Summary and Conclusions

Painting by: Ms. Hemalata Pradhan
6 SUMMARY AND CONCLUSIONS

The investigations presented in the thesis on evaluation of the recommended loci for DNA barcoding of orchid species were initiated at a time when the applicability of this technology to plants had become suspect because of (i) CO1, the locus proposed as a universal barcode for eukaryotes, was not found to be suitable for the plants, except for some microalgae and (ii) testing of a number of loci from the chloroplast genome and a locus from the nuclear genome in floristic assemblages of species did not result in recognition of any of the tested loci as a universal barcode for plants. Consequently, in 2006 the illusive universal DNA barcode for plants was considered akin to the “Holy Grail”, which is yet to be found in spite of centuries of search. However, the conclusion about the possibility of a multi-locus barcode that could provide a satisfactory resolution of plant species had emerged. With this background, the present investigations were initiated with two main objectives. These were (i) to check the applicability of the concept of DNA barcoding to plants and (ii) to identify the locus/loci, which individually or in combinations, could provide a core barcode for plants. During the period of this study (2007-2011), many recommendations, predominantly based on floristic studies, for the possible plant barcodes appeared and therefore, the additional objective of the thesis was (iii) evaluation of the recommended loci for the DNA barcoding of plants. The members of the family Orchidaceae are generally believed to be promiscuous in nature as evidenced by the existence of many natural hybrids and ease of facilitated hybridization. Moreover, occurrence of cryptic and sister species is also prevalent among the members of this family. In India, this family is adequately represented by having 1300-1600 species belonging to 186 genera. These include many which are of floricultural and medicinal importance and have a lot of demand in market. Because of their dwindling populations in nature, all orchids have been included under Appendix II of CITES with some even under Appendix I. Thus, this family was selected with the expectation that this would provide a robust test to the DNA barcoding as such and the selected loci in particular. Moreover, it was envisaged that a more stringent test would be if these were tested at multiple taxonomic levels. Therefore, species discrimination capabilities of the loci were tested at different taxonomical hierarchical levels within
the family Orchidaceae and also at interfamily level by including plant species from families other than Orchidaceae. Among a number of loci tested till 2007, five loci from the chloroplast genome and one from the nuclear genome had provided encouraging results and various combinations of these were suggested as core barcode for plants. These loci were \textit{matK}, \textit{rbcL}, \textit{rpoB}, \textit{rpoC1} and \textit{trnH-psbA} spacer from the chloroplast genome and \textit{nrITS} from the nuclear genome. From the above-mentioned six loci, four from the chloroplast genome (\textit{matK}, \textit{rbcL}, \textit{rpoB} and \textit{rpoC1}) and one from the nuclear genome (\textit{nrITS}) were selected for testing their ability in identifying members of Orchidaceae, either individually or in different combinations. The \textit{trnH-psbA} spacer was not tested due to its numerous limitations reported previously by various workers and repeated failure to obtain its bidirectional sequences from \textit{Dendrobium} species, which were being experimented in a parallel project in the laboratory during the course of this study.

6.1 PLANT MATERIALS AND METHODOLOGY

A total of 435 accessions belonging to 104 species of Orchidaceae were collected/procured from different locations in India. In addition, 95 individuals belonging to 68 other families of vascular plants were also included. To evaluate intraspecific variations, multiple accessions of 78 species were collected or obtained. While collecting different accessions of a species, it was ensured that there were no vegetative connections between two accessions. The number of accessions per species varied from 1 to 12 with three having more than 10 accessions each, 25 species were represented by 6-10 individuals, 39 by 3-5 individuals and 11 by 2 individuals. The remaining 26 had one accession each. Out of 104 collected/procured species, herbarium specimens were prepared for 50 collected species and voucher numbers for each were obtained after submitting these to Botanical Survey of India, Dehradun (BSD) or Delhi University Herbarium (DUH) at the Department of Botany, University of Delhi. Some orchids were maintained in a chick house on moss sticks or in baskets in the botanical garden of the Department, where some of these flowered. However, none of the plants survived beyond two years. The tissues, leaves, pseudo-bulbs or stem pieces from the collected plants, stored at -20°C, were used for isolation of
DNA, which was carried out either by CTAB method or using the DNA extraction kit. For tissues having high mucilage content, a modified CTAB method was used. After ascertaining the quality of DNA, the selected loci were amplified by using primers which were already available on Kew website or in related literature. The primers used for the loci from the chloroplast genome were slightly modified after aligning them with the whole chloroplast genome sequence of *Phalaenopsis aphrodite*, an orchid. The amplicons yielding single band were directly sequenced after purification by Exo-SAP method. In cases where multiple bands were obtained, the band having molecular weight closest to that of the targeted locus was gel extracted for sequencing. The purified amplicons were directly sequenced in both directions by Sanger’s di-deoxy chain termination method on an automated sequencer ABI prism 3700. The software ‘Phred’ was used for checking the quality of the sequences and only those with a minimum quality value of 20 were selected. For further analysis Sequencher software was used for making contigs and joining of two sequences (forward and reverse) of the amplicons. The identities of the sequences were confirmed by BLAST analysis each sequence on NCBI. Thereafter, the sequences of each locus were aligned using ClustalW and the aligned sequences were analyzed for phylogenetic distances by MEGA 4, which was also used for constructing Neighbor-joining (NJ) phylogenetic trees with 1000 bootstrap replicates. Intra-specific and inter-specific distances were calculated using Kimura-2-parameter (K2P) model. The percent species resolutions for each locus were calculated on the basis of K2P distances, phylogenetic tree and BLAST methods. In the first method, the species resolution was calculated by preparing a K2P distance matrix of all the species from their aligned DNA sequences of a particular locus. Two species were considered as distinct, if their inter-specific K2P distance was more than the maximum intra-specific distance, while in the phylogenetic tree method, the species resolution was based on the analyses of the Neighbor-Joining (NJ) trees constructed using the consensus sequences of a particular locus of each species, in which no intra-specific variation was observed, along with sequences of all the individuals of the species which exhibited intra-specific variation. The third method used for species resolution involved BLAST analysis of all the sequences individually as the query sequence. The first hit of the query sequence on the database was considered as the identity of the unknown
specimen. Three data sets were used for comparing the species resolution of the loci individually. Data set I comprised the sequences of 191 plant species belonging to different families of vascular plants for all the loci generated in the present study, while the data set II had sequences of each of the tested loci from 104 species of various genera of Orchidaceae investigated presently. In the Data set III the species discrimination ability of five loci was tested among congeneric species of 20 analyzed orchid genera.

### 6.2 RESULTS

The five loci, *matK, rbcL, rpoB* and *rpoC1*, from the chloroplast genome and ITS, from the nuclear genome, were compared for their amplification, sequencing and species discrimination success rates among 530 accessions belonging to 191 species of vascular plants. Among the loci tested, *rpoB* and *rpoC1* exhibited the highest (97%) amplification success rate, followed by *rbcL* (95.28%), ITS (84.73%) and *matK* (82.64%). Likewise, the sequencing success was the highest (91.71%) for *rpoC1*, and 91.68% for *rpoB* and the remaining three loci viz., *rbcL*, *matK* and ITS exhibited 89.9%, 88.12% and 85.52% sequencing success rates, respectively. A similar trend in amplification and sequencing success rates was observed for the orchid data set. A total of 2,174 barcode sequences were generated and submitted to GenBank, NCBI and accession numbers were obtained.

BLAST search revealed that all sequences generated in the present study were of only the targeted loci. Moreover, to confirm that the sequences generated in the present study were only of corresponding species and not of contaminations of fungi, especially in case of ITS or from the host tissue (some orchids being epiphytic), a BLAST analysis for sequences of each of the tested species was performed. It was observed that all the sequences closely matched with only the similar species or genus or recovered as unique sequence.

The results arrived at, based on the analysis of the three data sets, are summarized under their respective heads.
6.2.1 Analyses Based on Data Set I

6.2.1.1 Inter-specific Distances

The maximum average inter-specific K2P distance of 0.164 with a range of 0-0.408 was obtained for ITS sequences. Whereas, the \textit{rbcL} sequences had the minimum inter-specific variations with 0.071 (range 0-0.272) average inter-specific K2P distance. The average inter-specific K2P distances for \textit{matK}, \textit{rpoB} and \textit{rpoC1} were 0.147 (range 0-0.429), 0.106 (range 0-0.412) and 0.095 (range 0-0.393), respectively.

6.2.1.2 Species Resolution

Based on genetic distance method, ITS provided highest species resolution of 84.73% while, minimum was afforded by \textit{rpoC1} (66.47%). The percent species resolutions of \textit{matK}, \textit{rbcL}, and \textit{rpoB} were 82.6%, 73.83% and 69.89%, respectively.

The percent species resolutions provided by the five loci using tree based method, with a couple of exceptions, were exactly similar to the values obtained through genetic distance method, as all the individuals of the species exhibiting intra-specific variation clustered together. However, the values obtained through BLAST methods were slightly different and higher than the discrimination rates obtained with the former two methods for ITS, \textit{matK} and \textit{rpoB} loci. However, the species resolution values based on \textit{rbcL} and \textit{rpoC1} decreased in the BLAST analysis as compared to genetic distance and tree methods. The highest species resolution (92.02%) was provided by \textit{matK}, followed by ITS with slightly less resolution (91.6%). The species resolution values for \textit{rbcL}, \textit{rpoB} and \textit{rpoC1} were 66.86%, 70.43% and 61.93%, respectively. ITS provided the highest species discrimination rates using genetic distance and phylogenetic tree methods whereas, using BLAST analysis, the rate was slightly lower than \textit{matK}. Irrespective of the method used, the species resolution provided by \textit{matK}, among the loci tested from the chloroplast genome, was the maximum, while the success with \textit{rpoC1} was the minimum. Between, \textit{rpoB} and \textit{rbcL}, the former yielded better species resolution than the latter.
6.2.2 Analyses Based on Data Set II

6.2.2.1 Intra- and Inter-specific Distances

The computational analysis revealed that the individuals of 78 orchid species with multiple accessions exhibited intra-specific variations of variable range for all the five tested loci. The individuals of nine species exhibited intra-specific variations in the range of 0-0.004 in their *matK* sequences. On the other hand, ITS exhibited the highest intra-specific K2P distances up to 0.003 among individuals of four species. The highest intra-specific K2P distances for *rbcL*, *rpoB* and *rpoC1* were 0.002, 0.006 and 0.005, among individuals of two, four and three species, respectively.

Average inter-specific K2P distances calculated for ITS, *matK*, *rbcL*, *rpoB* and *rpoC1* among 86, 94, 102, 104 and 100 orchid species, were 0.247 (range 0-0.498), 0.066 (range 0-0.18), 0.023 (range 0-0.062), 0.04 (range 0-0.103) and 0.03 (range 0-0.09), respectively. Thus, average inter-specific distance was the highest for ITS and the lowest for *rbcL*, as was also obtained in data set I.

6.2.2.2 Species Resolution

The species resolutions for the tested loci were identical with both genetic distance and tree based methods. Out of 86 species analyzed with ITS, five species pairs exhibited distance estimates as zero. Therefore, these ten species could not be discriminated on the basis of ITS sequences. However, *Pholidota imbricata* and *P. pallida* collected as two different species and had formed species pair, on literature survey, were found to be synonyms and thus represented the same species. This provided an example of congruence of conventional taxonomy and DNA barcoding. ITS provided the highest species resolution (90.9%) for orchid species using genetic distance and tree methods.

Of the 94 species analyzed for *matK*, 76 could be successfully discriminated on the basis of K2P distances. Therefore, the species resolution was 80.85%. However, the species resolution was observed to be 55.82%, when *rbcL* sequences of 102 species were analyzed. Other two loci from the chloroplast genome, *rpoB* and *rpoC1* represented by 104 and 100 species, resolved 55.76% and 47% of the analyzed
species, respectively. Again the species resolution provided by \textit{rpoC1} was minimum among the chloroplast loci, while \textit{matK} resolved maximum species.

When BLAST method was used, the species resolution for ITS slightly increased to 91.86\% whereas with \textit{matK}, a significant increase was observed with 91.48\% analyzed orchid species being resolved. The remaining loci exhibited minor increase in species discrimination rates and the percent species resolution for \textit{rbcL}, \textit{rpoB} and \textit{rpoC1} were 58.82\%, 61.53\% and 50\%, respectively.

Multi-locus combinations were tested in 76 orchid species for which the sequences were obtained for all the five loci and the percent species resolution was calculated using genetic distance method. The two-locus combinations of ITS + \textit{matK}, ITS + \textit{rbcL}, ITS + \textit{rpoB}, ITS + \textit{rpoC1}, \textit{matK} + \textit{rbcL}, \textit{matK} + \textit{rpoB}, \textit{matK} + \textit{rpoC1}, \textit{rbcL} + \textit{rpoB}, \textit{rbcL} + \textit{rpoC1} and \textit{rpoB} + \textit{rpoC1} provided the species resolution of 94.74\%, 93.21\%, 92.11\%, 92.11\%, 90.79\%, 89.47\%, 92.11\%, 71.05\%, 65.79\% and 72.37\%, respectively. The three-locus combination - ITS+\textit{matK}+\textit{rbcL} resulted in further increase in species resolution and yielded 96.05\% species discrimination rate. The remaining three-locus combinations exhibited species discrimination rates above 90\% but less than 96.05\% with the exception of \textit{rbcL}+\textit{rpoB}+\textit{rpoC1} showing only 77.63\% discrimination success. The inclusion of one more locus to the above three-locus combination of ITS+\textit{matK}+\textit{rbcL} did not result in further increase of species resolution. The four- and five-locus combinations containing three loci viz., ITS, \textit{matK} and \textit{rbcL} barcodes in combination with either \textit{rpoB} or \textit{rpoC1} or both exhibited 96.05\% maximum species resolution only. The remaining four- and five-locus combinations showed 93-94\% species resolution.

\textbf{6.2.3 Analyses Based on Data Set III}

\textbf{6.2.3.1 Inter-specific Distances}

Average inter-specific distances among the congeneric species of 20 orchid genera analyzed separately were: (i) \textit{Aerides}: 0.072, 0.007, 0.01, 0.006 and 0.002, (ii) \textit{Coelogyne}: 0.054, 0.005, 0.001, 0.002 and 0.003, (iii) \textit{Conchidium}: 0.0395, 0.059, ...
0.028, 0.037 and 0.014, (iv) *Crepidium*: 0.057, 0.008, 0.002, 0.003 and zero, (v) *Cymbidium*: 0.123, 0.02, 0.008, 0.01 and 0.003, (vi) *Eria*: 0.168, 0.04, 0.016, 0.027 and 0.02, (vii) *Habenaria*: 0.175, 0.039, 0.007, 0.014 and 0.008, (viii) *Liparis*: 0.172, 0.024, 0.009, 0.013 and 0.007, (ix) *Nervilia*: 0.214, 0.033, 0.014, 0.034 and 0.039, (x) *Oberonia*: 0.066, 0.018, 0.005, 0.01 and 0.008, (xi) *Paphiopedilum*: 0.031, 0.009, 0.001, 0.001 and 0.001, (xii) *Peristylus*: 0.138, 0.06, 0.012, 0.006 and 0.014, (xiii) *Pleione*: 0.006, 0.001, zero, 0.003 and 0.002, (xiv) *Porpax*: 0.137, 0.019, 0.011, 0.011 and 0.009 and (xv) *Vanda*: 0.021, 0.006, 0.003, 0.003 and 0.001 for ITS, *matK*, *rbcL*, *rpoB* and *rpoC1*, respectively. The average inter-specific K2P distances for four tested loci except ITS for *Eulophia* and *Pholidota* species were 0.008, 0.002, 0.09 and zero; and 0.006, 0.002, 0 and 0, respectively and for *Goodyera* and *Pinalia* for all except *matK* were 0.111, 0.011, 0.017 and 0.012; and 0.087, 0.005, 0.011 and 0.012, respectively. The two species of *Platanthera* showed 0.003 distances for *matK* sequences and zero for the remaining four loci.

### 6.2.3.2 Species Resolution

Among the 73 congeneric species belonging to 20 different orchid genera, ITS and *matK* afforded 100% species resolutions for most of the genera. Among the congeneric species of *Paphiopedilum*, *matK* afforded 100% species resolution while, ITS could resolve only 50% of the sampled species. In contrast, ITS resolved all the analyzed species of *Nervilia* while, *matK* yielded 50% resolution. The remaining loci exhibited variable species discrimination rates.

### 6.3 CONCLUSIONS

The present study revealed ITS to be the best DNA barcode affording highest species resolution in both the assemblages (data set I and II). The first comprised species from the four of the five sub-families Orchidaceae along with non-orchid species from diverse families of plants and the second was represented by only orchids. These results apparently points towards suitability of ITS as the DNA barcodes for the land plants, as has been suggested recently by some other investigators too. However, this locus alone could not be a universal barcode for...
plants because of the several restricting reasons discussed in the thesis. Among the loci tested from the chloroplast genome, matK provided best species resolution values in both species assemblages. Even when applied to congeneric species of twenty orchid genera, these two loci provided best species resolutions. Rather, if taken together, these two loci yielded 100% species resolution. The other much hyped locus, rbcL, seems to be effective at higher taxonomic level, thus pointing towards its utility in multi-level barcodes.

Among the multi-locus combinations tested for 76 orchid species, the highest species resolution was attained by a three-locus combination of ITS+matK+rbcL with 96.05% species resolution. The two-locus combination from the chloroplast genome (matK+rbcL), suggested as universal barcode for land plants by CBOL Plant Working Group, provided 90.79% species discrimination success among 76 species. This is more than the species resolution of 72% obtained by them in a floristic assemblage of species on the basis of which this recommendation was made. However, ITS+matK exhibited still higher species resolution (94.74%) than the previous combination. The present study demonstrated that a three-locus combination (matK+rbcL+ITS), one of the combinations suggested recently by another group as DNA barcode for the land plants, could resolve 96.05% species, but there was no increase in species resolution when the remaining two loci viz., rpoB and rpoC1 from the chloroplast genome were added to the above combination as four- and/or five-locus combinations. These observations indicate the utility of including ITS in the core DNA barcode of matK+rbcL for orchid species.

The results presented in the thesis, adequately address to the concerns about the applicability of DNA barcoding to plants. However, quest for a perfect universal barcode for plants providing 100% species resolution across the plant kingdom appears to be unrealistic, as DNA barcoding, like any other technology, is not expected to be 100% perfect. However, within a taxonomic group 100% species resolution could possibly be obtained by taxa specific barcodes. Thus, the projection that DNA barcodes, once available for all the described species, would be able to provide a correct identity up to species level to any unknown sample, whether available in vegetative, fragmented or DNA form, or would indicate the discovery of a
new species does not hold true. Nevertheless, more than 90% success in species identification with single locus or two-/three-locus combinations emphatically demonstrates the efficacy of the technique. The instances of failure of DNA barcodes to correctly assign the species should encourage taxonomist to re-consider or re-investigate such taxa.