“Nothing in biology makes sense except in the light of evolution” - Dobzhansky (1973)

Everything makes more sense with evolutionary systematics and in the midst of revival stands systematics indicated by the common application of new analytical approaches including the molecular techniques. Molecular phylogenetic analyses have become familiar and they have provided matchless intuitions into relationships at all stages of plant phylogeny.

Molecular systematics is a comparatively new field and has made great developments in the last several years with over 50 wholly sequenced land plant chloroplast genomes and several more on the horizon. The impact of molecular data in the field of plant systematics is huge, reforming the concept of relationships and circumscriptions at all levels of the taxonomic hierarchy (Small, Cronn, and Wendel 2004). Information from macromolecules are used to infer the evolutionary relationships of the organisms under study which comes as a part of molecular systematics.

With the advent of molecular techniques our understanding on genetic variability of organisms located at various levels of the tree of life has advanced greatly (Avise 2012; Hillis, Mortiz, and Mable 1996; Hollingsworth, Bateman, and Gomall 1999; Mondini, Noorani, and Pagnotta 2009; Wen and Pandey 2005). Variety of molecular techniques are available to study genetic variability including PCR based techniques, DNA sequencing and cladistic analyses of the organellar genome of chloroplast (cpDNA) and the nuclear genome (nDNA) in plants. The chloroplast DNA or nuclear ribosomal DNA dominates as the two sources from which molecular data applied to plant systematics originates (Small, Cronn, and Wendel 2004).

**DNA sequencing**

The direct method of sequencing and the amplification of DNA are the *in vitro* techniques for the progression of molecular systematics and there has been an increasing interest in the analyses of molecular systematics since the use of DNA nucleotide sequence data. Such molecular approaches have been helpful in resolving
many phylogenetic questions in angiosperm systematics not solved by the phenotypic characters which is a common approach. Nucleotide sequencing is much valued in plant systematics as the nucleotides are the basic units of information encoded in organisms. Boundless systematically informative variations are present which is helpful and by using different genes or its regions the problems at various taxonomic levels can be resolved. Traditionally, the studies on variation of a species are based on morphological characters of homologous genes (Hamby and Zimmer 1992). When considering these variations at the molecular or DNA level, they are mainly due to the changes in DNA sequences. The major advantage of molecular data is that all known life forms are based on nucleic acids. Each nucleotide position, in theory, can be considered as a character and supposed to be independent. The morphological adaptations of an organism are reflected in its biomolecules and vice versa (Patwardhan, Ray, and Roy 2014).

The molecular methodologies depend on nucleotide sequences of RNA, DNA and sequences of amino acids of a protein determined using modern techniques. By equating homologous molecules from different organisms it is possible to establish their degree of resemblance thereby giving an insight into the hierarchy of relationship.

The “blueprints” of organisms except for few are written in DNA and hence can be used to study the evolutionary relationships by making a comparison of the DNA. Classical approaches used morphological and chemical characters.

However, molecular level comparison offers several advantages as explained by Nei and Kumar (2000).

- DNA can be used for comparing any groups of organisms (plants, animals, bacteria etc.) as they are basically made of Adenine, Thiamine, Guanine and Cytosine which is impossible in conventional approaches.
- Mathematical model can be used to formulate change and compare DNA's of distantly related organisms as DNA follows a regular pattern of evolutionary changes.
• Genomes of organisms are made of long nucleotide sequences and hence harbor more phylogenetic information than morphological and chemotaxonomical approaches.

**Phylogeny and Systematics**

Phylogenetics reveals evolutionary relationship and systematics involves assigning organisms to a specific taxonomic rank. However, both are closely associated with each other as classification is done to reveal evolutionary histories. Systematics and evolution are allied to one another and it is almost impossible to investigate one of these without looking at the other (Wanke 2006).
3.1. INTRODUCTION AND REVIEW OF LITERATURE

About 90% of land plant diversity are the seed plants, meeting majority of the human needs and dominating the terrestrial ecosystems. One of the common notion in taxonomy is that the species is its basic unit. A vital step to build up a reliable reference system of biological information is the unequivocal association of a scientific name to a biological entity (Wheeler 2004). About 1.7 million species have been formally described by taxonomists, over the last 250 years, since Carl Linnaeus classification system and it perhaps represent only a small section of the real biodiversity existing on the planet (Vernooy et al. 2010). Hebert, Ratnasingham, and de Waard (2003) proposed a molecular and bioinformatical tool called DNA barcoding to help discover the hidden biodiversity and to provide a useful standardized tool for species identification.

The term “DNA barcode” was coined by Hebert in 2003 (Hebert, Ratnasingham, and de Waard, 2003; Hebert, Cywinska, and Ball 2003) for global species identification and gained worldwide attention among the scientific community (Blaxter 2003; Gregory 2005; Marshall 2005; Miller 2007). Universality, specificity on variation and easiness on employment are the distinguishing features of a DNA barcode. The gene segment selected as barcode should be suitable for a wide range of taxa, having high variation between species but conserved within, so that the intra-specific variation will be irrelevant (Consortium for the Barcode of Life-CBOL 2009; Kress et al. 2005; Pennisi 2007). Ideal DNA barcode must be usually a uniform short sequence of DNA of 400-800 base pair (bp) (Savolainen et al. 2005). The reliability of barcoding is based on the supposition that genetic variation within a species is much smaller than the one between species. Hebert studies showed 98-100% species identification success rates and the suppositions were proved valid (Hajibabaei et al. 2006; Hebert, Stoeckle, et al. 2004, Hebert, Penton, et al. 2004; Ward et al. 2005).

3.1.1. Mitochondrial DNA (mtDNA)

The data from mitochondrial DNA can be very powerful in solving species-level phylogenies. The gene order in mitochondrion is variable separated by large
regions of noncoding DNA, the major non-coding area being the control region. It is involved in the regulation and initiation of mtDNA replication and transcription, for the regulation of heavy (H) and light (L) strand transcription and of H-strand replication. The mitochondrial genome rearranges itself repeatedly and hence many rearranged forms occur in the same cell. The use of mtDNA has become increasingly popular in phylogenetics and population genetic studies. Some of the mitochondrial genes used are COI/II, mitochondrial 12S, cytochrome b (Patwardhan, Ray, and Roy 2014). These regions have been widely used for studying phylogenetic relationship of insects, beetles, Indian leopards, birds etc.

DNA barcoding is well established in animals using coxI gene (cytochrome oxidase I) in mtDNA. "Folmer region" at the 5' end of coxI was used in the pioneer work and also for the further studies. This gene is 648 nucleotide base pairs long and is flanked by regions of conserved sequences, making it comparatively easy to PCR, amplify sequence and analyze. In animals the mtDNA is exploited as it is highly conserved regarding the gene content and order, but possess high nucleotide substitution rates (Brown, George, and Wilson 1979; Boore 1999; Moritz, Dowling, and Brown 1987). For algae, coxI or COI—has been developed as a universal DNA barcode (Hebert, Ratnasingham, and de Waard 2003).

COI is not a suitable barcode in advanced plants because of its slower rate of evolution and moreover, the plants quickly change their mitochondrial genome structure (John et al. 2011). Rearrangements, transfer of genes to the nuclear genome and incorporation of foreign genes are very common in higher plant mitochondrial genomes (Mower et al. 2004; Palmer et al. 2000).

In contrast, plants have an additional genome - the chloroplast genome (cpDNA) apart from the nuclear (nDNA) and mitochondrial (mtDNA) genomes. The two regions most commonly exploited for plant DNA barcoding are chloroplast and nuclear genome.

### 3.1.2. Chloroplast DNA (cpDNA)

In comparison with mitochondria and nuclear genome, cpDNA is the smallest. The chloroplast genome is well suited for evolutionary and phylogenetic
IV. 3. DNA Barcoding

Introduction and review of literature

Phytochemical and Molecular Systematics of Aristolochia spp. studies particularly above the species level, because cpDNA is comparatively an abundant component of the total DNA on which wide molecular informations are available. It is highly conserved and not altered by evolutionary processes and has a conservative rate of nucleotide substitution. It contains single copy genes and can be isolated easily. Majority of the works related to plant systematics are concentrated on the chloroplast genome mainly because of its comparatively simple genetics.

3.1.2.1. Organization of the chloroplast genome

The size of the chloroplast genome range from 120-217 kilo base pairs (kb) and the green alga Floydia terrestis has a size of 521.168 bp (Gyulai et al. 2012). In land plants there is fairly high degree of conservation in size, structure, gene content and linear order of the genes (Downie and Palmer 1992). cpDNA is closed circular region that contains inverted repeats. Mostly two inverted repeats (approximately 25 kb each) are present that are mirror images of one another in terms of gene complement. One large (LSC) and one small single-copy (SSC) region separates the inverted repeats and these repeats accumulate point mutations slower than the single-copy regions (Curtis and Clegg 1984; Gaut 1998; Wolfe, Li, and Sharp 1987; Wolfe 1991). According to Perry and Wolfe (2002) the nucleotide substitution rate is 2.3 times greater in the single-copy regions compared to inverted repeats. Single-copy regions have got attention and became the focus point as most regions studied (e.g., rbcL, atpB, trnL-trnL-trnF) are located in the LSC and inverted repeats evolve at a relatively slower rate. All cpDNA molecules carry basically the same set of genes that include the ones for ribosomal RNA, transfer RNA, ribosomal proteins and about 100 different polypeptides and subunits of enzyme capturing CO₂, but arranged differently in different species of plants. cpDNA is relatively abundant and is maternally inherited in angiosperms

Three functional categories present in the chloroplast genome include (1) protein-coding genes, (2) introns and (3) intergenic spacers; the latter two referred to as noncoding regions, do not encode proteins. Almost 43% of the LSC and SSC are noncoding according to the Nicotiana chloroplast map (Wakasugi et al. 1998). Fifteen introns constitute around 10.6% of the single-copy chloroplast DNA, while 92 intergenic spacers comprise 32.3%. The plant phylogenies using chloroplast DNA
sequences began with the pivotal publications of Taberlet et al. (1991), Clegg et al. (1994), Morton and Clegg (1993), and Gielly and Taberlet (1994) in the early 1990s. These studies were simplified by the three chloroplast genomes of *Marchantia polymorpha* (Ohyama et al. 1986), *Nicotiana tabacum* (Shinozaki et al. 1986) and *Oryza sativa* (Hiratsuka et al. 1989) that had been fully sequenced.

In sequence related investigations only few noncoding chloroplast DNA regions have been compared (Taberlet et al. 1991). The comparative phylogenetic efficacy of 21 noncoding chloroplast regions was studied by Shaw et al. in 2005. Finding and documentation of other noncoding plastid region of higher variability include explaining the origin of domesticated species (Wills and Burke 2006), locating biogeographic movements (Ickert-Bond and Wen 2006; Schonswetter, Popp, and Brochmann 2006) and clarifying complex relationships among species (Shaw and Small 2005). Rapidly evolving plastid regions are also useful for species identification via molecular barcoding or microarray analysis; imparting themselves to studies of simulating sequence evolution (Cartwright 2005) and ultimately may lead to a better understanding of the functions of noncoding DNA or the mechanisms for cpDNA evolution on the basis of frequency comparison of various types of mutational events. Because of the lack of broad comparison of various noncoding portions of the chloroplast genome, only little is known about the different benefits of the many potential chloroplast markers. Considered similar to the mitochondrial gene in animals, the chloroplast genes are mostly exploited for barcoding plants. Chloroplast genes in plants have slower rate of evolution; uniparental inheritance, non-recombination and structural stability in both the genic and intergenic regions (Vijayan and Tsou 2010).

### 3.1.2.2. Various plastid loci used for barcoding studies

Plastid loci verified for DNA barcoding of plants with some success were *rbcL* (Kress et al. 2007; Newmaster, Fazekas, and Ragupathy 2006), *matK* (Lahaye et al. 2008), *rpoB* and *rpoC1* (Newmaster, Fazekas, and Ragupathy 2006; Sass et al. 2007; Seberg and Petersen 2009). Nine intergenic spacers was studied by Kress et al. (2005) which included *trnH-psbA*, *rp136-rps8*, *atpB-rbcL*, *trnK-rps16*, *trnL-F*, *trnC-ycf6*, *ycf6-psbM*, *psbM-trnD* and *trnV-atpE* and were found to meet the barcode
norms as being the most variable regions. According to Taberlet et al. (2007) the trnL (UUU) intron and its shorter P6 loop (10-143 bp) could be a suitable candidate as plant barcode. Other widely used plastid barcoding markers include accD, atpF-atpH, psbK-psbl, ycf5 and trnL-trnF. However, lack of sufficient variation makes these markers unsuitable for plant DNA barcoding at lower taxonomic levels (Li et al. 2015). Genes rpoB, rpoC1 and rpoC2 encode three out of the four subunits of the chloroplast RNA polymerase (Shinozaki 1986; Serino and Maliga 1998). The plastid accD gene encodes the β-carboxyl transferase subunit of acetyl-Co A carboxylase and is present in most flowering plants, except in grasses (Hajdukiewicz, Allison, and Maliga 1997). ycf5 is the only gene from the small single-copy region being seriously studied for its suitability in DNA barcoding. Ycf5 encodes a protein containing 313 amino acids (Yasui and Ohnishi 1998). The plastid ndh gene complex, identified originally from the tobacco plastid genomes (Ledford 2008) and liverwort (Ohyama et al. 1986), codes for subunits of a functional respiratory protein complex of size 550kDa within the mature chloroplast (Burrows et al. 1998). The genes atpF and atpH encode ATP synthase subunits CFO I and CFO III respectively (Drager and Hallick 1993). psbK and psbl genes encode two low molecular mass polypeptides, K and I respectively, for the photosystem II (Meng, Wakasugi, and Sugiura 1991) and are conserved from algae to land plants (Knauf and Hachtel 2002; McNeal et al. 2007). The trnL (UAA) – trnF (GAA) locus contains the trnL (UAA) gene, its intron and the intergenic region between trnL (UAA) and trnF (GAA). Taberlet et al. (1991) employed trnL intron for the first time in plant systematic studies.

Dong et al. (2015) reported that two regions of the plastid gene ycf1, ycf1a and ycf1b can serve as core barcode in land plants as they were the most variable loci superior to the current plastid candidate barcodes.

A number of genes and intergenic spacers are being used for barcoding studies. Some of the commonly used regions are described further.
**IV. 3. DNA Barcoding**

*Introduction and review of literature*

3.1.2.2.1. **rbcL** (Rubisco Large sub-unit)

The plastid-encoded rbcL gene is the most commonly and extensively used gene to provide sequence data and hence more than 50,000 rbcL sequences are available in the GenBank (Chase et al. 1993; Chase et al. 2007; Donoghue et al. 1993; Li et al. 2015; Newmaster, Fazekas, and Ragupathy 2006; Ritland and Clegg 1987). rbcL was the first gene sequenced from the plants (Zurawski et al. 1981) and is the best characterized gene among the plastid genes.

3.1.2.2.1.1. **Structure and Function**

This single copy gene is approximately 1428 (Newmaster, Fazekas, and Ragupathy 2006) base pairs in length, free from length mutations except at the far 3’ end and has a fairly conservative rate of evolution (Fig. 49). It is located between the atpB and trnR coding regions (Yoshinaga et al. 1996). It encodes the large subunit of the photosynthetic enzyme RUBISCO (Ribulose, 1, 5-bisphosphate carboxylase/oxygenase).

![Fig. 49. Structure of rbcL gene](image)

3.1.2.2.1.2. **rbcL and Plant Systematics – advantages and limitations**

The gene is the right choice for interpretation of phylogenetic relationships at higher taxonomic levels (Duvall et al. 1993; Les, Garvin, and Wimpee 1991; Plunkett et al. 1995). rbcL sequences evolve slowly because of its slow synonymous nucleotide substitution rate in comparison with nuclear genes and its functional constraint cuts down the evolutionary rate of non-synonymous substitution (Wolfe, Li, and Sharp 1987); it shows the lowest divergence among the plastid genes in flowering plants (Kress et al. 2005). For the taxa that diverged from 10’s of million years rbcL seems suitable because of its substitution rate (Zurawski, Clegg, and Brown 1984). The gene worked out well in many families of angiosperms like Cucurbitaceae (Reddy 2009) Orchidaceae (Cameron et al. 1999), Rhamnaceae (Richardson et al. 2000), Araucariaceae (Stefanovic et al. 2009) etc. for solving intergeneric and interspecific relationships. Except for parasites, the gene occurs in...
all plants with many copies available; is fairly long with no difficulties of alignment. The sequence data of the rbcL gene are extensively used in the renewal of phylogenies all over the seed plants.

Soltis and Soltis (1998) opined that rbcL is best suited to reconstruct relationships down to the generic levels, but is not suitable for specific levels due to modest discrimination ability (Fazekas et al. 2008; Lahaye et al. 2008) and according to Doebley et al. (1990) its ability to resolve phylogenetic relationships below the family level is often meager. An ideal DNA barcode should be short enough to amplify from degraded DNA and analyzed via single-pass sequencing and hence it forms a major limitation for rbcL (Chase et al. 2007). It has been rectified in most taxa by developing primers for PCR amplification and sequencing of such short sequence within the rbcL gene (Fay, Swensen, and Chase 1997; Kress et al. 2007).

Owing to ease in PCR amplification across a wide array of plant groups and the availability of sequence information in many plant groups, the CBOL working group (2009) has recognized rbcL as one of the most potential gene sequences for DNA barcoding in plants. Majority of the investigating groups opinioned that rbcL should be used in combination with other markers because of its low species discriminatory power (CBOL 2009; Chase et al. 2007; Hollingsworth et al. 2009; Soltis and Soltis 1998).

### 3.1.2.2.2. matK (maturase K)

matK is one of the most quickly evolving genes among the various plastidial genes.

#### 3.1.2.2.2.1. Structure and Function

It is having a length of about 1550 bp (Hilu, Alice, and Liang 1999; Wolfe 1991) coding around 500 amino acids for the translated protein product and encodes the enzyme maturase which is involved in the splicing of type-H introns from RNA transcripts (Neuhaus and Link 1987). Since matK is embedded in the group II intron of the lysine gene trnK, PCR-amplification is made easier with a primer set designed from the conserved regions of the genes trnK, rso 16 and psbA. In majority of land plants, the matK gene is nested between the two exons (Fig. 50).
of tRNA-lysine (trnK). This gene contains indels of different length and number (Hilu et al. 2003; Hilu and Alice 1999). Throughout the matK open reading frame (ORF), the substitution rates of the nucleotides are not consistent and possess regions exhibiting high mutation rates (Hilu and Liang 1997). The gene has neutral or purifying selection suggested so because the third codon position tends to have a slightly higher mutation rate than the first and second (Young and dePamphilis 2000). A highly conserved region close to the 3′ end of the gene that lacks indels have shown up in various plants obtained from the analysis of the amino acid sequence predicted (Hilu and Liang 1997). This region resembles a conserved functional domain seen in mitochondrial group II intron maturases and is 448 bp in length (Neuhaus and Link 1987; Sugita, Shinozaki, and Sugiura 1985).

With the exception of some parasitic species in *Cuscuta*, the chloroplasts of land plants (Funk et al. 2007; McNeal et al. 2009) possess a single maturase gene matK, expressed at least in the green tissue (Barthet and Hilu 2007; Jardin et al. 1994; Vogel, Borner, and Hess 1999). matK as positioned within the trnK gene, is believed to be involved in the splicing of the trnK precursor. The retention of free-standing matK reading frame in several chloroplast genomes of parasitic plants and ferns even though they have lost trnK, suggests added tasks for matK, perhaps in splicing other introns (Duffy, Kelchner, and Wolf 2009; Funk et al. 2007; Gao et al. 2009; McNeal et al. 2009; Wolfe, Morden, and Palmer 1992). Thus, it has been suggested that other chloroplast introns are also targeted by matK.

Fig. 50. The structure of the matK gene (Wolfe, Morden, and Palmer 1992)

### 3.1.2.2.2. matK and Plant Systematics - advantages and limitations

matK because of its universal presence and rapid evolution in plants, has been used as a marker in systematic studies (Hilu and Liang 1997; Kelchner 2000). matK was found to be reliable in solving intergeneric as well as interspecies relationships in many angiosperms; has suitable length and low transition/
transversion rate (Min and Hickey 2007; Selvaraj, Sarma, and Sathishkumar 2008; Soltis and Soltis 1998; Johnson and Soltis 1995) and has been proposed either alone or in combination with other loci to resolve systematic problems. Starr, Naczi, and Chouinard (2009) advocated the use of matK alone as a universal barcode for land plants after trying out matK, rbcl, rpoCB, rpoC1 and trnH-psbA as barcodes in Cyperaceae. 100% success was achieved using specific matK primers (Cuenoud et al. 2002) in 1660 plant samples in a study by Lahaye et al. (2008). In a study by Kress and Erickson (2007) over 90% of species in Orchidaceae was identified by matK. The CBOL-Plant Working Group (2009) tested matK in around 550 plant species and found that nearly 90% of the angiosperm samples were easily amplified and sequenced using a single primer pair, though the success was limited in gymnosperms (83%) and much poor in cryptogams (10%).

With the available primer sets matK is difficult to amplify universally which is a problem. Different taxonomic groups required different primer pairs (Chase et al. 2007). In certain taxonomic groups as reported by Wolfe, Li, and Sharp (1987), new sets of primers were developed that work good in majority of the taxonomic groups and amplifies a DNA fragment of 930 bp between positions 429 and 1313 of the matK sequence (Cuenoud et al. 2002; Schmitz-Linneweber et al. 2001). Discrimination of less than 49% in the nutmeg family was found in a study by Newmaster et al. (2008). Fazekas et al. (2008) tried to identify 92 species from 32 genera with matK, but the success percentage was only 56.

3.1.2.2.3. trnH-psbA (Transfer RNA for histidine - D1 protein of photosystem II)

This is an intergenic spacer and is one of the most variable genome segments in the chloroplast of angiosperms. psbA is the gene encoding a 32-kilodalton thylakoid membrane protein (Curtis and Clegg 1984).

3.1.2.2.3.1. Structure and function

The literatures suggest that its length varies from 296 to 1120 bp and with an average length of 450 bp (CBOL 2009; Chase et al. 2007; Hollingsworth et al. 2009; Shaw et al. 2005). This intergenic spacer has highly conserved coding sequences on both sides (Fig. 51) which makes the design of universal primers
possible (Shaw et al. 2005) with a single primer pair probably able to amplify nearly all angiosperms (Shaw et al. 2007).

The psbA - trnH intergenic spacer possesses 2 regions that vary in their evolutionary conservation. The psbA 3’UTR in control of regulation of gene expression and the psbA-trnH non-transcribed intergenic spacer that has no specific function but show variability across angiosperms.

The psbA gene encodes the D1 reaction center protein of photosystem II, the expression of which depends on physiological state, light intensity and the development stage of the plant (He et al. 1998). It accumulates at high levels in chloroplasts and low levels in amyloplasts (Deng and Gruissem 1987).

Fig. 51. The Structure of Intergenic Spacer psbA-trnH Intergenic (Storchova and Olson 2007)

3.1.2.3.2. trnH-psbA and Plant Systematics - advantages and limitations

The locus has been effectively PCR amplified across a diverse group including angiosperms and gymnosperms, ferns, mosses and wild liverworts using the primers. trnH-psbA is a suitable candidate for next generation sequencing because of its high sequence variability and pcr recovery.

One of the primary worries on the use of trnH-psbA as a DNA barcode marker is the presence of mononucleotide repeats that interfere with sequencing (CBOL 2009; Devey, Chase, and Clarkson 2009). The presence of repeat motifs is problematic as sequence quality weakens immediately following large repeat motifs. The other limitations include problems on alignment of sequences in certain taxa because of high length variations and difficulty in obtaining high-quality bidirectional sequences. Another matter of concern is the occurrence of rps 19 gene or pseudogene within the trnH-psbA region (Chang 2006). In some conifers and monocots (Chase et al. 2007; Hollingsworth et al. 2009) the trnH-psbA sequence is greater than 1000 bp which is supposed to be due to the presence of duplicated loci.
and a pseudogene. In some cases it is exceedingly short, less than 300 bp (Kress et al. 2005) and shorter than 100 bp in bryophytes (Stech and Quandt 2010). Extremely short length of trnH-psbA forms a limitation as it may not exhibit adequate sequence variation for the discrimination of the taxa (Schindel and Miller 2005). According to Devey, Chase, and Clarkson (2009) developing amplification strategies for the matK gene are better than resolving the problem of mononucleotide repeats in trnH-psbA.

The CBOL Plant Working Group (2009) proposed trnH-psbA as the most preferred supplementary locus. They found the species discriminatory power of trnH-psbA to be maximum (69%) among the seven loci tested.

In pteridophytes, a success rate of 90% was found when trnH-psbA was used (Ma et al. 2010).

CBOL (2009) recommends the use of gene regions from other genomic compartments such as the nuclear genome as they may provide highly successful solutions for barcode identifications within specific taxa or in specific situations (Soininen et al. 2009; Taberlet et al. 2007).

3.1.3. Nuclear DNA (nDNA)

Nuclear genome sizes, measured by the number of base pairs of DNA, of different plant species vary enormously between species, although each species has a characteristic and relatively constant genome size. The plant nuclear genome consists of repetitive DNA which includes tandem repeats. The tandem repeats possess rDNA, 45S rRNA and 5S rRNA genes.

Nuclear DNA segment is likely to offer more information on species identity as it is biparentally inherited and provides evidence on hybridization events. The only nuclear DNA region that have been tested for suitability as barcodes in plants is the ITS region of ribosomal DNA (Chase et al. 2007; Kress et al. 2005; Sass et al. 2007). Nuclear ribosomal genes are used as markers in phylogenetic studies and include 16S rRNA, 5S rRNA and 28S rRNA. Gradually evolving rRNA genes are used at higher taxonomic levels (Kuzoff et al. 1998); external and internal intergenic spacers at lower taxonomic levels (Alvarez and Wendel 2003). Only a few genes are tested for reliability probably because of low species discrimination due to
conservation of functional genes across large lineages and the difficulty in obtaining high universality of PCR amplification of single or low-copy genes particularly from degraded and low quality DNA. In comparison to nuclear loci, the chloroplast loci shows relatively low rate of evolution and hence in most studies, DNA barcoding technology relies mainly on chloroplast genes (Dong et al. 2012).

3.1.3.1. ITS (Internal transcribed spacer)

The nucleic acid core of the ribosome is encoded by the rDNA cistron which is a multigene family. The rDNA units are structured into large blocks in the chromosome called the nucleolar organizer regions by repetition of these units thousands of times (Appels, Honeycutt, and Dutta 1986; Hemleben et al. 1988, Mondini, Noorani, and Pagnotta 2009). The individual unit of this multiple gene family of rDNA evolves in a concerted manner such that higher level of overall sequence similarity exists among copies of the rDNA within a species, but differs among different species (Brown, Wensink, and Jordan 1972; Hershkovitz and Zimmer 1996). This high sequence similarity is achieved through a process termed as concerted evolution (Hood, Campbell, and Elgin 1975), which involves unequal crossing over and gene conversions.

3.1.3.1.1. Structure and Function

The rDNA is organized as tandem repeat units with 18S, 5.8S, 26S coding region and two internal transcribed spacers (ITS1 and ITS2) present on either side of 5.8S region (Vijayan and Tsou 2010). ITS is a piece of non-functional RNA situated between structural ribosomal RNAs precursor transcript (Persson 2000). The ITS1 and ITS2 spacers along with the 5.8s nrDNA are referred to as ITS region (Fig. 52). The entire ITS region is less than 700 bp and ITS1 and ITS2 are individually around 300 bp in length. Within angiosperms 5.8S subunit is almost invariant in length (163-164 bp). ITS1 and ITS2 form a part of the transcription unit and seem to function in the maturation process but are not incorporated into mature ribosomes (Baldwin 1992).
3.1.3.1.2. ITS and Plant systematics - advantages and limitations

Currently, nrITS is considered as a useful marker for both plants and animals, because of its universal nature, biparental inheritance and relatively higher evolutionary variations due to less functional constrains (Alvarez and Wendel 2003; Baldwin et al. 1995; Rogers and Bendich 1987; Zimmer et al. 1980). ITS from nuclear genome as a DNA barcode for flowering plants was proposed by Kress et al. (2005). ITS as a barcode is not restricted to green plants but are utilized across the taxonomic diversity covering almost the tree of life, including fungi and lichenicolous fungi (Cubero et al. 2004; Martin et al. 2003; Wu et al. 2000), unicellular and pluricellular algae (Leclerc et al. 1998; Oppen et al. 2005), non-arthropod invertebrates (Dumont et al. 2005), arthropods (Harris and Crandall 2000) and even vertebrates (Booton et al. 1999). The amplification and sequencing of ITS is possible without internal primers because of its moderate size (Gernandt, Liston, and Pinero 2001). ITS1 and ITS2 regions can be PCR amplified independently by anchoring primers in the conserved coding genes which enables easy amplifications of ITS even from poor quality or degraded DNA (Kress et al. 2005). PCR amplification of ITS even from herbarium specimens is made possible by the presence of multicopy structure.

However, certain amount of intra-individual variations among the copies of ITS1 and ITS2 sequences have been reported in tree plants and asexually propagated plants. Recent hybridization, lineage sorting, recombination among copies, high mutation rate and pseudogene formation of cistrons are considered to be the reasons for such variations (Alvarez and Wendel 2003; Buckler, Ippolito, and Holtsford 1997; Rogers and Bendich 1987; Wendel, Schnabel, and Seelanan 1995;
Zimmer et al. 1980). ITS does not encode any proteins, frequently have indels and may be hard to align across unrelated species leading to incorrect homology assessments. Other phenomena that affect its utility are multiple rDNA arrays, concerted evolution, pseudogenes, fungal contamination, secondary structure, difficulties of amplification and sequencing and compensatory base changes (Feliner and Rossello 2007; Hollingsworth, Graham, and Little 2011).

However, nrITS is still considered to be a powerful barcoding tool at the species level (Vijayan, Zhang, and Tsou 2009). When tested for its suitability as barcode in plants, nrITS showed better universality (88%) and species discrimination in comparison with nine other chloroplast loci (Kress et al. 2005). Considering the availability of universal primers, presence of multiple copies in cells, high universality and good species discriminatory power, nrITS was proposed as a potential candidate for barcoding in plants (Kress et al. 2005; Sass et al. 2007). Hollingsworth et al. (2009) later suggested that nrITS can be considered for barcoding of species that have limited variations in the plastid genome. Okuyama and Kato (2009) used ITS as a barcode for identifying a reproductively isolated and cryptic species of Asimitellaria from its close relatives. Based on this finding, it was suggested that nrDNA can be used for accurate and efficient demarcation of plant biological species in lineages with various life history traits (annuals, perennials, trees, aquatics) and evolutionary backgrounds (recent and old radiations, oceanic island endemics).

Wang et al. (2015) viewed ITS2 to be treated as a separate barcode as it is a subregion of ITS. Chen et al. (2010) reported on the use of ITS2 to identify medicinal plants and their close relatives and proved the potential of this nuclear gene as a barcode for plants. The ITS2 region has also been tested and extensively used as the ideal DNA barcode for algae and plants (Buchheim et al. 2011; Gao et al., “Evaluating feasibility……,” (2010); Gao et al., “Identification of……,” (2010); Garcia-Robledo et al. 2013; Gu et al. 2013; Hassel, Segreto, and Ekrem 2013; Kuzmina et al. 2012; Pang, Luo, and Sun 2012), fungi (Heinrichs, Hoog, and Hase et al. 2012; Kelly et al. 2011), protists (Gou et al. 2012) and animals (Salvi and Mariottini 2012; Yao et al. 2010).
Moreover, the China Plant BOL group (2011) suggested that ITS should be included as the core barcode in plants and wherever problem of sequencing occurs ITS2 can be used as a backup because of its conserved sequence characters that diminish the amplification and sequencing problems.

Because of the problems arising from paralogous sequences, pseudogenes, intragenomic variability and difficulties in direct sequencing of PCR products, the CBOL Plant Working Group (2009) has not regarded nrITS suitable for a universal plant DNA barcode, but as a supplementary locus for taxonomic groups which have less resolution with cpDNA and where direct sequencing of nrITS is possible. However in fungi, the internal transcribed spacers (ITS) of nuclear ribosomal DNA have been selected as the universal barcode region (Horton and Bruns 2001; Schoch et al. 2012). The various gene sequences used as barcodes from three genomic regions are given (Fig. 53).

Fig. 53. Genes from three genomes that are used as barcodes (adopted from Chen et al. 2010)

Based on the information available from reviews, a number of gene sequences, both coding and non-coding sequences from the chloroplast DNA along with a gene from nuclear DNA have been examined for their suitability as barcodes.
3.1.4. Different approaches in barcoding

Barcoding of plants are going on and much experimentations being done to single out a universal barcode for plants. Various concepts and approaches have been tried out from “gene to genome” like single locus barcodes, multi locus barcodes, super-barcoding etc.

3.1.4.1. Single-locus barcodes

This makes use of a single region for the identification or discrimination of species. Deletion of gene is suggested to be most important limiting factor for single loci technique (Li et al. 2015).

3.1.4.2. Multi-locus barcodes

Multi-locus approach has been suggested by many authors which according to them will be required to obtain sufficient species discrimination (CBOL Plant Working Group 2009; Chase and Fay 2009; Chase et al. 2007; Erickson et al. 2008; Hebert, Penton, et al. 2004; Kane and Cronk 2008; Kress and Erickson 2007; Lahaye et al. 2008; Pennisi 2007). Some of these combinations included rbcL+trnH-psbA (Kress and Erickson 2007), rpoC1+rpoB+matK (Chase et al. 2007) and matK+atpF-atpH+psbK-psbI (Pennisi 2007). Ford et al. (2009) proposed a combination of rpoC1+rpoB+matK as the most likely combination for barcoding land plants. The CBOL Plant Working Group (2009) recommended matK in combination with rbcL as the standard two-locus barcode for plants because of its high universality and species discrimination. trnH-psbA has high species discriminatory power and according to Kress et al. (2005) proposed it can be used along with nrITS for DNA barcoding in plants. The use of trnH-psbA mostly in combination with matK in barcoding land plants have been proposed by many (Chase et al. 2007; Erickson et al. 2008; Fazekas et al. 2008; Newmaster et al. 2008; Pennisi 2007; van de Wiel et al. 2009). Although majority of literatures support the use of three locus barcode (rbcL + matK + trnH-psbA) along with ITS as a supplementary locus, many other genes are being tested for their utility as potential markers. matK+atpFatpH+trnH-psbA was a combination suggested by Pennisi (2007). trnH-psbA can thus be used in a three-locus barcode system wherever the two-locus barcode system fails to deliver ample
IV. 3. DNA Barcoding

Phytochemical and Molecular Systematics of Aristolochia spp.

resolution. trnH-psbA can be used in two-locus (Kress et al. 2005) or three locus rpoC1+matK+trnH-psbA (Chase et al. 2007) barcode systems to provide sufficient resolution. CBOL (2009) proposed the use of a three locus DNA barcode for land plants. Improved results were observed using the three marker combination (rbcL + matK + trnH-psbA) in comparison to two marker (rbcL + matK), particularly for the basal angiosperms, with overall congruence to prospects based on Angiosperm Phylogeny Group II (APG II 2003) being remarkably high. The three locus marker provides a better estimate of species identity.

3.1.4.3. Super-barcoding

The whole-plastid genome sequence in plant identification referred to as “super-barcoding” have been proposed by many (Erickson et al. 2008; Nock et al. 2011; Parks, Cronn, and Liston 2009; Sucher and Carles 2008; Yang et al. 2013) though not yet universally accepted; the major concerns being the sequencing cost and difficulties involved in obtaining complete plastid genome sequences in comparison to the use of single-locus barcodes. According to Kane and Cronk (2008) the complete chloroplast-genome (cp-genome) had as much variation as the CO1 locus in animals and hence may be used as a plant barcode. Since the complete cp-genome has a conserved sequence ranging from 110 to 160 kbp, it helps in providing more variation to distinguish close species as against the commonly used barcodes.

The simplest method of species identification is to determine whether a gene exists in either of two species which can be possible by the super-barcoding (Hebert, Penton, et al. 2004). This is because super-barcoding is more efficient in spotting gene loss and outlining gene order than traditional barcoding (Luo et al. 2008, 2009). However, Hollingsworth, Graham, and Little (2011) claimed that the full plastid haplotype does not always track species boundaries and hence not a good marker.

3.1.3.4. Specific barcodes

Single-locus barcodes is said to lack sufficient variations while the use of fully annotated super-barcodes are not economical and may be very complicated for
IV. 3. DNA Barcoding

Introduction and review of literature

Phytochemical and Molecular Systematics of Aristolochia spp.

laboratories that lack the needed experience. Single-locus barcodes lack adequate variations, while fully annotated super-barcodes can be costly and may be over complicated for laboratories that lack the necessary experience. According to Ahmed et al. (2013), discovering a universal barcode locus for land plants is not possible especially in the chloroplast wherein lineage-specific evolution and non-random spatial patterns of substitution can occur is not feasible.

Li et al. in 2015 has put forward the concept of “specific barcode” i.e., a fragment DNA sequence that has an adequately high mutation rate to facilitate species identification within a given taxonomic group (Table 18). Several options to be selected as markers includes genes, intergenic spacers, partial gene sequences, partial intergenic spacers and even sequences including partial gene sequences and partial intergenic spacers.

Table 18. Barcodes selected for suitability at higher taxonomic levels (Li et al. 2015)

<table>
<thead>
<tr>
<th>Genera/Family</th>
<th>Barcode markers used</th>
<th>Success rate of unique identification (%)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solanum</td>
<td>trnS-trnG/ndhF</td>
<td>100</td>
<td>Zhang et al. (2013)</td>
</tr>
<tr>
<td>Fabaceae</td>
<td>matK</td>
<td>96</td>
<td>Gao et al. (2011)</td>
</tr>
<tr>
<td>Taxus</td>
<td>trnL-F/ITS</td>
<td>100</td>
<td>Liu et al. (2011)</td>
</tr>
<tr>
<td>Lemnaceae</td>
<td>atpF-atpH</td>
<td>92.85</td>
<td>Wang et al. (2010)</td>
</tr>
<tr>
<td>Rutaceae</td>
<td>ITS2</td>
<td>100</td>
<td>Luo et al. (2010)</td>
</tr>
<tr>
<td>Medicinal plants</td>
<td>ITS2</td>
<td>99.8</td>
<td>Chen et al. (2010)</td>
</tr>
<tr>
<td>Pteridophytes</td>
<td>psbA-trnH</td>
<td>90.2</td>
<td>Ma et al. (2010)</td>
</tr>
<tr>
<td>Hydrocotyle</td>
<td>trnH-psbA</td>
<td>100</td>
<td>van de Wiel et al. (2009)</td>
</tr>
<tr>
<td>Dendrobium</td>
<td>psbA-trnH</td>
<td>100</td>
<td>Yao et al. (2009)</td>
</tr>
<tr>
<td>Aspalathus</td>
<td>trnT-trnL</td>
<td>100</td>
<td>Edwards et al. (2008)</td>
</tr>
<tr>
<td>Macrozamia</td>
<td>ITS</td>
<td>100</td>
<td>Sass et al. (2007)</td>
</tr>
<tr>
<td>Cycas</td>
<td>ITS</td>
<td>91.7</td>
<td>Sass et al. (2007)</td>
</tr>
</tbody>
</table>
IV. 3. DNA Barcoding

**Introduction and review of literature**

Requirements for applying specific barcodes are,

(i) Rich database of cp-genome sequences (though they need not represent the fully annotated genome of the target taxa).

(ii) Database with primers for each plant group resulting from the exploration of these cp-genomes.

The primer database with specific barcodes could be used to distinguish the known species. The two tier ‘1+1’ approach for unknown species is that firstly, unknown species are classified using single locus barcodes (e.g., rbcl) at the family or genus levels and then the corresponding specific barcodes are chosen from cp-genome datasets to accomplish discrimination at the species level. This model gives flexibility in the selection of barcode rather than using the routine selection procedure and in the run may find regions similar to COI. This model is aimed to share the specific barcodes at higher taxonomic levels as given in table which makes it more appealing.

3.1.5. **Applications of plant barcoding**

The main application of DNA barcoding is species identification. It helps taxonomists to quickly identify novel/ambiguous and polymorphic species (Lahaye et al. 2008; Miwa et al. 2009; Xiao, Moller, and Zhu 2010). It helps in detecting the invasive species (Bleeker et al. 2008; van de Wiel et al. 2009), for checking illegal trade of endangered species of both plants (Jeanson, Labat, and Little 2011; Muellner, Schaefer, and Lahaye 2011; Yesson et al. 2011), identifying plants at juvenile or mature stages (Gonzalez et al. 2009), identifying adulteration in herbal drugs (Chen et al. 2014; Techen et al. 2014), act as genetic resource tags and in turn would be of help in conservation of genetic diversity (Yesson et al. 2011), in forensic analyses, for distinguishing the poisonous plants from edible ones etc. DNA barcoding have the capacity to completely revolutionize our awareness of diversity of life and the relationship of living things to nature. By harnessing technological advances in electronics and genetics, DNA barcoding will enable people to rapidly and economically recognize species and retrieve information about them and also in...
the speed discovery of thousands of species yet to be named. Barcoding has the potential to provide a vital new tool for appreciating and managing the Earth’s immense biodiversity (Ali et al. 2014).

### 3.1.6. Selection of the candidate barcodes

The barcoding technique proves to be efficient, but a universal primer barcode has not been identified in plants as in animals. The search for a universal barcode has been compared with that of “Holy Grail” (Rubinoff, Cameron, and Will 2006) and is still in progress. In plants a universal barcode is yet to be evolved.

A number of studies claim the use of a combination of barcodes for plants. rbcL and matK - the chloroplast genes, forming the two-locus barcode were most commonly used as universal barcode in most studies and was approved as the barcode region for land plants by CBOL (2009). These were also concluded as the universal plant DNA barcodes in the 3rd International Barcode of Life (iBOL) conference held at Mexico in 2009. The use of non-coding intergenic spacer trnH-psbA along with rbcL and matK to form the three-locus barcode was also proposed by the CBOL working group in 2009. At the 4th International Barcode of Life Conference (iBOL) held in Adelaide in 2011 concluded rbcL, matK and ITS as the universal barcode candidates. The inclusion of ITS in the core barcode was also advocated by the China Plant BOL Group (2011). The use of nr ITS as a supplementary locus has an added advantage against others as it exhibits biparental inheritance. In plants, ITS2 loci have been advocated as a universal DNA barcode and in animals as a complementary locus (Chen et al. 2010; Yao et al. 2010).

The selection of barcodes was done on the basis of the most common barcodes used at the onset of the study. Thus, five regions rbcL, matK, trnH-psbA, ITS and ITS2 was selected. In a study Techen et al. (2014) using the chemical abstracts service database, “Scifinder” listed out the major barcodes mentioned mostly in the literature from 2010-2013. The major regions cited were in the order of ITS, trnH-psbA, matK followed by rbcL which included the same loci selected for the present investigation.
DNA barcoding enables rapid, accurate and automatable species identifications by using short, standardized gene regions as internal species tags which will make the Linnaean taxonomic system more accessible, with benefits to ecologists, conservationists etc. More broadly, DNA barcoding allows a day to be envisioned when curious mind, from professional biologists to school children, will have an easy access to the names and the biological attributes of any species on the planet (Hebert and Gregory 2005). Once fully developed, DNA barcoding has the potential to completely change not only how biologists understand and monitor biodiversity, but also, as emphasized by Janzen (2005), the “relationship of the general public to nature”. If implemented successfully, barcoding will provide a vital new tool for appreciating and managing the Earth’s immense and changing biodiversity.
3.2. MATERIALS AND METHODS

3.2.1. Genomic DNA isolation and PCR amplification

Total genomic DNA was isolated from fresh leaf tissues. The protocol used is described under different sections.

3.2.1.1. DNA isolation

DNA isolation was carried out using GenElute Plant Genomic DNA Miniprep Kit (Sigma). About 100 mg of the tissue was ground into fine powder in liquid nitrogen by a mortar and pestle and was transferred to a microcentrifuge tube.

The samples were kept in ice until the addition of lysis solution. 350µl of lysis solution A and 50µl of lysis solution B were added to the tube. After thoroughly mixing by vortexing and inverting the mixture was incubated at 65°C for 10 minutes with occasional inversion. 130µl of precipitation solution was added to the mixture, mixed completely by inversion and the sample was placed on ice for 5 minutes. The sample was centrifuged at 14,000 rpm (Eppendorf Centrifuge 5804R) for 5 minute to pellet the cellular debris, proteins, and polysaccharides. The supernatant was transferred to the GenElute filtration column tube and centrifuged at 14,000 rpm for 1 minute. This removed any cellular debris not removed in the previous step. The filtration column was discarded and 700µl of binding solution was added directly to the flow through liquid and mixed thoroughly by inversion. 700µl of this mixture was added into GenElute nucleic acid binding column and centrifuged at 14,000rpm for 1 minute.

The flow through liquid was discarded and the collection tube was retained. The column was returned to the collection tube and the remaining sample was applied to the column. Centrifugation was repeated as above and the flow through liquid and the collection tube were discarded. The binding column was placed into a fresh 2ml collection tube. 500 µl ethanol-added wash solution was added to the binding column and centrifuged at 14,000rpm for 1 minute. The flow through liquid was discarded and the collection tube was retained. The wash was repeated once more. The binding column was transferred to a new collection tube. 30 µl of elution
solution (pre-warmed to 65°C) was added to the binding column and centrifuged at 14,000rpm for 1 minute. The elution was repeated once more with 30 µl of the elution solution. The stock DNA was properly labelled and stored at 4°C.

3.2.1.2. Agarose Gel Electrophoresis for DNA Quality check

The quality of the DNA isolated was checked using agarose gel electrophoresis. 1µl of 6X gel-loading buffer (0.25% bromophenol blue, 30% sucrose in TE buffer pH-8.0) was added to 5µl of DNA. The samples were loaded to 0.8% agarose gel prepared in 0.5X TBE (Tris-Borate-EDTA) buffer containing 0.5 µg/ml ethidium bromide. Electrophoresis was performed with 0.5X TBE as electrophoresis buffer at 75 V until bromophenol dye front has migrated to the bottom of the gel. The gels were visualized in a UV transilluminator (Genei) and the image was captured under UV light using Gel documentation system (Bio-Rad).

3.2.1.3. PCR Analysis

PCR amplification reactions were carried out in a 20 µl reaction volume which contained 1X PCR buffer (150mM Tris HCl, pH-8; 500mM KCl), 0.2mM each dNTPs (dATP, dGTP, dCTP and dTTP), 2.5mM MgCl₂, 20ng DNA, 1 unit of AmpliTaq Gold DNA polymerase enzyme (Applied Biosystems), 0.1 mg/ml BSA and 4% DMSO, 5pM of forward and reverse primers (Table 19).

<table>
<thead>
<tr>
<th>Target</th>
<th>Primer Name</th>
<th>Direction</th>
<th>Sequence (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>matK</td>
<td>matK_xf</td>
<td>Forward</td>
<td>TAATTTAGTGATTTACTTCATT</td>
</tr>
<tr>
<td></td>
<td>matK_MALPR1</td>
<td>Reverse</td>
<td>ACAAGAAAGTCTGAGTAT</td>
</tr>
<tr>
<td>rbcL</td>
<td>rbcLa_f</td>
<td>Forward</td>
<td>ATGTCCACACAAACAGAGACTAAGC</td>
</tr>
<tr>
<td></td>
<td>rbcL724_rev</td>
<td>Reverse</td>
<td>GAAAAATCGAATTCCAGCGGG</td>
</tr>
<tr>
<td>trnH-psbA</td>
<td>psbA-F</td>
<td>Forward</td>
<td>GTATGCATGAAGTCATCAGAGT</td>
</tr>
<tr>
<td></td>
<td>trnH2</td>
<td>Reverse</td>
<td>GCGCATGGATGATTCCAGCAATCC</td>
</tr>
<tr>
<td>ITS</td>
<td>ITS-F</td>
<td>Forward</td>
<td>AATTGTCGCGGTAATGTTTC</td>
</tr>
<tr>
<td></td>
<td>ITS-R</td>
<td>Reverse</td>
<td>CTGCGGTATATTGAGAGGAAAT</td>
</tr>
<tr>
<td>ITS2</td>
<td>ITS2-F</td>
<td>Forward</td>
<td>CGCATTGACATCGAGTCCTT</td>
</tr>
<tr>
<td></td>
<td>ITS-R</td>
<td>Reverse</td>
<td>CGCATTGAGTTGAGAGGAAAT</td>
</tr>
</tbody>
</table>

The PCR amplification was carried out in a PCR thermal cycler (GeneAmp PCR System 9700, Applied Biosystems).
3.2.1.4. PCR amplification profiles

<table>
<thead>
<tr>
<th></th>
<th>matK</th>
<th>rbcL</th>
<th>psbA-trnH, ITS &amp; ITS2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95 °C - 5.00 min</td>
<td>94 °C - 5.00 min</td>
<td>95°C - 5.00 min</td>
</tr>
<tr>
<td></td>
<td>95 °C - 0.30 min</td>
<td>94 °C - 0.30 min</td>
<td>95°C - 0.30 min</td>
</tr>
<tr>
<td></td>
<td>45 °C - 0.40 min, 10 cycles</td>
<td>55 °C - 0.30 min, 40 cycles</td>
<td>58°C - 0.30 min, 40 cycles</td>
</tr>
<tr>
<td></td>
<td>72 °C - 1.00 min</td>
<td>72 °C - 0.30 min</td>
<td>72 °C - 0.30 min</td>
</tr>
<tr>
<td></td>
<td>95 °C - 0.30 min</td>
<td>72 °C - 5.00 min</td>
<td>72 °C - 5.00 min</td>
</tr>
<tr>
<td></td>
<td>51 °C - 0.40 min, 30 cycles</td>
<td>4 °C - ∞</td>
<td>72 °C - 5.00 min</td>
</tr>
<tr>
<td></td>
<td>72 °C - 1.00 min</td>
<td></td>
<td>4 °C -∞</td>
</tr>
<tr>
<td></td>
<td>72 °C - 7.00 min</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 °C - ∞</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The PCR products were checked in 1.2% agarose gels prepared in 0.5X TBE buffer containing 0.5 µg/ml ethidium bromide. 1 µl of 6X loading dye was mixed with 5 µl of PCR products and was loaded and electrophoresis was performed at 75V power supply with 0.5X TBE as electrophoresis buffer for about 1-2 hours, until the bromophenol blue front had migrated to almost the bottom of the gel. The molecular standard used was a 2-log DNA ladder (NEB). The gels were visualized in a UV transilluminator (Genei) and the image was captured under UV light using Gel documentation system (Bio-Rad).
3.2.1.5. ExoSAP-IT Treatment

ExoSAP-IT (GE Healthcare) consists of two hydrolytic enzymes, Exonuclease I and Shrimp Alkaline Phosphatase (SAP), in a specially formulated buffer for the removal of unwanted primers and dNTPs from a PCR product mixture with no interference in downstream applications.

Five micro litres of PCR product is mixed with 2 µl of ExoSAP-IT and incubated at 37°C for 15 minutes followed by enzyme inactivation at 80°C for 15 minutes.

3.2.2. Sequencing of amplified product

3.2.2.1. Sequencing using BigDye Terminator v3.1

Sequencing reaction was done in a PCR thermal cycler (GeneAmp PCR System 9700, Applied Biosystems) using the BigDye Terminator v3.1 Cycle sequencing Kit (Applied Biosystems, USA) following protocol.

The PCR mix consisted of the following components:

- PCR Product (ExoSAP treated) - 10-20 ng
- Primer - 3.2 pM (either Forward or Reverse)
- Sequencing Mix - 0.28 µl
- 5x Reaction buffer - 1.86 µl
- Sterile distilled water - make up to 10µl

The sequencing PCR temperature profile consisted of a 1st cycle at 96°C for 2 minutes followed by 30 cycles at 96°C for 30 sec, 50°C for 40 sec and 60°C for 4 minutes for all the primers.

3.2.2.2. Post Sequencing PCR Clean up

1. Make master mix I of 10µl milli Q and 2 µl 125mM EDTA per reaction
2. Add 12µl of master mix I to each reaction containing 10µl of reaction contents and are properly mixed.
3. Make master mix II of 2 µl of 3M sodium acetate pH 4.6 and 50 µl of ethanol per reaction.
4. Add 52 µl of master mix II to each reaction.
5. Contents are mixed by inverting.
6. Incubate at room temperature for 30 minutes
7. Spin at 14,000 rpm for 30 minutes
8. Decant the supernatant and add 100 µl of 70% ethanol
9. Spin at 14,000 rpm for 20 minutes.
10. Decant the supernatant and repeat 70% ethanol wash
11. Decant the supernatant and air dry the pellet.

The cleaned up air dried product was sequenced in ABI 3500 DNA Analyzer (Applied Biosystems).

3.2.3. Sequence Analysis
The sequence quality was checked using Sequence Scanner Software v1 (Applied Biosystems). Sequence alignment and required editing of the obtained sequences were carried out using Geneious Pro v5.6 (Drummond et al. 2012). Further analyses of the sequences were carried out to determine the length of various barcodes, molecular weight, nucleotide composition and percentage of GC content.

3.2.4. GenBank Submission
The sequences obtained were submitted in GenBank database, an annotated collection of all publicly available DNA sequences. This was done using Sequin Application Version 13.05.

3.2.5. BLAST
The basic local alignment search tool (BLAST) was used to determine the similarity of sequences obtained with those in the database. Nucleotide blast was performed (BLASTN).

3.2.6. Multiple sequence alignment
Multiple sequence alignment (MSA) establishes positional correspondence in evolution and only a successful sequence alignment produces a genealogically
related tree. MSA was performed using Clustal W in the BioEdit version 7.0.9.0. (Hall 1999), which support a wide array of file types.

3.2.7. Pairwise alignment

Pairwise alignment was performed to determine the percentage of variation.

3.2.7.1. Single locus

The percentage variation for each of the barcodes was calculated and represented graphically using bardiagram.

3.2.7.2. Multilocus combination

3.2.7.2.1. Concatenation of gene sequences

Different gene combinations were combined to analyze the variation percentage. The gene sequences were concatenated using Geneious 8.1.6 software (Kearse et al. 2012)

The variation percentage using two-locus, three-locus and four-locus combinations were tried out.

3.2.8. Construction and analyses of the dendrogram

The selection of the tree building method was important. Among the various tree building methods Unweighted Pair Group Mean Average-UPGMA (Sokal and Sneath 1973), a distance based method which is based on the degrees of difference between pairs of sequences. Such distance will be used to construct the distance matrix between individual pairs of taxa. UPGMA is a clustering or phenetic algorithm and it is not strictly an evolutionary distance method (Li 1997). The tree was constructed using MEGA 4.0 (Tamura et al. 2007). All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option).

Model adopted was p-distance, which is the proportion (p) of nucleotide sites at which two sequences being compared are different. It is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared (Nei and Kumar 2000). Substitutions to include were d: Transitions + d: Transversions; homogenous pattern among lineages and uniform rates among sites.
3.3. RESULTS AND DISCUSSION

DNA barcoding is a novel technique used for the identification of species using standard DNA segments. DNA barcoding in plants plays an important role in the evaluation of biodiversity especially in the assessment of hotspots. Apart from the routine identification of the rare species, it is also useful in many other aspects.

3.3.1 Genomic DNA isolation and PCR amplification

The images captured after visualization of the genomic DNA isolated and that of the PCR products of the five barcodes rbcL, matK, trnH-psbA, ITS and ITS2 are shown (Fig. 54; A-F).

Fig. 54 (A-F). Images of gel
A. Isolated genomic DNA of the four *Aristolochia* species
B. PCR amplicons of the candidate barcode rbcL in four species
C. PCR amplicons of the candidate barcode matK in four species
D. PCR amplicons of the candidate barcode trnH-psbA in four species
E. PCR amplicons of the candidate barcode ITS in four species
F. PCR amplicons of the candidate barcode ITS2 in four species
3.3.2 Sequencing

The sequences obtained for the five barcodes are given in appendix I.

3.3.3 Sequence analysis

Sequences of the five barcodes obtained were analyzed. The length of rbcL sequenced in all the species remained the same as in the study by Li et al. (2014). Noticeably steady length of rbcL was also reported by Ferri et al. (2015). matK sequences varied in length from 880 to 922. In a study by Li et al. (2014), matK region ranged in length from 837-840 in three different species of Aristolochia, whereas Murata et al. (2001) reported the length of matK from 1,194 to 1,224 bp among the 70 samples of Aristolochia. The length of mat K was found highest among the five barcodes analyzed. Longer fragments for matK were also reported by Ferri et al. (2015) and Li et al. (2012). trnH-psbA ranged from 259-280 bp in the present study, while Li et al. (2014) reported it to range from 230-302 bp. ITS ranged from 622 to 727 bp and ITS2 from 364-396 bp. The length and molecular weight of the five barcodes in the four species are given in table 20.

Sequences of rbcL, mat K and trnH-psbA were of good quality. ITS and ITS2 sequences had a high GC content. Hence, the percentage of GC content in the four species using the five barcodes was calculated to note the variability (Fig. 55). On calculating the GC percentage in the four Aristolochia species, all of them showed a high GC content when ITS and ITS2 are used (more than 65%). In A.indica on using ITS2, GC turned upto 75% but ITS sequencing did not work out. The comparison of the nucleotide composition in the four species is given in table 21. In a study by Li et al. (2014) GC content was found highest with rbcL in A. contorta, A. fangchi and A. manshuriensis followed by trnH-psbA and matK.

Thus, the major results derived from the analysis of the raw sequences were that, matK gave the longest sequence, whereas trnH-psbA gave the shortest sequence. Regarding the GC content, it was found that ITS and ITS2 showed the highest GC percentage and matK with the lowest. These observations were found similar with the studies by Castro et al. (2015) in fig cultivars.
### Results and Discussion

Phytochemical and Molecular Systematics of *Aristolochia* spp.

**Table 20. Size and molecular weight of the five barcodes in *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra* (**bp**-basepair, **D**-dalton)**

<table>
<thead>
<tr>
<th>Barcode</th>
<th><strong>A. indica</strong></th>
<th><strong>A. tagala</strong></th>
<th><strong>A. ringens</strong></th>
<th><strong>A. krisagathra</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>rbcL</strong></td>
<td>Length (bp)</td>
<td>697</td>
<td>697</td>
<td>697</td>
</tr>
<tr>
<td></td>
<td>Molecular weight (D)</td>
<td>423002</td>
<td>422951</td>
<td>422917</td>
</tr>
<tr>
<td><strong>matK</strong></td>
<td>Length (bp)</td>
<td>898</td>
<td>898</td>
<td>922</td>
</tr>
<tr>
<td></td>
<td>Molecular weight (D)</td>
<td>532863</td>
<td>532795</td>
<td>547351</td>
</tr>
<tr>
<td><strong>trnH-psbA</strong></td>
<td>Length (bp)</td>
<td>268</td>
<td>280</td>
<td>259</td>
</tr>
<tr>
<td></td>
<td>Molecular weight (D)</td>
<td>162410</td>
<td>169619</td>
<td>156964</td>
</tr>
<tr>
<td><strong>ITS</strong></td>
<td>Length (bp)</td>
<td>-</td>
<td>628</td>
<td>622</td>
</tr>
<tr>
<td></td>
<td>Molecular weight (D)</td>
<td>-</td>
<td>365618</td>
<td>371714</td>
</tr>
<tr>
<td><strong>ITS2</strong></td>
<td>Length (bp)</td>
<td>364</td>
<td>386</td>
<td>396</td>
</tr>
<tr>
<td></td>
<td>Molecular weight (D)</td>
<td>222859</td>
<td>218424</td>
<td>224486</td>
</tr>
</tbody>
</table>

**Table 21. Nucleotide composition of the five barcodes in *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra**

<table>
<thead>
<tr>
<th>Barcode</th>
<th>Nucleotide</th>
<th><strong>A. indica</strong></th>
<th><strong>A. tagala</strong></th>
<th><strong>A. ringens</strong></th>
<th><strong>A. krisagathra</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>rbcL</strong></td>
<td>A</td>
<td>194</td>
<td>197</td>
<td>196</td>
<td>196</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>145</td>
<td>142</td>
<td>140</td>
<td>144</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>160</td>
<td>160</td>
<td>160</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>198</td>
<td>198</td>
<td>201</td>
<td>198</td>
</tr>
<tr>
<td><strong>matK</strong></td>
<td>A</td>
<td>252</td>
<td>255</td>
<td>265</td>
<td>252</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>173</td>
<td>171</td>
<td>175</td>
<td>172</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>142</td>
<td>140</td>
<td>146</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>313</td>
<td>314</td>
<td>318</td>
<td>313</td>
</tr>
<tr>
<td><strong>trnH-psbA</strong></td>
<td>A</td>
<td>79</td>
<td>85</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>49</td>
<td>50</td>
<td>48</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>53</td>
<td>53</td>
<td>51</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>87</td>
<td>92</td>
<td>85</td>
<td>89</td>
</tr>
<tr>
<td><strong>ITS</strong></td>
<td>A</td>
<td>-</td>
<td>103</td>
<td>105</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>-</td>
<td>223</td>
<td>230</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>-</td>
<td>197</td>
<td>196</td>
<td>193</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>-</td>
<td>75</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td><strong>ITS2</strong></td>
<td>A</td>
<td>47</td>
<td>50</td>
<td>54</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>145</td>
<td>139</td>
<td>143</td>
<td>144</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>128</td>
<td>120</td>
<td>120</td>
<td>119</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>44</td>
<td>48</td>
<td>50</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>
3.3.4 Accession numbers

19 sequences were submitted in the GenBank. The accession numbers of these sequences are given in Table 22.

<table>
<thead>
<tr>
<th>Species</th>
<th>GenBank Accession Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aristolochia indica</strong></td>
<td>KF498583, KF498586, KP763859, -</td>
</tr>
<tr>
<td><strong>Aristolochia tagala</strong></td>
<td>KF498584, KF498587, KP763862, KP763869, KP763864</td>
</tr>
<tr>
<td><strong>Aristolochia ringens</strong></td>
<td>KF498585, KF498588, KP763860, KP763867, KP763865</td>
</tr>
<tr>
<td><strong>Aristolochia krisagathra</strong></td>
<td>KF498589, KF476063, KP763861, KP763868, KP763866</td>
</tr>
</tbody>
</table>

3.3.5 BLAST results

Comparing nucleotide sequences of the same or different species is a very powerful tool and an important aspect in molecular studies. This is usually performed with the blast search. The sequences obtained were compared to the database in NCBI with the nucleotide blast. BLASTN of rbcL gene showed maximum identity of 99% in **A. indica**, **A. tagala** and **A. krisagathra** to **A. zollingeriana**; and
A. ringens to A. reniformis. matK in A. ringens showed 100% identity to A. ringens in GenBank whereas A. indica, A. tagala, A. krisagathra showed 99% similarity to A. tagala from GenBank. For trnH-psbA the identity percentage for the first hits in A. indica, A. ringens and A. krisagathra were 92%, 98% and 91% respectively with A. littoralis. But A. tagala showed 88% similarity with A. clematitis. For ITS, A. tagala (86%) and A. krisagathra (85%) showed similarity with A. foveolata; A. ringens showed 83% similarity with an unknown species of Aristolochia in GenBank. For ITS2, the identity percentage shown was 86%, 88%, 80% and 87% for A. indica, A. tagala, A. ringens and A. krisagathra respectively to A. clematitis.

3.3.6 Multiple sequence alignment

Multiple sequence alignment is the procedure of comparing more than two sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. The significance of sequence alignment is that it allows comparison of individual homologous positions in different sequences. Multiple sequence alignment of the four species of Aristolochia for each barcode is given (Fig. 56 to Fig. 60).

Multiple sequence alignment showed variation of the nucleotides in the four species for the five regions sequenced. Hajibabaei et al. (2007) has mentioned the importance of variable/polymorphic sites and opined that species uniqueness can be validated from a small number of polymorphic sites within the barcode candidate gene. Some of the important variations that can act as unique regions for a particular barcode in identifying a particular species from the other three after multiple sequence alignment are discussed further.

For the rbcL gene (Fig. 56), A. krisagathra did not have any single distinct variable nucleotides to distinguish it from other three species. A. indica had a single variable nucleotide at position 421, A. tagala with two nucleotides (258 and 398) and A. ringens having the maximum of ten. For matK (Fig. 57), A. indica had only a single nucleotide varying (21st). A. tagala did not have any such variable nucleotide, A. krisagathra had 6. The aligned length among the four species for matK was 904
bp. The nucleotides from 220 to 234 and 637 to 645 in mat K gene can be used a distinct barcode region to differentiate *A. ringens* from the other 3 species. For trnH-psbA gene (Fig. 58), 5 nucleotides showed distinct variation in *A. indica* and 7 in *A. krisagathra*. Unique region for *A. tagala* present in trnH-psbA gene were the nucleotides from 199-203 and 226 to 232 and for *A. ringens* they were found at regions from 112 to 118, 204 to 206 and 242 to 244 along with other variable nucleotides. trnH-psbA produced thirteen mononucleotide ‘T’ repeats at region 204 to 216, in *A. tagala* it showed sixteen ‘T’ repeats (201 to 216). Presence of such long mononucleotide repeats, in which one DNA base recurs excessively within the spacer was mentioned by Thomas (2009). For ITS (Fig. 59), in *A. ringens* the variable regions were the nucleotides from 118-122 and 352 to 373. At these positions, nucleotides were completely absent for *A. tagala* and *A. krisagathra* (*A. indica* sequencing did not work). Hence this region can act as a unique identifier for *A. ringens*. 14 bases showed variation in this region for *A. tagala* and 24 bases in *A. krisagathra*. For ITS2 (Fig. 60), nucleotides from positions 87 to 103 forms a unique identifier of *A. ringens* as in other species these regions have gaps and 283 to 288 forms another distinct region. *A. indica* (6 bases), *A. tagala* (12 bases) and *A. krisagathra* (12) also showed some variable nucleotides.

Thus, among the various barcode sequences analyzed after multiple sequence alignment, it was found that rbcl showed the least variability whereas ITS and ITS2 showed higher variability in general. Similar results were obtained by Li et al. (2014) where rbcl was found to have the highest conserved sites among the various barcodes analyzed. Maximum variability of bases shown among the five barcodes for *A. indica* was ITS2, *A. tagala* - trnH-psbA (14bases) and *A. krisagathra* - ITS (24 bases). *A. ringens* showed distinct regions in matK, trnH-psbA, ITS and ITS2 that can definitely serve as unique region for its identification from other three species.
Fig. 56. MSA of rbcL sequences of *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra* with the highlighted variable sites.
Fig. 57. MSA of matK sequences of *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra* with the highlighted variable sites
Fig. 58. MSA of trnH-psbA sequences of *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra* with the highlighted variable sites.
Fig. 59. MSA of ITS sequences of *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra* with the highlighted variable sites.
Fig. 60. MSA of ITS2 sequences of *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra* with the highlighted variable site
3.3.7 Pairwise alignment

3.3.7.1. Single-locus barcode

Pairwise alignment is the procedure of comparing two sequences by searching for a series of individual characters or character patterns that are in the same order in both sequences. The percentage variations between the different sequences for each barcode determined are given (Fig. 61). The results displayed that ITS and ITS2 showed better discrimination, but since ITS sequencing did not work in *A. indica* it cannot be considered reliable. The order of variation percentage shown by the four barcodes excluding ITS was in the order of ITS2 > trnH-psbA > matK > rbcL. Almost a similar result was obtained by Li et al. (2014) in a study of herbal medicinal materials confused with *Aristolochia* herbs where the variation percentage was in the order of trnH-psbA > matK > rbcL. Similar results were also observed by Ferri et al. (2015) at the genus level in forensic botany and Chen et al. (2010). The results by Hou et al. (2013), Song et al. (2012) and Xin et al. (2013) indicated that the ITS2 barcode exhibited higher interspecific variations.

3.3.7.2. Multilocus combinations

The use of multilocus barcodes is considered most effective strategy to single locus studies (Chase et al. 2005; Kress et al. 2005; Newmaster et al. 2008; CBOL Plant Working Group 2009). The percentage variation using the multilocus combinations - two loci (Fig. 62); three and four loci (Fig. 63) were determined. Among the two-locus combination the highest variation was shown by trnH-psbA+ITS2 combination, of which the highest was between *A. ringens* and *A. krisagathra* (17.85%). matK+trnH-psbA+ITS2 showed the highest variation percentage among the three-locus combinations; highest variation between *A. ringens* and *A. krisagathra* (11.99%). When all the four loci were combined and studied, the variation was found lesser than when using two or three locus together.

Though multi-locus combinations have been approved as better to single locus barcode, the percentage variation between the species was lesser than with the single locus barcodes.
Fig. 61. Percentage variation of the pairwise alignment of single locus barcodes between *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra*

Fig. 62. Percentage variation of the pairwise alignment of two locus combination of barcodes between *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra*

Fig. 63. Percentage variation of the pairwise alignment of three and four locus combination of barcodes between *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra*
3.3.8 Construction and analysis of the dendrogram

The most popular form of data for cladistics is the sequences and they have surpassed morphological and other characters used for grouping and classification. The dendrogram was constructed using UPGMA, the distance based method. In this method by grouping sequences (which forms the unit of classification) that are most similar to each other, termed as operational taxonomic units, a tree is built in a stepwise manner. Similar identities are recognized until only two OTUs remain. Rooted trees were produced, having a single node from which other nodes were derived (Fig. 64 A-E).

The rooted tree bifurcated into two branches forming a dichotomous tree. One of these main branches (A) was further again divided into two branches and the other (B) culminated in A. ringens which remained distinct and distant from the other species. These branching patterns of the dendrogram were similar for all the barcodes analyzed.

In rbcL gene, the branch A of the dichotomy formed a node and gave rise to two branches thus producing two branches, one (D) having A. indica and the other branch (C) further subdivided into two with A. tagala and A. krisagathra. Thus A. tagala and A. krisagathra grouped in a clade and formed the sister taxons whereas A. indica formed the sister clade. The optimal tree with the sum of branch length 0.0198 was obtained (Fig. 64-A). Similar branching patterns were obtained for ITS2 except that here A. indica and A. krisagathra formed the sister taxons thereby making a clade (branch C) and A. tagala (branch D) formed the sister clade to this. The sum of branch length obtained was 0.197 (Fig. 64-E).

In the case of matK (Fig. 64-B) and trnH-psbA (Fig. 64-C) similar trees were obtained. In both, the second main branch (A) further got split up into two branches. Of this, one of the branches (D) had A. krisagathra as the leaf; the other branch (C) partitioned into two with A. tagala and A. indica forming the sister taxons and sister clade to A. krisagathra. For matK the optimal tree with the sum of branch length 0.0454 and for trnH-psbA 0.1018 were obtained.

For ITS, the tree obtained had the sum of branch length as 0.184. Since A. indica sequences were not obtained, the second main branch (A) was subdivided only once with A. tagala and A. krisagathra as its leaf (Fig. 64-D).
IV. 3. DNA Barcoding

Phytochemical and Molecular Systematics of Aristolochia spp.

Results and Discussion

Thus, from all these dendrograms it was observed that all the three species *A. indica*, *A. tagala* and *A. krisagathra* formed a part of main bifurcation. *A. krisagathra* was found as a sister taxon to *A. indica* in ITS2 and as a sister taxon to *A. tagala* in rbcL. In trnH-psbA and matK it was obtained as a sister clade to the clade of *A. indica* and *A. tagala*. Considering the morphology of *A. krisagathra*, it was found to be more similar to *A. indica* and may be misidentified by non-taxonomists. On considering the dendrograms build by using various barcodes, it was found that, compared to other barcodes; the tree generated by using ITS2 backed the morphology. In ITS2 *A. indica* and *A. krisagathra* was best closely placed forming the sister taxons.

On the basis of the sequences and dendrograms generated from the present study, it was found that ITS2 could serve as a reliable barcode for the selected
Aristolochia species. Similar conclusions on the utility of ITS2 as reliable barcode on analyzing the multiple barcode regions was obtained in the studies by Chen et al. (2010), He et al. (2012) and Selvaraj et al. (2012). They opined that ITS2 showed a higher discrimination rate compared to other loci. The applicability of ITS2 in discriminating a wide array of plants from the families of Asteraceae, Rutaceae, Rosaceae etc have also been discussed (Gao et al. 2010; Luo et al. 2010; Yao et al. 2010; Pang et al. 2011). Han et al. (2013) studied on the herbarium specimens of 90 year old wherein ITS2 was found reliable with higher identification efficiency. Sun et al. (2012) also advocated ITS2 as an effective barcode region. The use of ITS2 as a barcode region is less studied in Aristolochia. Chen et al. (2010) reported ITS2 as a reliable barcode for medicinal plants and in the present study also ITS2 was found effective which included medicinal species of Aristolochia. Li et al. (2010) had reported trnH-psbA to be the useful barcode for identifying Aristolochia species. However, in the present study ITS2 showed higher variations between closely related Aristolochia species in comparison to other barcodes. Hence, though the number of species taken and analyzed is limited, the study showcases the suitability of ITS2 as a potential candidate barcode for Aristolochia.

Shifting the archetype of species identification is now happening because of plant DNA barcoding. Hence, whatever knowledge that comes out of such barcoding studies can turn out to be helpful in some way or the other. As Ali et al. (2014) rightly said “molecularization of taxonomy is the need of the hour”.
4.1. INTRODUCTION AND REVIEW OF LITERATURE

The term “phylogenetics” derived from the Greek terms phyle and phylon means “tribe” and “race”; and the term “genetikos” imply “relative to birth”, from “genesis” i.e. “birth”. Phylogenetics is the study of evolutionary relatedness amongst groups of organisms e.g., species, populations (Roy, Dasgupta, and Bagchi 2014). The study of the evolutionary history of organisms from bacteria to humans is important and phylogenetic analysis from DNA or protein sequences forms an important tool for such studies. The rate of evolution differs with gene or DNA segment (Dayhoff 1972; Nei 1987; Wilson, Carlson, and White 1977).

According to Roy, Dasgupta, and Bagchi (2014), the term “evolution” can be defined in versatile ways in different contexts. From the biologist perspective, evolution is the development of a biological form from other pre-existing forms or its origin to the current existing form through natural selections and modifications i.e., change across successive generation. The dynamo behind evolution is natural selection in which “unfit” forms are discarded through changes in environmental conditions or sexual selection so that only the fittest are selected (Darwinism). Different mechanisms of acquiring variation have led to the biodiversity of today. These mechanisms include duplication of genes, reorganization of genomes, and genetic exchanges such as recombination, reassortment and lateral gene transfer. The basic mechanism of evolution is genetic mutations that occur spontaneously and mutations are the most often used ones to infer the relationships between genes. Genetic multiplicity forms the source of raw material for the natural selection to act on.

Phylogenetic analysis is the key for illustrating the evolutionary pattern of multigene families (Atchley, Fitch, and Bronner-Fraser 1994; Goodwin, Baumann, and Berger 1996; Ota and Nei 1994) as well as for understanding the adaptive evolution at molecular level (Chandrasekharan et al. 1996; Jermann et al. 1995). According to evolutionary theory, all organisms evolved from one common ancestor, going back to the origin of life. Phylogenetic analysis establishes the relationships
between genes or gene fragments, by inferring the common history of the genes or gene fragment.

The classical way of estimating the relationship between species is to compare their morphological characters (Linnaeus 1758). Taxonomy is still based largely on morphology. The increasingly available molecular information, such as nucleotide or amino acid sequences and restriction fragment length polymorphisms (RFLPs) can also be used to infer phylogenetic relationships. Whether the morphological or molecular approach is preferable is an important topic of debate. However, the use of molecular data for inferring phylogenetic trees has now gained considerable interest among biologists of different disciplines, and it is often used in addition to morphological data to study relationships in further detail.

Molecular data has become one of the most powerful sources for revealing the evolutionary history among organisms (Baldauf 1999; Brown 2001; Graham and Olmstead 2000; Hassanin 2006; Mathews and Donoghue 1999, Nozaki et al. 2003; Soltis, Soltis, and Chase 1999). In most cases, however, only a few molecular markers are employed for reconstructing evolutionary history. Chloroplast gene and multicopy nuclear ribosomal DNA regions are prevalent.

Aristolochiaceae is a conspicuous family with distinctive and showy flowers positioned among the basal angiosperms (APG III 2009; APG II 2003). The family is divided with a total of four to twelve genera and 500 to 600 species, depending on classification; 400 to 500 of those species belong to the subfamily Aristolochiioideae and nearly 100 to the Asaroideae (Neinhuis et al. 2005; Wanke, Gonzalez, and Neinhuis 2006). The latter are distributed mainly in northern temperate regions with East Asia as centre of diversity, the former mostly in tropical to subtropical species-rich areas in Central and South America (Schmidt 1935). Geographic areas richest in endemism are China, Mexico, Brazil, and Hispaniola (Pfeifer 1966; González et al. 2014). The distribution of Aristolochiaceae is given (Fig. 65). Aristolochia comprises about 120 (Huber 1993) or 400 to 500 (Wanke, Gonzalez, and Neinhuis 2006) species depending on the differentiation within the subgenera. The plants grow as perennial rhizomatous herbs or deciduous or evergreen shrubs and lianas.
IV. 4. Phylogenetic Relationship

Introduction and Review of Literature

Phytochemical and Molecular Systematics of Aristolochia spp.

Fig. 65. The distribution of Aristolochiaceae (Jinshuang 1990)

4.1.1. Limitations on circumscription of the genus

The genus *Aristolochia* has been treated in its broad sense by many authors (Duchartre 1854, 1864; Hoehne 1942; Hou 1984; Ma 1989; Nardi 1984, 1991; Pfeifer 1966, 1970). However, many generic segregates have been proposed and the splitting primarily based on floral, and fruit characters such as the morphology of the gynostemium, the gross shape of the perianth, the dehiscence of fruits and the morphology of the seeds (Gonzalez and Stevenson 2002).

The circumscription of the genus has been a significant taxonomic problem (Gonzalez and Stevenson 2002). In some treatments, a single genus *Aristolochia* sensu lato (s.l.), consisting of three subgenera, *Aristolochia* (*OrthoAristolochia*), *ParAristolochia* (Hutchinson and Dalziel) O. C. Schmidt, and *Siphisia* (Raf.) Duch., has been recognized (Schmidt 1935; Ma 1989; Gonzalez 1999; Gonzalez and Stevenson 2000). Hutchinson and Dalziel (1954), Poncy (1978), and Parsons (1996) recognized two genera, *Aristolochia* and *ParAristolochia* Hutchinson and Dalziel (1927), and Klotzsch (1859) recognized five genera. In contrast, Huber (1993) recognized eight genera in two subtribes, *Aristolochia* sensu stricto (s.s.), *Einomeia*
IV. 4. Phylogenetic Relationship

Introduction and Review of Literature

Phytochemical and Molecular Systematics of Aristolochia spp.

Raf., *Euglypha* E. Chodat & Hassl., *Holostylis* Duch., *Howardia* Klotzsch, and *ParAristolochia* in the subtribe Aristolochiinae H. Huber, and *Endodeca* Raf. and *Isotrema* Raf. in the subtribe Isotrematinae H. Huber. Gonzalez and Stevenson further in 2002 examined the relationships within *Aristolochia* s.l. based on a cladistics analysis of a number of morphological characters. Their results showed that *Einomeia*, and the monotypic *Euglypha* and *Holostylis*, which were accepted as distinct genera by Huber (1993), were nested within *Aristolochia* s.s. They also indicated that the genus s.l. consists of two major lineages that were congruent with the subtribes Aristolochiinae and Isotrematinae of Huber (1993). They proposed that *Aristolochia* s.l. should be divided into four subgenera in two subtribes: *Aristolochia* s.s. and *ParAristolochia* in Aristolochiinae, and *Endodeca* and *Isotrema* in Isotrematinae. Kelly and Gonzalez (2003) substantiated the data, and also demonstrated closer relationships between *Aristolochia* s.s. and *ParAristolochia*, and between *Endodeca* and *Isotrema* based on a more detailed cladistic analysis of Aristolochiaceae. Comparing this system with that in which three subgenera of *Aristolochia* s.l. are recognized, *Aristolochia* s.s. and *ParAristolochia* of the subtribe Aristolochiinae correspond to the subgenera *Aristolochia* and *ParAristolochia* respectively, and the subtribe Isotrematinae is congruent with the subgenus *Siphisia*. Ohi-Toma et al. (2006) studied the molecular phylogeny of *Aristolochia* s.l. and proposed two lineages corresponding to the subtribes Aristolochiinae and Isotrematinae each further divided into two forming the four subgenera in congruence with the morphology based phylogeny proposed by Gonzalez and Stevenson (2002). Thus, the taxonomic frameworks are largely identical in several systems, but the recognitions of their taxonomic ranks are different.

4.1.2. **Generic segregates of *Aristolochia***

The taxonomic treatment of the genus *Aristolochia* is ambiguous. Several generic segregates have been proposed for the genus. Most literatures however propose four subgenera within two subtribes (Gonzalez and Stevenson 2002; Kelly and Gonzalez 2003; Ohi-Toma et al. 2006). The genus *Aristolochia* s.l. comes under the tribe Aristolochiaceae with two subtribes Isotrematiiinae and Aristolochiinae. The
subtribe Isotrematiinae encloses two subgenera *Endodeca* and *Isotrema*; Aristolocholochiineae subtribe with *ParAristolochia* and *Aristolochia*.

### 4.1.2.1. Subgenus *Endodeca*

Herbaceous group in North America, with reduced subtending leaves, clasping bracts, short internodes in the laterals, and prostrate inflorescence (Huber 1993; Gonzalez and Stevenson 2002).

### 4.1.2.2. Subgenus *Isotrema* (=Siphisia)

Comprises of 70 species, 50 of which grow in Eastern Asia (mainly the Sino-Japanese province), including Japan, Hainan and Taiwan to the east and the Eastern Himalayan and the Manchurian provinces to the west (Takhtajan 1986). Two species are found in eastern Malaysia (*A. singalanguensis* Korthals ex. Ding Hou in Sumatra, *A. coadunata* Back. in Sumatra and Java) and one (*A. punjabensis* Lace) in the Western Himalayan province (western Pakistan, northwestern India and Nepal). In the Americas, four species (*A. macrophylla, A. reticulata, A. serpentaria* and *A. tomentosa*) are restricted primarily to the Eastern Deciduous Forest and the Coastal Plain (Graham 2011) and one (*A. californica*) is restricted to the Californian province; the remaining 15 species grow in south-eastern Mexico, Guatemala, Belize, Honduras, El Salvador, Nicaragua and Panama. Thus, *Isotrema* has a primarily intercontinental Asia to North/Central America disjunct distribution. Within the genus *Aristolochia*, shrubs and shrub-like growth forms are limited to a handful of species recorded within *Isotrema* (Pfeifer 1966); characterized by basally dehiscent capsules, a trilobed perianth tube and a gynostemium with three lobes each accompanying two anthers (Murata et al. 2001)

Sometimes, *Siphisia* is used for the combination of *Endodeca* and *Isotrema* or for *Isotrema* alone.

### 4.1.2.3. Subgenus *Aristolochia*

Contains approximately 350 species of mostly woody tropical climbers and are widely distributed. However, herbaceous forms, including species from the Mediterranean region, are also known (Neinhuis et al. 2005). The group is characterized by an apically dehiscent capsule, lobes of the perianth unilaterally

Phytochemical and Molecular Systematics of *Aristolochia* spp.
IV. 4. Phylogenetic Relationship

Introduction and Review of Literature

Phytochemical and Molecular Systematics of Aristolochia spp. 

appressed in the bud and breaking up into one to three segments, six or fewer lobes of the gynostemium, and six or fewer anthers. They are morphologically quite diverse and several lower taxonomic ranks recognized (Schmidt 1935; Gonzalez 1999).

4.1.2.4. Subgenus ParAristolochia

The smallest group (approximately 35 species), distributed in tropical Australia, Asia and Africa, contains exclusively climbers. About 33 species in the Southern Hemisphere are recognized, one in Kalimantan, 22 in Papua New Guinea and northeast Australia, and about 10 in western tropical Africa (Gonzalez 1999; Gonzalez and Stevenson 2002; Hutchinson and Dalziel 1954; Parsons 1996). The subgenus was proposed as the most primitive group in Aristolochia s.l. because of its morphological similarity to Thottea s.l. by Ma (1989).

It has fleshy indehiscent fruits, a three-lobed perianth, valvate in bud, sometimes with one or three long tails, six to 12 lobes of the gynostemium, six to 24 anthers.

The phylogeny of A. indica, A. tagala, A. ringens and A. krisagathra with the different segregate members were analyzed to study the relationships.
4.2. MATERIALS AND METHODS

The phylogenetic relationships of the four major groups proposed both by Gonzalez and Stevenson (2002) based on morphological data and by Neinhuis et al. (2005) based on molecular data was examined.

In order to infer the phylogenetic relationship of the Aristolochia species, the statistics of various barcodes in the NCBI database was searched. The statistics of the five barcode sequences deposited in NCBI database was noted (Fig. 66)

Fig. 66. Percentage of various barcode sequences deposited in GenBank

As the number of sequences for trnH-psbA (3), ITS (7 of which 3 sequences were of A. mollisima) and ITS2 (8) was very few compared to rbcL and matK. Being the “agreed upon – core barcodes” by the CBOL PWG (2009), these were used for the phylogenetic analyses in the present investigation.

From the available rbcL and matK sequences of Aristolochia in NCBI database, the selection criteria was

- Common species for which sequences were available for both rbcL and matK
- Unidentified species were avoided
- When more than one sequence was available for a particular species, those which had the longer sequence length was selected.

Thus a total of 26 species were selected and these along with the four species A. indica, A. tagala, A. ringens and A. krisagathra were used for phylogenetic analysis.
IV.4. Phylogenetic Relationship

The accession numbers of the sequences as retrieved from GenBank is given in table 23. These species were grouped according to the classification as proposed by Gonzalez and Stevenson (2002) and is given in table 24. The genus Aristolochia is monophyletic and is sister to Thottea (Ohi-Toma 2006). Hence 3 species of Thottea (T. tomentosa, T. borneensis and T. macrantha) was selected as outgroup. These made a final set of 33 sequences.

4.2.1. Phylogenetic analysis

The number of conserved sites, variable sites and parsimony-informative sites were determined using Mega 4.0 (Tamura et al. 2007).

4.2.1.1. Distance matrix and interspecific distance

As part of the phylogenetic analysis, the pairwise genetic distance was calculated using Kimura-2-parameter (K2P) model in Mega 4.0 (Tamura et al. 2007). K2P model takes into consideration the transitional and transversional substitution rates, though assumes that the four nucleotide frequencies are the same and that rate of substitution do not vary among sites (Nei and Kumar 2000). The distance matrix obtained was used to determine the overall average interspecific distance and species refinement ability.

4.2.1.2. Construction and analysis of the phylogenetic tree

The phylogenetic tree construction with the coding nucleotide sequences was done using the neighbour joining, a distance based method in Mega 4.0 (Tamura et al. 2007). Distance-based methods use the amount of dissimilarity (the distance) between two aligned sequences to derive trees. Neighbour Joining is clearly the fastest procedure and generally yields a tree close to the Minimum Evolution tree (Rzhetsky and Nei 1992; Li 1997). All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). Codon positions included were 1st+2nd+3rd+Noncoding. Model adopted was K2P; substitutions to include were d:Transitions + Transversions; homogenous pattern among lineages and uniform rates among sites. The reliability of the inferred tree was evaluated using bootstrap analysis with 1000 bootstrap replicates. Bootstrap values greater than 50% have only been indicated. It has been considered to be a measure of
Phylogenetic Relationship

Materials and Methods

accuracy—a biologically significant parameter that gives the probability that the true phylogeny has been recovered. The simulation studies have suggested that, under favourable conditions (roughly equal rates of change, symmetric branches), bootstrap values greater than 70% correspond to a probability of greater than 95% that the true phylogeny has been found (Hillis and Bull 1993).

Table 23. Accession numbers of rbcL and matK sequences retrieved from GenBank

<table>
<thead>
<tr>
<th>Species</th>
<th>Accession number (rbcL)</th>
<th>Accession number (matK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.serpentaria</td>
<td>KJ773281.1</td>
<td>AB211558.1</td>
</tr>
<tr>
<td>A.convolvulacea</td>
<td>AB205589.1</td>
<td>AB211557.1</td>
</tr>
<tr>
<td>A.tomentosa</td>
<td>AB205598.1</td>
<td>AB071814.1</td>
</tr>
<tr>
<td>A.manshuriensis</td>
<td>JX944488.1</td>
<td>AB060748.1</td>
</tr>
<tr>
<td>A.californica</td>
<td>GQ248551.1</td>
<td>AB071813.1</td>
</tr>
<tr>
<td>A.arborea</td>
<td>AB205586.1</td>
<td>AB060739.1</td>
</tr>
<tr>
<td>A.macrophylia</td>
<td>KF724349.1</td>
<td>AB060742.1</td>
</tr>
<tr>
<td>A.kaempferi</td>
<td>AB205593.1</td>
<td>AB180166.1</td>
</tr>
<tr>
<td>A.promissa</td>
<td>AB205602.1</td>
<td>AB211590.1</td>
</tr>
<tr>
<td>A.clematitidis</td>
<td>HE963329.1</td>
<td>AB060767.1</td>
</tr>
<tr>
<td>A.pistolochia</td>
<td>AF543711.1</td>
<td>AF543724.1</td>
</tr>
<tr>
<td>A.debilis</td>
<td>JX944485.1</td>
<td>AB060768.1</td>
</tr>
<tr>
<td>A.contorta</td>
<td>JX944486.1</td>
<td>AB211581.1</td>
</tr>
<tr>
<td>A.lutea</td>
<td>HE963332.1</td>
<td>HE970672.1</td>
</tr>
<tr>
<td>A.fordiana</td>
<td>KP094727.1</td>
<td>KP093790.1</td>
</tr>
<tr>
<td>A.zollingeriana</td>
<td>AB205599.1</td>
<td>AB071817.1</td>
</tr>
<tr>
<td>A.micrantha</td>
<td>AB205595.1</td>
<td>AB071815.1</td>
</tr>
<tr>
<td>A.burelæ</td>
<td>AB205587.1</td>
<td>AB060788.1</td>
</tr>
<tr>
<td>A.maxima</td>
<td>AB205594.1</td>
<td>AB060784.1</td>
</tr>
<tr>
<td>A.pilosa</td>
<td>KJ594107.1</td>
<td>AB060802.1</td>
</tr>
<tr>
<td>A.littoralis</td>
<td>GU135225.1</td>
<td>AB211571.1</td>
</tr>
<tr>
<td>A.gigantea</td>
<td>AB205591.1</td>
<td>AB060793.1</td>
</tr>
<tr>
<td>A.grandiflora</td>
<td>AB205592.1</td>
<td>AB071812.1</td>
</tr>
<tr>
<td>A.cordiflora</td>
<td>KJ594102.1</td>
<td>KJ593770.1</td>
</tr>
<tr>
<td>A.tondzii</td>
<td>KJ594108.1</td>
<td>JQ586748.1</td>
</tr>
<tr>
<td>A.reniformis</td>
<td>AB205600.1</td>
<td>AB211568.1</td>
</tr>
<tr>
<td>T.borneensis</td>
<td>AB205604.1</td>
<td>AB211554.1</td>
</tr>
<tr>
<td>T.tomentosa</td>
<td>AB205606.1</td>
<td>AB060738.1</td>
</tr>
<tr>
<td>T.macrantha</td>
<td>AB205605.1</td>
<td>AB211555.1</td>
</tr>
</tbody>
</table>
Table 24. Various generic segregates of *Aristolochia* spp. used to infer phylogenetic relationships

<table>
<thead>
<tr>
<th>Species</th>
<th>Tribe</th>
<th>Subtribe</th>
<th>Subgenera</th>
<th>Section</th>
<th>Subsection</th>
<th>Series</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. serpentina</em></td>
<td><strong>ARISTOLOCHIEAE</strong></td>
<td>ISOTREMATINAE</td>
<td>ENDODECA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. convolvulacea</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. tomentosa</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. manshuriensis</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. californica</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. arborea</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. macrophylla</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. kaempferi</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. promissa</em></td>
<td>ARISTOLOCHIEAE</td>
<td></td>
<td>PARARISTOLOCHIA</td>
<td>(Old</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>world)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. clematits</em></td>
<td>ARISTOLOCHINAE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. pistolochia</em></td>
<td>ARISTOLOCHINAE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. debilis</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. contorta</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. lutea</em></td>
<td>ARISTOLOCHINAE</td>
<td></td>
<td>ARISTOLOCHIA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. fordiana</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. zollingeriana</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. tagala</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. indica</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. krisagathra</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. micrantha</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. burelae</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. cordiflora</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. pilosa</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. littoralis</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. grandiflora</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. reniformis</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. gigantea</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. ringens</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. tonduzii</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. maxima</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.3. RESULTS AND DISCUSSION

The two “core barcodes” rbcL and matK was used to infer the phylogenetic relationship of Aristolochia species. Studies on the phylogeny of Aristolochia species using the molecular data have focussed on mainly the chloroplast regions matK, rbcL, phy A, trnL-trnF sequences (Murata et al. 2001; Neinhuis et al. 2005; Ohi-Toma et al. 2006; Silva Brandao et al. 2006; Watanabe, Kajita, and Murata 2006).

4.3.1. rbcL Phylogeny

The rbcL sequences generated for Aristolochia were aligned and the total aligned length including the sequences of the outgroup species was 506 forming the final data set. The number of conserved sites was 464. Out of the 45 variable sites 32 were parsimony informative.

4.3.1.1. Distance matrix

Overall average interspecific K2P distance for the 33 sequences of rbcL analyzed was 0.0196. The highest interspecific distance was observed for A. kaempferi and A. pilosa (0.0363). A. pistolochia - A. zollingeriana, A. contorta - A. debilis and A. littoralis - A. grandiflora-A. cordiflora - A. gigantea - A. reniformis formed clusters with zero interspecific distance.

4.3.1.2. Neighbour-joining tree based on rbcL

The phylogenetic tree of rbcL showed that Aristolochia is monophyletic and embraces two major clades indicating two major lineages (Fig. 67) but without strong support. One of the major clade was Aristolochiinae and the other major clade further subdivided with ParAristolochia in one branch and Isotrematinae (Endodeca and Isotrema) as the other. The Isotrematinae with Endodeca and Isotrema was monophyletic with a strong support as observed in a study by Ohi-Toma et al. (2006). The placement of ParAristolochia represented by a single species A. promissa in the study, within the Isotrematinae was found odd and against the
classification by Gonzalez and Stevenson (2002) and other related studies (Kelly and Gonzalez, 2003; Neinhuis et al. 2005; Ohi-Toma et al. 2006). This relationship gains no significant support from bootstrapping also. The clade *Endodeca* included *A. serpentaria* and *A. convolvulacea* which showed zero interspecific distance in the distance matrix. *Isotrema* with 6 species - *A. arborea*, *A. kaempferi*, *A. manshuriensis*, *A. macrophylla*, *A. californica* and *A. tomentosa* of which *A. manshuriensis*, *A. macrophylla*, *A. californica* and *A. tomentosa* formed a single cluster with no interspecific distance.

The other major clade treated as Aristolochiinae showed several subclades. *A. indica*, *A. tagala*, *A. ringens* and *A. krisagathra* were grouped in the Aristolochiinae clade with *A. indica* and *A. krisagathra* forming the sister taxons and sister clade to *A. tagala*. *A. indica*, *A. tagala* and *A. krisagathra* came under the Diplolobus section coming under the Old World and Podanthemum subsection. However, Aristolochia subsection showed splitting in the phylogenetic tree and were placed in different clades. *A. ringens* was placed in Haplolobus section coming under New World. *Thottea* was basal to the other taxa and is supported strongly (bootstrap value – 100).

The species pairs which showed zero distance were *A. gigantea X A. grandiflora*, *A. littoralis X A. grandiflora*, *A. cordiflora X A. grandiflora*, *A. littoralis X A. gigantea*, *A. reniformis X A. gigantea*, *A. cordiflora X A. gigantea*, *A. cordiflora X A. littoralis*, *A. macrophylla X A. manshuriensis*, *A. californica X A. manshuriensis*, *A. debilis X A. contorta*, *A. californica X A. macrophylla*, *A. tomentosa X A. macrophylla*, *A. tomentosa X A. californica*, *A. zollingeriana X A. pistolochia*, *A. cordiflora X A. reniformis*. These involved 15 of the 33 species studied and the species discrimination rate was 45.45%.
IV. 4. Phylogenetic Relationship

Results and Discussion

Phytochemical and Molecular Systematics of Aristolochia spp.

205

4.3.2. matK phylogeny

The matK sequences generated were aligned and the total aligned length forming the final data set including those of outgroup species was 768 bp. The number of conserved sites was found to be 539. 218 variable sites were obtained in the dataset of which 140 were parsimony informative.

Fig. 67. Neighbour joining tree of 33 plant species based on rbcL (sequences generated from the present study)
4.3.2.1. **Distance matrix**

The distance matrix obtained using K2P model with 33 matK sequences showed an overall average interspecific distance of 0.0623. All the species were successfully discriminated by matK sequences.

4.3.2.2. **Neighbour-joining tree based on matK sequences**

The matK tree generated (Fig. 68) was much strongly supported with higher bootstrap values. The tree obtained was in congruence with the classification as proposed by Gonzalez and Stevenson (2002). The subfamily Aristolochioideae divided into two tribes Bragantiaeae (with *Thottea*) and Aristolochiaeae (with *Aristolochia*). *Thottea* treated as outgroup was basal to *Aristolochia* which is strongly supported. *Aristolochia* was found monophyletic and gave rise to two major clades representing two strongly supported lineages Isotrematinae and Aristolochiinae as of Huber (1993).

In matK phylogeny, the Isotrematinae had two subclades representing *Endodeca* and *Isotrema*; and Aristolochiineae subdivided into *ParAristolochia* and *Aristolochia*. The subgenera *Aristolochia* was further divided into two, one with the series Thyrsicae. The other included the section Diplolobus and other members of Gymnolobus which was further divided into three, one with a member of series Pentandrae (*A. micrantha*), Diplolobus members and other Gymnolobus members. In the Diplolobus section the members of the subsections Podanthemum and *Aristolochia* was correctly placed forming two major clades of Diplolobus.

The results also supported the sister group relationship of *Endodeca* and *Isotrema*; subgenera *ParAristolochia* and *Aristolochia* also showed sister group relationship and were found to be monophyletic. These results were in congruence with the studies of Wanke (2006). It was also found correlated with the studies by González and Stevenson (2002), Kelly and González (2003) and Neinhuis et al. (2005) who supported the monophyly of the generic segregates *Endodeca* and *Isotrema*. Murata et al. (2001) conducted a phylogenetic analysis of *Aristolochia* s. l. based on matK sequences. Although *ParAristolochia* was not sampled, their results supported the monophyly of subgenera *Isotrema* and *Aristolochia*. 

Phytochemical and Molecular Systematics of *Aristolochia* spp.
IV. 4. Phylogenetic Relationship

Results and Discussion

Phytochemical and Molecular Systematics of Aristolochia spp.

Fig. 68. Neighbour joining tree of 33 plant species based on matK (🌟 sequences generated from the present study)

Two morphological features supporting the sister-group relationship within the two main clades of subgenus Isotrema, the segregates Isotrema and Endodeca explained by Gonzalez and Stevenson (2000) and Wanke (2006) are a three-lobed gynostemium, and an annulus, a ring-like structure around the mouth of the...
perianth that is variously modified. The sister group relationship between subgenera *ParAristolochia* and *Aristolochia* is supported by the presence of conical trichomes, especially throughout the inner surface of the perianth tube which are partially responsible for the trap mechanisms of the flowers.

The *Aristolochia* clade was further divided into subclades forming section Diplolobus and Gymnolobus; the section Diplolobus with subsection Podanthemum and subsection Aristolochia. Podanthemum and Aristolochia showed sister relationship and found monophyletic as in previous studies (Ohi-Toma et al. 2006; Wanke 2006). Most of these lineages gained strong support from bootstrapping and also were found correlated with the studies by Ohi-Toma et al. (2006). All the species studied were fully resolved using matK phylogeny.

The phylogenetic analysis using rbcl and matK core barcodes revealed matK to be more useful in resolving species compared to rbcl as the natural groups were clustered together mostly. Moreover, all the species were fully resolved using matK. matK tree was more strongly supported with the bootstrap values. matK also showed a higher number of parsimony-informative characters in comparison to rbcl. A better measure of phylogenetic information is made available through the parsimony-informative characters (Mort et al. 2007). On comparing the phylogenetic trees, placing *ParAristolochia* within the subtribe Isotrematineae by using the rbcl phylogeny was a major drawback. However, matK phylogeny almost resembled the pattern of classification as proposed by Stevenson and Gonzalez (2002), than with the classification in which *Aristolochia* s.l. is deemed to be one large genus with three subgenera (Ma 1989; Gonzalez 1999; Gonzalez and Stevenson 2000; Schmidt 1935). Both the rbcl and matK phylogenies made clear distinction between *Endodeca* and *Isotrema*; stated monopyletic origin, with strong support. Ma (1989) proposed that subgenus *ParAristolochia* is the most primitive group in *Aristolochia* s.l. because of its morphological similarity to *Thottea* s.l. However, this is not supported by the molecular phylogeny as observed by Ohi-toma et al. (2006). The subgenera *Isotrema* was placed as the primitive group. In rbcl and matK phylogenies, *A. krisagathra* was placed as the sister taxon to *A. indica* with strong support.