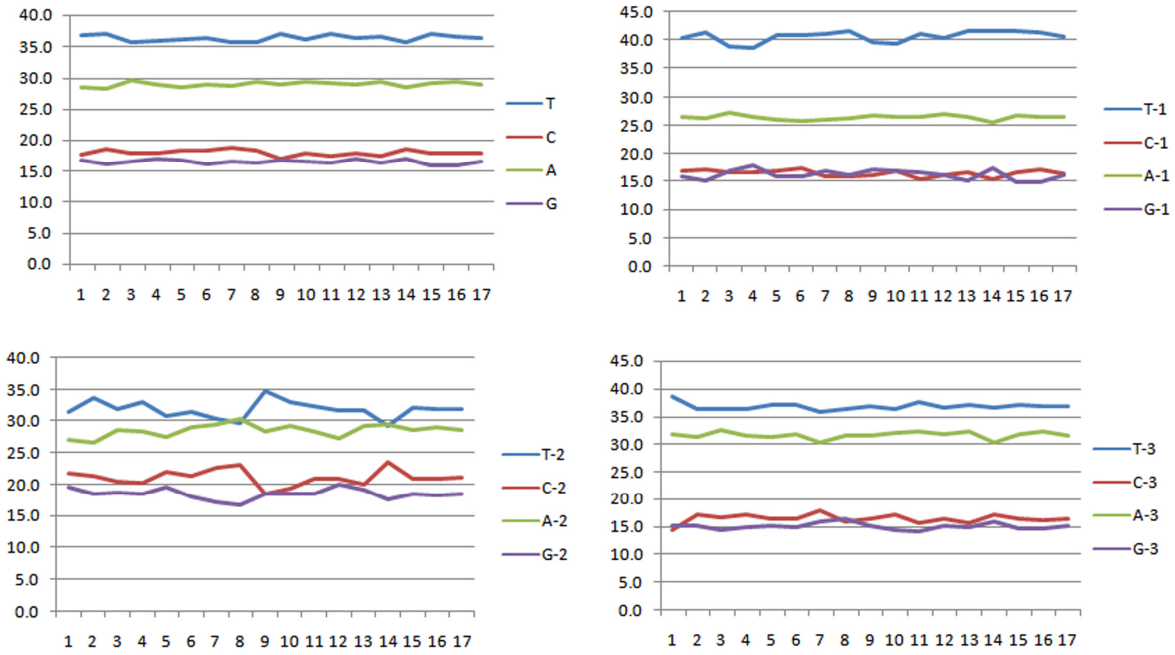


Figure - 4.19 Nucleotide compositions of ~758bp partial *matK* for the different species of Apocynaceae plants. The frequencies of nucleotide in sequences are presented as the total average value for the all the codon positions and for each codon position separately with the accuracy to tenths of a percent. (A, T, G, C shown average value for all codon positions. A-1, T-1, G-1, C-1 shown average value for first codon position. A-2, T-2, G-2, C-2 shown average value for second codon position. A-3, T-3, G-3, C-3 shown average value for third codon position. A+T, A1+T1, A2+T2, A3+A3 represent the average value of A+T bias of total and each codon position

Species	T	C	A	G	A+T	T-1	C-1	A-1	G-1	A1+T1	T-2	C-2	A-2	G-2	A2+T2	T-3	C-3	A-3	G-3	A3+T3
C. roseus	36.9	17.7	28.5	16.9	65.4	40.4	17.1	26.6	15.9	67.0	31.6	21.7	27.1	19.6	58.7	38.6	14.3	31.9	15.1	70.5
V. minor	37.1	18.5	28.2	16.3	65.3	41.3	17.1	26.4	15.2	67.7	33.7	21.2	26.7	18.4	60.4	36.3	17.1	31.4	15.1	67.7
A. microphylla	36.3	18.5	28.4	16.9	64.7	40.9	17.1	26.2	15.9	67.1	30.8	22.0	27.6	19.6	58.4	37.1	16.3	31.5	15.1	68.5
A. scholaris	36.5	18.3	29.0	16.2	65.5	40.9	17.5	25.8	15.9	66.7	31.6	21.2	29.2	18.0	60.8	37.1	16.3	31.9	14.7	68.9
A. curassavica	35.8	18.8	28.7	16.7	64.5	41.1	15.9	26.0	17.1	67.1	30.5	22.7	29.7	17.2	60.2	35.8	17.9	30.4	16.0	66.2
C. gagantia	35.8	18.3	29.4	16.5	65.2	41.5	15.9	26.4	16.3	67.9	29.7	23.0	30.5	16.8	60.2	36.2	16.0	31.5	16.3	67.7
A. cathartica	37.1	17.0	29.0	16.9	66.1	39.5	16.3	27.0	17.3	66.5	34.8	18.4	28.4	18.4	63.2	36.9	16.3	31.7	15.1	68.6
A. schotti	36.2	17.8	29.4	16.6	65.6	39.3	17.1	26.6	17.1	65.9	32.9	19.3	29.3	18.5	62.2	36.3	17.1	32.3	14.3	68.6
T. peruviana	37.0	17.4	29.1	16.5	66.1	41.0	15.6	26.5	16.8	67.5	32.3	20.8	28.4	18.5	60.7	37.7	15.7	32.4	14.1	70.1
T. ahouai	36.3	17.8	28.8	17.1	65.1	40.2	16.3	27.2	16.3	67.4	31.8	20.8	27.3	20.0	59.1	36.7	16.3	31.8	15.1	68.5
T. divericata	36.7	17.4	29.4	16.4	66.1	41.4	16.8	26.6	15.2	68.0	31.6	19.9	29.3	19.1	60.9	37.1	15.6	32.4	14.8	69.5
N. oleander	36.6	18.0	29.4	16.0	66.0	41.2	17.2	26.6	15.0	67.8	31.9	20.7	29.1	18.3	61.0	36.8	16.1	32.5	14.7	69.3
Average	36.4	18.0	29.0	16.6	65.4	40.6	16.6	26.5	16.3	67.1	31.9	21.0	28.6	18.5	60.5	36.8	16.4	31.8	15.0	68.6

Table- 4.10 Nucleotide compositions of ~758bp partial *matK* for the different species of Apocynaceae plants. The frequencies of nucleotide in sequences are presented as the total average value for the all the codon positions and for each codon position separately with the accuracy to tenths of a percent.

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
DQ660536	3.45	1.48	3.45	3.94	9.85	2.46	3.45	5.42	7.88	13.30	2.96	6.40	3.94	3.45	4.93	7.88	2.46	5.42	2.46	5.42
AUMP5	2.96	1.97	2.96	4.93	9.85	1.97	3.45	5.42	6.90	14.29	2.96	6.90	3.45	2.96	5.42	8.87	2.46	4.93	1.97	5.42
Z70191	3.45	1.48	3.45	3.94	10.34	2.46	3.45	5.42	7.88	13.30	2.46	6.40	3.94	3.45	4.93	7.88	2.46	5.42	2.46	5.42
Z70187	3.32	1.42	3.32	3.32	10.90	3.32	2.84	6.16	8.06	12.80	1.90	7.58	2.84	3.32	4.74	8.53	2.37	5.69	1.90	5.69
AM295076	2.96	0.99	2.96	3.94	11.33	3.45	2.46	4.43	5.91	13.30	1.97	7.39	2.96	4.43	5.42	8.87	2.96	5.91	2.46	5.91
DQ660548	3.32	1.42	2.84	3.32	10.90	3.79	2.84	5.69	7.58	13.27	1.90	8.06	3.32	3.32	5.21	8.06	2.37	5.69	1.90	5.21
DQ660550	3.33	1.43	2.86	3.33	10.95	3.33	2.86	6.19	7.62	12.86	1.90	8.10	2.38	3.81	4.76	8.57	2.38	5.71	1.90	5.71
Z70188	2.99	1.99	3.48	3.48	9.95	1.99	1.99	5.47	6.97	13.93	1.49	7.96	3.48	3.48	6.97	7.96	2.49	6.47	1.99	5.47
AUMP1	3.45	1.97	2.96	3.94	10.34	1.97	2.96	5.91	6.90	13.30	1.48	7.39	3.45	3.45	6.40	7.39	2.46	6.40	1.97	5.91
Z70189	2.96	1.97	2.96	4.93	9.85	1.97	3.45	5.42	6.90	14.29	2.96	6.90	3.45	2.96	5.42	8.87	2.46	4.93	1.97	5.42
FJ449631	2.96	1.97	2.96	4.93	9.85	1.97	3.45	5.42	6.90	14.29	2.96	6.90	3.45	2.96	5.42	8.87	2.46	4.93	1.97	5.42
GU135061	3.45	0.99	3.45	3.94	10.84	2.96	3.45	4.93	6.40	13.79	1.48	6.90	2.96	3.94	5.42	8.87	2.46	6.40	2.46	4.93
DQ660495	3.47	1.49	3.47	3.96	10.40	2.48	3.47	4.95	7.43	12.87	3.47	6.93	3.47	3.47	4.95	7.92	2.48	5.45	2.48	5.45
AUMP36	2.96	0.99	3.45	3.94	11.33	3.45	2.96	5.42	6.40	13.79	1.97	7.88	2.96	2.96	5.42	7.39	1.48	6.40	2.96	5.91
AUMP32	2.84	1.42	3.32	3.32	10.90	3.32	3.32	6.16	7.58	13.27	2.37	7.58	2.84	3.32	5.21	8.53	1.90	5.69	1.90	5.21
AUMP33	2.96	1.48	3.45	3.94	11.82	2.96	2.96	5.42	6.40	12.32	2.96	6.90	2.96	3.94	5.42	8.37	2.46	5.42	2.46	5.42
AUMP34	2.96	1.48	3.45	3.94	12.32	2.96	2.96	5.42	6.40	12.32	2.96	6.90	2.96	3.94	5.42	7.88	2.46	5.42	2.46	5.42
EF456295	3.94	1.97	2.46	3.45	9.85	2.46	3.94	5.91	6.40	14.29	2.46	7.39	2.96	3.94	5.42	6.90	2.96	4.93	1.97	6.40
DQ660553	2.96	0.99	2.96	3.94	11.33	3.45	3.45	4.43	5.91	13.30	1.97	7.39	2.96	3.94	5.42	8.87	2.96	5.91	2.46	5.42
AM295068	2.96	0.99	3.45	3.94	11.33	3.45	2.96	5.42	5.91	13.79	1.97	7.88	2.96	3.45	5.42	7.39	1.48	6.40	2.96	5.91
GQ997641	3.94	1.97	2.46	3.45	9.85	2.46	3.94	5.91	6.40	14.29	2.46	7.39	2.96	3.94	5.42	6.90	2.96	4.93	1.97	6.40
DQ026716	3.83	1.44	2.39	2.87	9.09	2.87	3.83	5.74	7.66	14.35	1.91	6.22	3.35	2.87	7.18	8.61	3.83	4.78	1.44	5.74
AUMP29	2.96	1.48	3.45	3.94	11.82	2.96	2.96	5.42	6.40	12.32	2.96	6.90	2.96	3.94	5.42	8.37	2.46	5.42	2.46	5.42
AUMP22	2.39	2.39	2.39	3.35	8.61	2.39	3.83	5.74	7.66	14.35	1.91	6.22	3.35	2.87	7.18	7.66	4.31	5.74	1.44	6.22
AUMP3	2.96	0.99	3.45	3.94	11.33	3.45	2.96	5.42	6.40	13.79	1.97	7.88	2.96	2.96	5.42	7.39	1.48	6.40	2.96	5.91
DQ660507	2.96	0.99	3.45	3.94	11.33	3.45	2.96	5.42	6.40	13.79	1.97	7.88	2.96	2.96	5.42	7.39	1.48	6.40	2.96	5.91
GU135060	3.45	0.99	3.45	3.94	10.84	2.96	3.45	4.93	6.40	13.79	1.48	6.90	2.96	3.94	5.42	8.87	2.46	6.40	2.46	4.93
AJ429321	2.96	1.97	2.96	4.93	9.85	1.97	3.45	5.42	6.90	14.29	2.96	6.90	3.45	2.96	5.42	8.87	2.46	4.93	1.97	5.42
DQ660506	3.45	1.97	2.96	3.94	10.34	1.97	2.96	5.91	6.90	13.30	2.96	7.39	2.96	2.96	5.91	7.88	2.46	5.42	2.46	5.91
Average	**3.19**	**1.52**	**3.12**	**3.88**	**10.60**	**2.78**	**3.21**	**5.48**	**6.88**	**13.55**	**2.31**	**7.22**	**3.16**	**3.44**	**5.53**	**8.13**	**2.48**	**5.65**	**2.24**	**5.62**

Table- 4.11 Amino Acid compositions of ~204aa partial *matK* protein for the different species of Apocynaceae plants. The frequencies of amino acid in sequences are presented as the total average value.

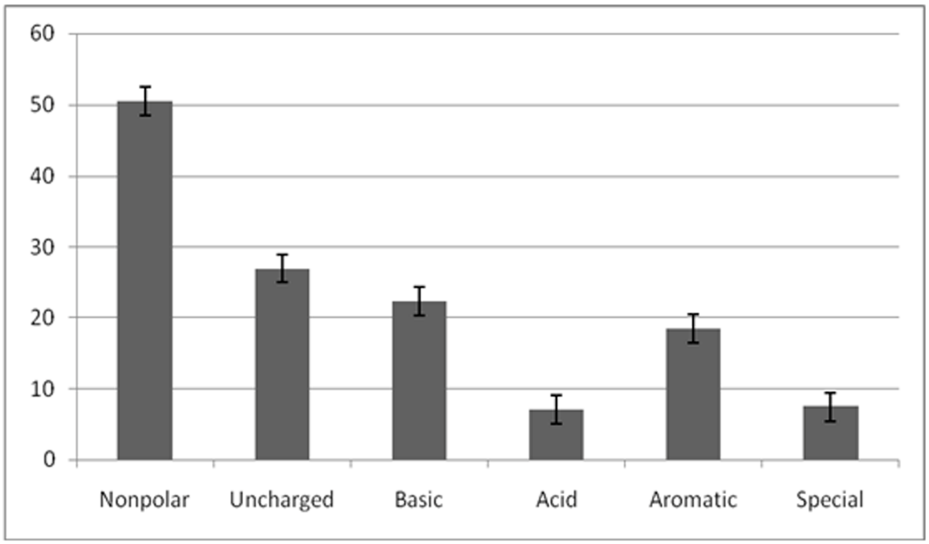


Figure - 4.20 Average side chain composition of the partial matK ORF. Amino acids were categorized as nonpolar, uncharged polar at pH 7, basic, acid, aromatic, or ‘‘special.’’ Categories consisted of the following amino acids: nonpolar (A, G, V, L, I, P, F, M, W, C); uncharged, pH 7 (N, Q, S, T, Y); basic (K, R, H); acidic (D, E); aromatic (F, W, Y); and ‘‘special’’ (C, G, P).

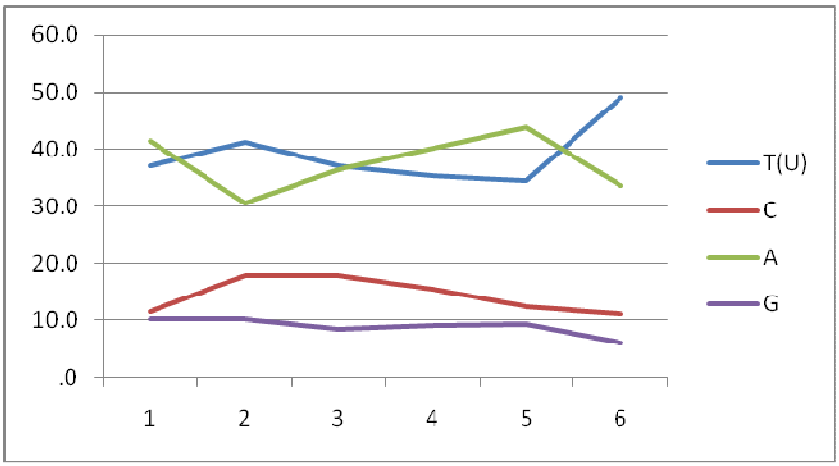


Figure - 4.21 Nucleotide compositions *trnH-psbA* sequence for the different species of Apocynaceae plants. The frequencies of nucleotide in sequences are presented as the total average value with the accuracy to tenths of a percent.

Nucleotide composition of *trnH-psbA* sequences of Apocynaceae was shown in Table-4.12. It shows that the *trnH-psbA* sequences had an almost equal percentage of T (average 38.0) and A (average 38.2) (Figure- 4.21). In *C. gigantea,* maximum percentage of T (49.2%) and minimum percentage of C (11.1) and G (6.1) was found. The highest percentage of C (18) was found in *A. cathartica*. In *N. oleander,* the highest percentage of G (10.6) and the lowest percentage of A (30.8) was found. *A. scholaris* had 43.8 (highest) of A.

Species Name	T	C	A	G	Total
T. peruviana	37.0	11.4	41.4	10.2	411.0
N. oleander	41.2	17.9	30.6	10.3	291.0
A. cathertica	37.0	18.0	36.5	8.6	373.0
C. roseus	35.4	15.5	40.0	9.1	407.0
A. scholaris	34.4	12.4	43.8	9.4	363.0
C. gigantea	49.2	11.1	33.7	6.1	380.0
Average	38.0	14.5	38.2	9.2	357.1

Table - 4.12 Nucleotide compositions *trnH-psbA* sequence of the different species of Apocynaceae plants. The frequencies of nucleotide in sequences are presented as the total average value with the accuracy to tenths of a percent.

4.2.2.2 Distance analysis (Barcode gap) :

K2P distances were used following the guidelines of the Consortium for the Barcoding of Life (CBOL) to evaluate performance barcoding locus (http://www.barcoding.si.edu/protocols.html). The sequence divergence (K2P) among the six medicinal plants of Apocynaceae based on *matK* sequence reveals that the maximum divergence (0.119) is between *Catharanthus roseus* and *Calotropis gigantea* (Table- 4.13). *Calotropis gigantea* had a maximum divergence with other five genera of Apocynaceae because *C. gigantea* belongs to subfamily Asclepiadoideae others five belong to subfamily Rauvolfioideae of the family Apocynaceae. Maximum mean divergence within species (0.029) was found in *Thevetia peruviana*. The sequence divergence (K2P) among the six medicinal plants

of the Apocynaceae reveals that the maximum divergence (0.478) is between *Alstonia scholaris* and *Nerium oleander* and minimum divergence (0.191) is between *Catharanthus roseus* and *Alstonia microphylla*. Maximum mean divergence within species (0.000) was found in maximum species (Table- 4.13).

	Species	**1**	**2**	**3**	**4**	**5**	**6**
1	***Catharanthus roseus***	**0.001**					
2	***Alstonia scholaris***	0.068	**0.000**				
3	***Calotropis gigantean***	0.119	0.094	**n/c**			
4	***Allamanda cathertica***	0.089	0.068	0.109	**0.001**		
5	***Tabernaemontana divaricata***	0.075	0.065	0.113	0.080	**0.005**	
6	***Thevetia peruviana***	0.083	0.068	0.103	0.070	0.073	**0.029**

Tabel- 4.13 Pairwise comparisions between *matK* sequences among the six species of Apocynaceae. Mean divergence (K2P) within (bold number on diagonal) and among (below diagonal). n/c indicates not comparable due to only one accession number).

The sequence divergence (K2P) among the six medicinal plants of Apocynaceae based on *trnH-psbA* reveals that the maximum divergence (0.654) is between *Alstonia scholaris* and *Calotropis gigantea* and minimum divergence (0.191) is between *Catharanthus roseus* and *Alstonia microphylla*. *Calotropis gigantea* had a maximum divergence with other four genera of Apocynaceae because *C. gigantea* belongs to subfamily Asclepiadoideae others five belong to subfamily Rauvolfioideae of the family Apocynaceae. Maximum mean divergence within species (0.000) was found in maximum species (Table- 4.14).

	Species Name	1	2	3	4	5	6
1	***Thevetia peruviana***	**n/c**					
2	***Alstonia scholaris***	0.287	**0.000**				
3	***Nerium oleander***	0.475	0.470	**0.000**			
4	***Allamanda cathertica***	0.610	0.576	0.274	**n/c**		
5	***Catharanthus roseus***	0.395	0.422	0.489	0.620	**0.000**	
6	***Calotropis gigantea***	0.654	0.615	0.354	0.493	0.608	**n/c**

Table- 4.14 Mean divergence (K2P) within (bold number on diagonal) and among (below diagonal) the six species of Apocynaceae from southern Assam based on *trnH-psbA* intergenic spacer (n/c indicates not comparable due to only one accession number).

4.2.3.4. Phylogenetic analysis:

Due to length variations in the *matK* sequences, 758 aligned nucleotide positions were used in sequence and phylogentic analysis. The 189 variable and 157 parsimony-informative positions contribute a considerable number of characters for resolving the phylogeny of the medicinal plants of the Apocynaceae. The overall transition/transversion bias was 0.769 in the partial *matK* sequence region. Total 28 partial *matK* sequences were belonging to 14 taxa from subfamily Rauvolfioideae and 2 taxa from subfamily Asclepiadaceae of Apocynaceae. We also determined 6 sequences of *trnH-psbA intergenic spacer* of 5 different species of Apocynaceae. Very few numbers of highly similar sequences or close member of this family were available in Genbank.

The phylogenetic tree was constructed on the basis of 28 partial *matK* sequences belonging to 16 taxa of Apocynaceae. Medicinally important members of Apocynaceae are *Catharanthus roseus, Vinca minor, Alstonia scholaris, A. microphylla, Tabernaemontana bufalina, T. divaricata, Thevetia peruviana, T.*

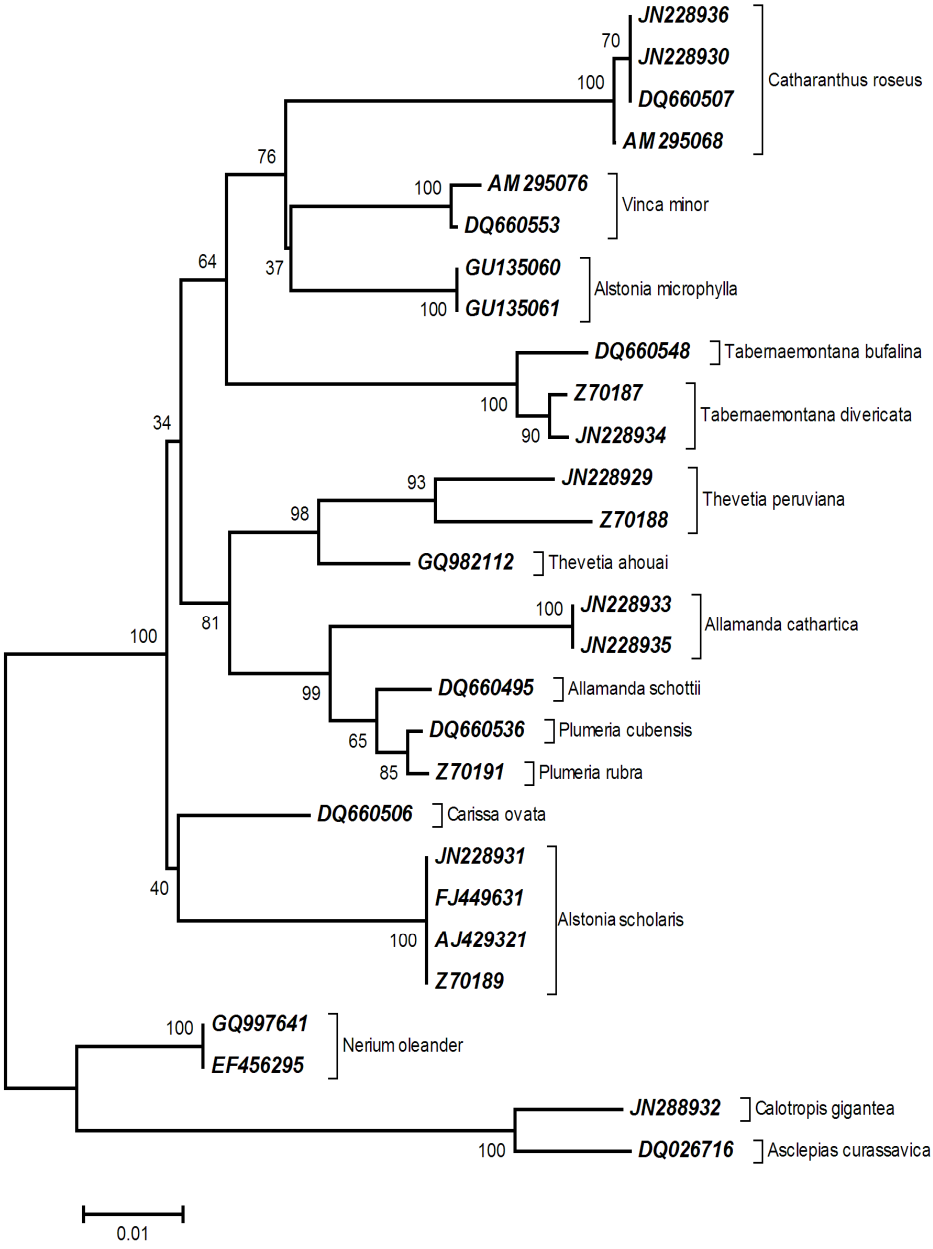


Figure - 4.22 Neighbour-Joining analysis of Kimura2-parameter (K2P) distance of *matK* sequences of Apocynaceae ~758 aligned nucleotide positions of *matK* (Nt. 520-1278) were used in phylogenetic analysis. A total of 1000 bootstrap replicates were calculated for the NJ tree construction. The indicated scale represents 0.01 nucleotide substitution per site.

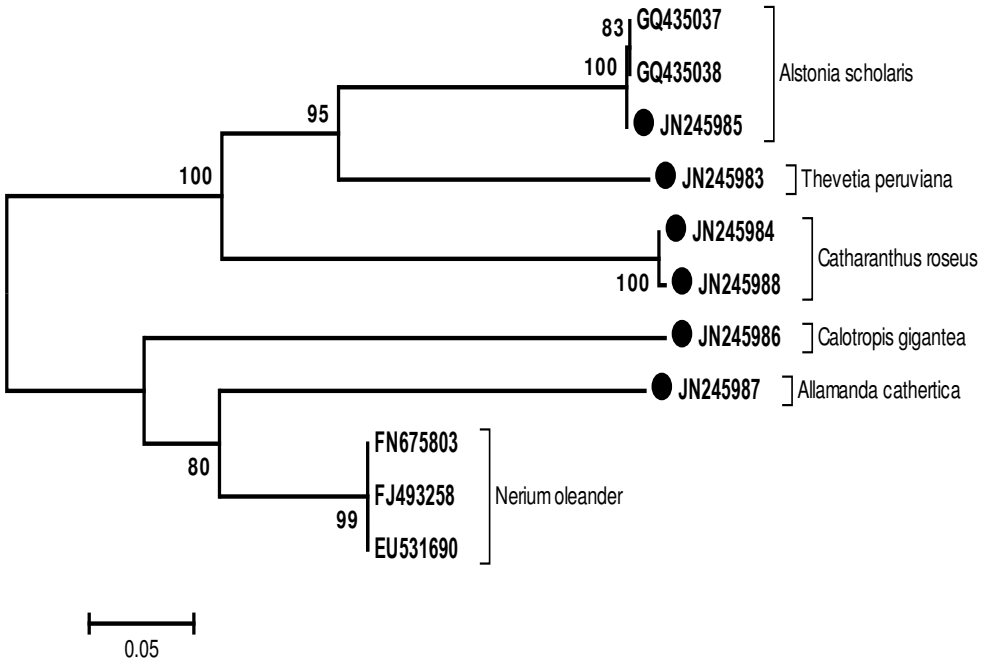


Figure- 4.23 Phylogenetic tree from partial *matK* (435 bp) of six species of Apocynaceae family. Bootstrap values (%) are shown on each branch. The indicated scale represents 0.005 nucleotide substitution per site.

- Generated sequence from this study

ahouai, Allemanda cathertica, A. schotti, Plumeria cebensis, P. rubra, Carissa ovate, Neruim oleander, Calotropis gigantea, and Asclepias curassavica. The different species formed distinctive clusters (Figure- 4.22). It is evident that all the 4 specimens of *Catharenthus roseus* with different accession numbers are in the same cluster. Similarly, *Thevetia peruviana, Tabernaemontana divaricata, Allemanda cathartica, Alstonia scholaris* with their different accession numbers are also formed different clusters. The member of the Asclepiadaceae subfamily, *Calotropis gigantea* and *Asclepias curassavica* located at basal position and used as out the group of the phylogentic tree.

The phylogenetic tree was constructed on the basis of 11 partial *trnH-psbA* intergenic spacer sequences belonging to 6 taxa of important medicinal plants of the Apocynaceae family. The different species formed distinctive clusters (Figure- 4.23). It is evident that all the 2 specimens of *Catharanthus roseus* with different accession numbers are in the same cluster. Similarly, *Alstonia scholaris, Nerium oleander, Alstonia microphylla* with their different accession numbers were also formed different clusters.

Chapter 4.3: Identification of Single Nucleotide polymorphism (SNP)/ Insertion and deletion (indels) for the development of DNA passport

4.3.1 Diagnostics Single Nucleotide polymorphism:

Character based method was applied for species delineation as an alternative to the genetic divergence based approach, as diagnostic characters prevent the loss of information characteristic to distance approach. Identification and confirmation of unique characters were accomplished in a straight forward manner with respect to the respective reference sequence from the database by visual inspection. Diagnostic characters were identified using at least two or more accession per species. In case of *matK,* maximum intraspecies variable site was found in *Thevetia peruviana* (17 variable site) and minimum intraspecies variable site were found in *Alstonia scholaris, Hibiscus rosa-sinensis, Senns obtusifolia, Citrus maxima, Citrus limon, Anisomeles indica* (non variable site) (Table- 4.15). In the variable site, mainly C>T and T>C changes were found within the species. In case of *trnH-psbA*, very few numbers of sequences were available in database. There was no variation in *Catharanthus roseus, Alstonia scholaris, Citrus maxima, Citrus limon, Citrus limetta, Citrus paradisi, Senna hirsuta, Senns obtusifolia, Senna occidentalis* except *Senna siamea* (268 A>T, 267 A>T, 270 A>T).

4.4.2 Inverted repeats:

The simple alignment of sequences between nucleotide site 27- 80 region contain a 54bp exact inverted repeat or palindrome sequence (5'-AAATAACGGATCAATACTGACCCCCAGCTGGGGGTCAGTATTGATCCGTT ATTT-3') within the *trnH-psbA* sequences. A BLASTN search in NCBI showed 100 % similarity with *Citrus* genus only (Table - 4.16). The inverted repeat sequence was conserved within the *Citrus* genus and authenticated the genus level identification of *Citrus*. There was no inversion in the flanked inverted region of *trnH-psbA* intergenic spacer.

Species	No of sequences	No of SNPs within intra species	SNPs position and changes
Catharanthus roseus	4	1	547 C>A
Alstonia scholaris	4	-	-
Thevetia peruviana	2	17	123 T>C, 126 C>T, 439 C>T, 453 T>G, 454 C>G, 455 C>G, 471 A>G, 477 A>G, 484 A>C, 493 G>T, 524 T>G, 536 G>T, 537 T>G, 555 C>A, 562 T>G, 573 T>3, 624 C>T
Allamanda cathertica	2	2	64 G>T, 782 A>T
Tabernaemontana divaricata	2	4	37 G>A, 528 C>A, 683 A>T, 711 C>G,
Aloe vera	7	1	106 T>G
Cynodon dactylon	2	1	109 T>C
Acmella oppositifolia	2	9	203 C>T, 238 C>T, 272 C>T, 275 C>T, 500 C>T, 507 A>C, 553 C>T, 758 C>T, 807 T>C
Cajanus cajan	2	11	165 G>T, 226 T>C, 416 T>A, 423 C>T, 514 T>C, 555 T>C, 582 A>G, 624 T>A, 719 A>T, 720 A>T, 722 G>A
Hibiscus rosa-sinensis	2	-	-
Senna hirsuta	2	4	417 A>T, 435 A>T, 445 A>T, 624 A>T
Senns obtusifolia	2	-	-
Citrus maxima	5	-	-
Citrus limon	2	-	-
Anisomeles indica	2	-	-

Table- 4.15 Diagnostic characters for each species in *matK*. Diagnostic characters for *matK* were identified with reference to GenBank sequence of particular species. For *matK*, the first nucleotide of start codon of reference sequence was considered as position 1.

Accession	Description	Max score	Total score	Query coverage	E value	Max identity
HE966751	*Citrus trifoliata* chloroplast DNA containing psbA-trnH	100	201	100%	5e-19	100%
JN315369	*Citrus limetta* voucher AUS-MP54 trnH-psbA intergenic spacer	100	201	100%	5e-19	100%
JN315368	*Citrus limetta* voucher AUS-MP50 trnH-psbA intergenic spacer	100	201	100%	5e-19	100%
JN315367	*Citrus x paradisi* voucher AUS-MP55 trnH-psbA intergenic spacer	100	201	100%	5e-19	100%
JN315366	*Citrus x paradisi* voucher AUS-MP49 trnH-psbA intergenic spacer	100	201	100%	5e-19	100%
JN315365	*Citrus maxima* voucher AUS-MP45 trnH-psbA intergenic spacer	100	201	100%	5e-19	100%
JN315363	*Citrus limon* voucher AUS-MP46 trnH-psbA intergenic spacer	100	201	100%	5e-19	100%
JN315362	*Citrus limon* voucher AUS-MP40 trnH-psbA intergenic spacer	100	201	100%	5e-19	100%
HQ415528	*Glycosmis parviflora* trnH-psbA intergenic spacer	100	201	100%	5e-19	100%
GQ435454	*Citrus medica* voucher PS1616MT03 psbA-trnH intergenic spacer,	100	201	100%	5e-19	100%
GQ435453	*Citrus medica* voucher PS1616MT02 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ435450	*Citrus limon* voucher PS1609MT01 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ435446	*Citrus maxima* voucher PS1600MT09 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ435445	*Citrus maxima* voucher PS1600MT08 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ435444	*Citrus maxima* voucher PS1600MT06 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ435443	*Citrus maxima* voucher PS1600MT05 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ435442	*Citrus maxima* voucher PS1600MT04 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ435439	*Citrus trifoliata* voucher PS1597MT01 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ435438	*Citrus reticulata* voucher PS1596MT01 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ435437	*Citrus medica* var. sarcodactylis voucher PS1595MT01 psbA-trnH	100	201	100%	5e-19	100%

Accession	Description	Max score	Total score	Query coverage	E value	Max identity
GQ435434	*Citrus maxima* voucher PS1590MT02 psbA-trnH intergenic spacer,	100	201	100%	5e-19	100%
GQ435433	*Citrus maxima* voucher PS1590MT01 psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ267065	*Citrus maxima* isolate HZY-08 photosystem II protein D1 (psbA)	100	201	100%	5e-19	100%
GQ267064	*Citrus medica* var. sarcodactylis isolate psbA-trnH intergenic spacer	100	201	100%	5e-19	100%
GQ267061	*Citrus maxima* isolate YO-3 photosystem II protein psbA-trnH	100	201	100%	5e-19	100%
GQ267060	*Citrus maxima* isolate YO-2 photosystem II protein psbA-trnH	100	201	100%	5e-19	100%
GQ267059	*Citrus maxima* isolate YO-1 photosystem II psbA-trnH	100	201	100%	5e-19	100%
GQ267057	*Citrus maxima* isolate HZY-14 photosystem II protein psbA-trnH	100	201	100%	5e-19	100%
GQ267056	*Citrus maxima* isolate HZY-13 photosystem II protein psbA-trnH	100	201	100%	5e-19	100%
GQ267055	*Citrus maxima* isolate HZY-12 photosystem II protein psbA-trnH	100	201	100%	5e-19	100%
GQ267054	*Citrus maxima* isolate HZY-11 photosystem II protein psbA-trnH	100	201	100%	5e-19	100%
GQ267053	*Citrus maxima* isolate HZY-01 photosystem II protein psbA-trnH	100	201	100%	5e-19	100%
GQ248268	*Citrus reticulata* voucher USDA PI 109635 psbA-trnH	100	201	100%	5e-19	100%
GQ248267	*Citrus aurantium* voucher USDA PI 128347 psbA-trnH	100	201	100%	5e-19	100%
EF590679	*Citrus aurantium* voucher USDA PI128347 psbA-trnH	100	201	100%	5e-19	100%
EF590680	*Citrus reticulata* voucher USDA PI109635 psbA-trnH	100	201	100%	5e-19	100%
DQ864733	*Citrus sinensis* chloroplast, complete genome	100	201	100%	5e-19	100%

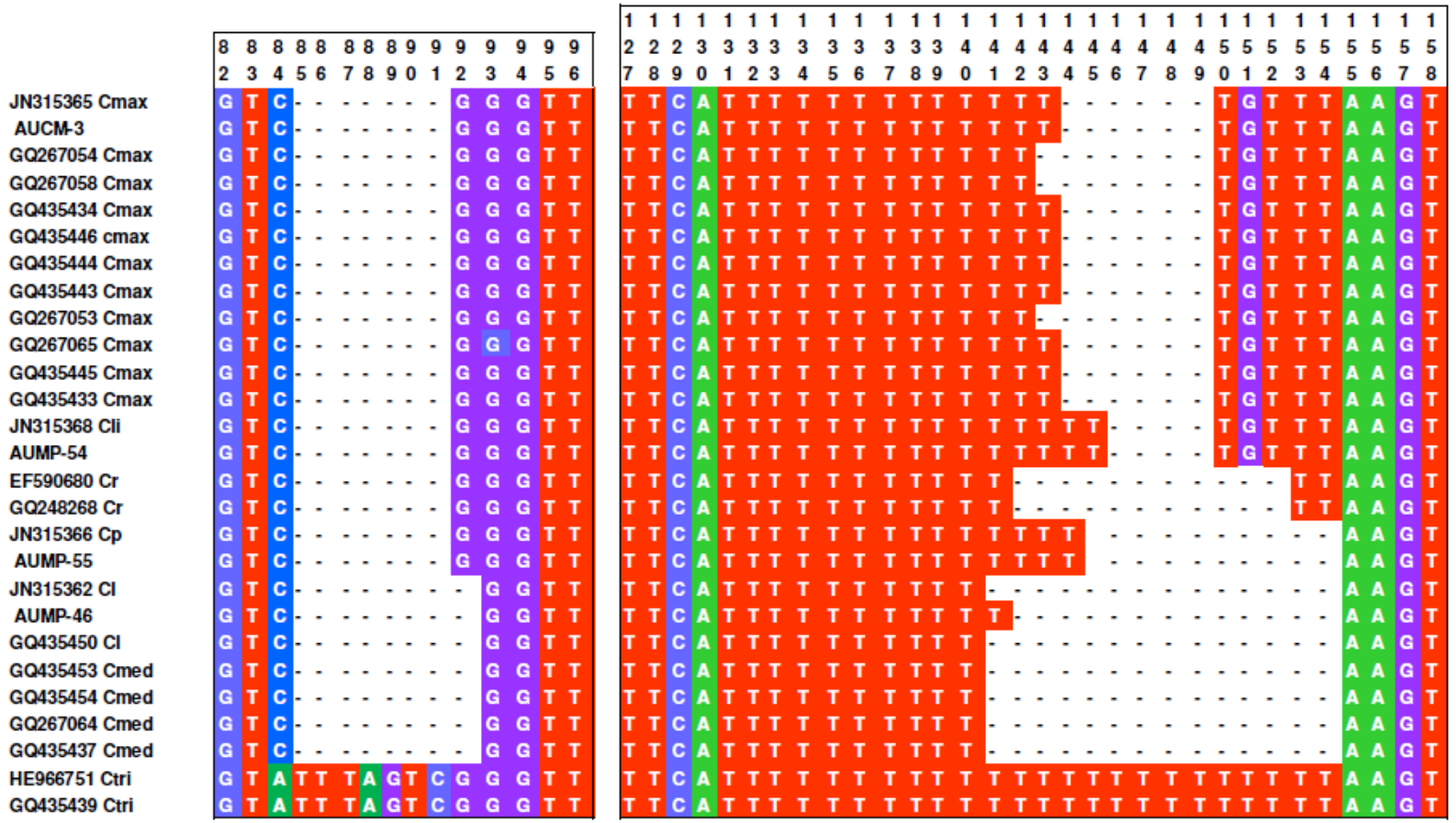
Table- 4.16 BLAST results of 54 bp exact inverted repeat or palindrome sequence. Maximum Score, Maximum identity, E-value, Total score are also given.

4.4.3 Diagnostics insertion and deletion (Indels):

The sequences within the group taxa of *Citrus* were aligned using the ClustalX program (Thompson et al., 1997) with a default setting of 15gap opening penalty, 6.66 gap extension penalty and 0.50 DNA transitions un-weighted. With few obvious misalignments were corrected manually using the similarity criterion (Simmons and Ochoterena, 2000). Identical sequences were considered as the same haplotype and representative each haplotypes were taken in alignment. GenBank database was searched using megablast during July-August, 2012 with default parameter adjusted to retrieving 5000 sequences.

Due to length variations in the sequences, 452bp aligned nucleotide positions of *trnH-psbA* were used in sequence analysis. We recognized 5 highly indels polymorphic regions within 85-91, 140-154, 186-198, 240-247 and 248-292. In 85-91 region, we found 7 bp insertions in *Citrus trifoliata* and 6 bp deletion common in *Citrus maxima*, *Citrus limetta*, *Citrus reticulata*, and *Citrus paradisi*. In 140-154 region, we identified 6-7 bp deletion in *Citrus maxima*, 4 bp deletion in *Citrus limetta*, 11 bp deletion in *Citrus reticulata*, 10 bp deletion in *Citrus paradisi*, 14 bp deletion in *Citrus limon* and *Citrus medica,* and only 14 bp insertion in *Citrus trifoliata*. A 13 bp insertion (within 186-198 region) was found in *Citrus paradisi*. In 240-247 region, 8 bp deletion was recognized in *Citrus limon, Citrus medica* and *Citrus trifoliata*. Another 9 bp deletion (within 284-292 regions) was found in *Citrus maxima*, *Citrus limetta* and 1bp insertion (290 positions) in *Citrus paradisi* (Figure- 4.24).

matK sequences exhibited indels in multiple of three at 5´ end where a 12bp insertion (641-652 region) was found in *Tabernaemontana divaricata, Tabernaemontana bufalina, Calotropis gigantea* and *Asclepias curassavica*; next 12bp insertion (677-688 region) was found in *Tabernaemontana divaricata* and *Tabernaemontana bufalina*, while the other 6bp insertion (1124-1129 region) was found only in *Calotropis gigantea* and *Asclepias curassavica* of a subfamily Asclepiadaceae (Figure- 4.25).


JN315365 Cmax
AUCM-3
GQ267054 Cmax
GQ267058 Cmax
GQ435434 Cmax
GQ435446 cmax
GQ435444 Cmax
GQ435443 Cmax
GQ267053 Cmax
GQ267065 Cmax
GQ435445 Cmax
GQ435433 Cmax
JN315368 Cli
AUMP-54
EF590680 Cr
GQ248268 Cr
JN315366 Cp
AUMP-55
JN315362 Cl
AUMP-46
GQ435450 Cl
GQ435453 Cmed
GQ435454 Cmed
GQ267064 Cmed
GQ435437 Cmed
HE966751 Ctri
GQ435439 Ctri

Cont…

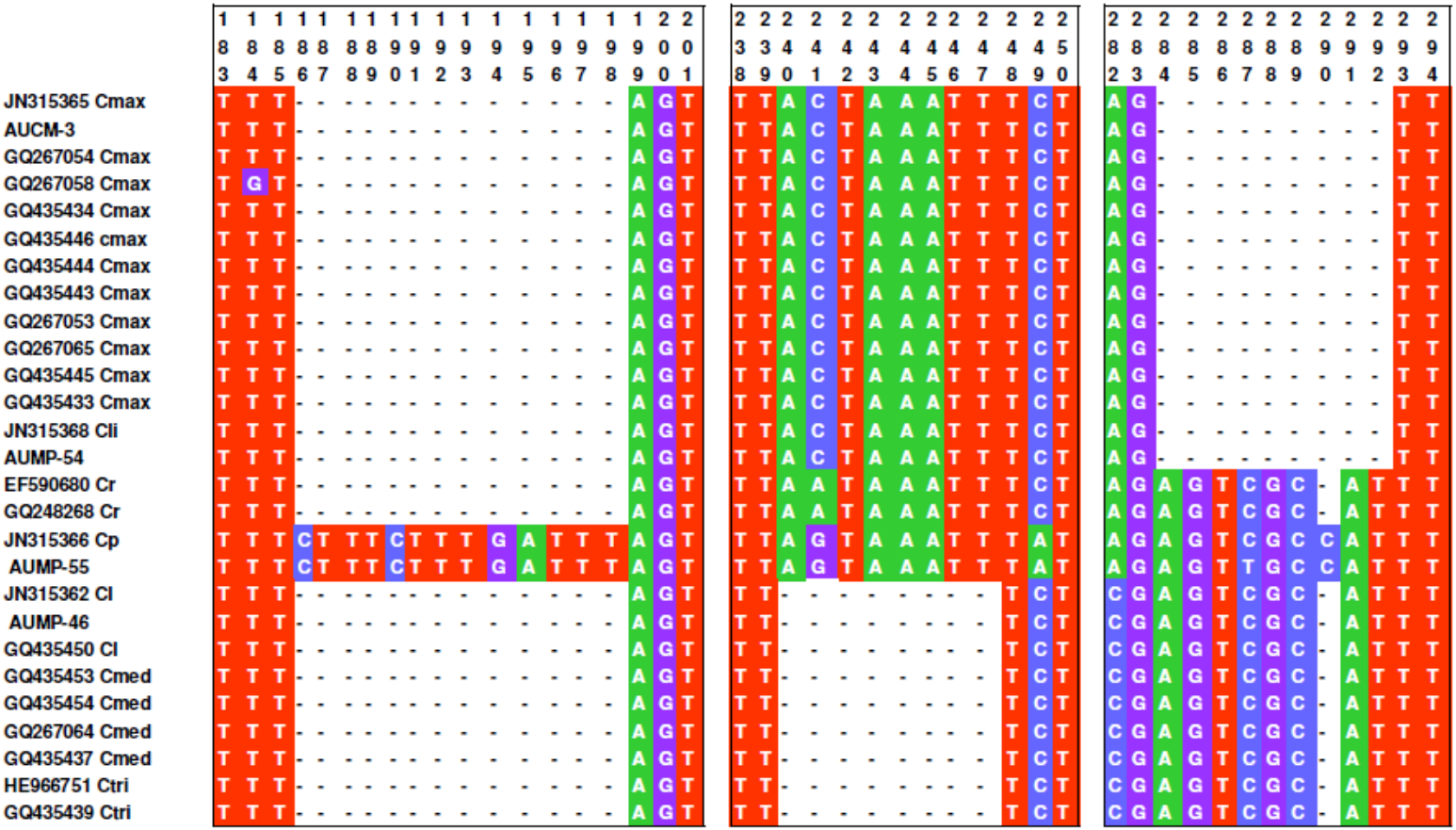


Figure- 4.24 Indels sites used as diagnostic character in the *trnH-psbA* intergenic spacer region of selected citrus species. Alignment of 23 sequences of the *trnH-psbA* intergenic spacer including reference and sample sequences. Each dot ‘‘-’’ in the sequence indicates an indels. Alignment positions of Indels and Accession number are also given. Abbreviations of *Citrus* species are Cmax-*Citrus maxima*, Cli-*Citrus limetta*, Cl- *Citrus limon*, Cp- *Citrus paradisi*, Cr-*Citrus reticulata*, Cmed- *Citrus medica*, Ctri-*Citrus trifoliate.*

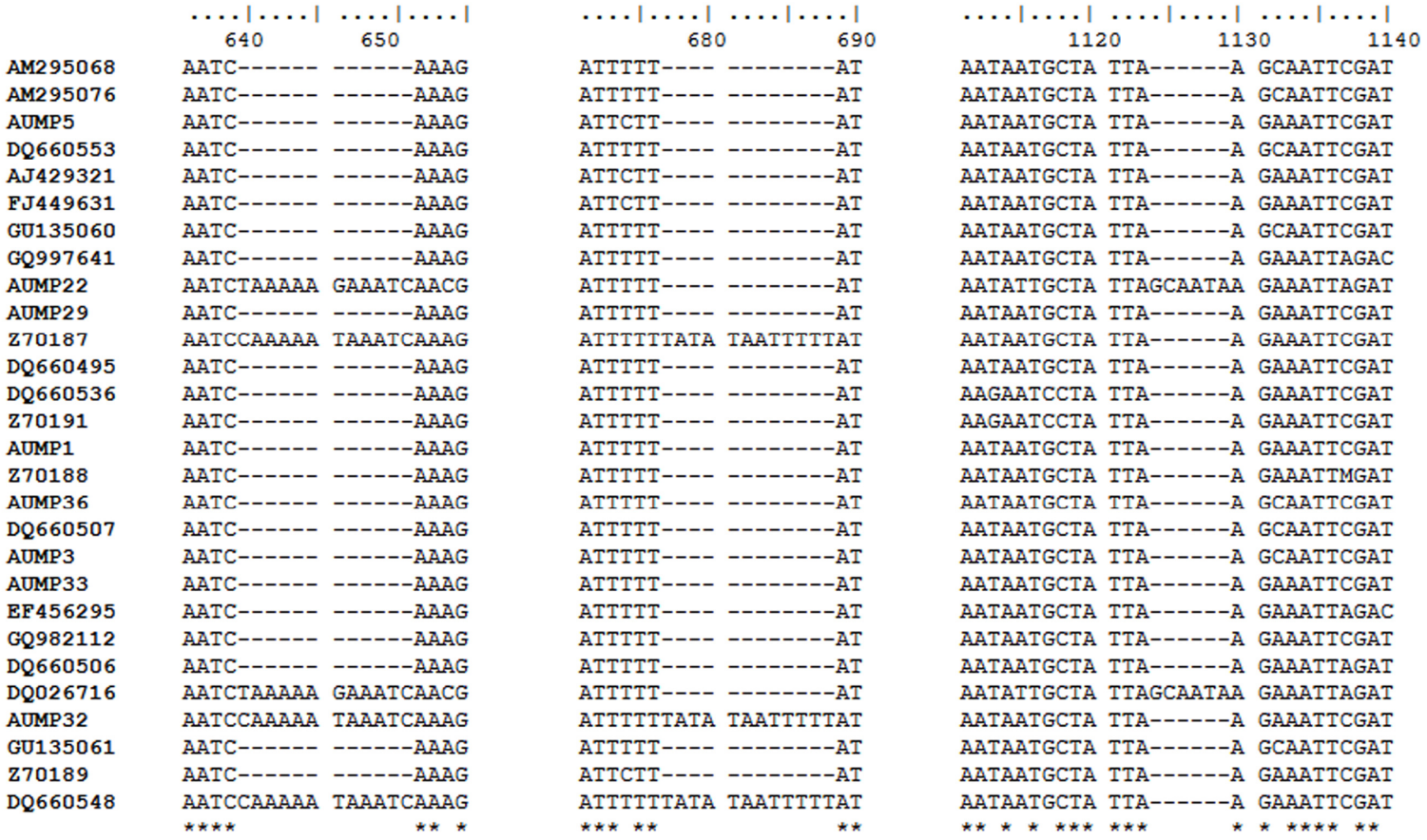


Figure- 4.25 Showing Alignment of 28 sequences of *matK* of Apocynaceae, containing indels of three different regions. A 12bp insertion found in *Tabernaemontana divericata* (Z70187, JN228934), *Tabernaemontana bufalina* (DQ660548), *Calotropis gigantea* (JN228932), *Asclepias curassavica* (DQ026716) (641-652 region) and in *Tabernaemontana divericata* (Z70187, JN228934), *Tabernaemontana bufalina* (DQ660548) (677-688 region) and 6bp insertion in *Calotropis gigantea* (JN228932), *Asclepias curassavica* (DQ026716) (1124-1129 region). * indicate conserve nucleotide.