Chapter 5. DISCUSSION

5.1 Status of DNA barcoding of freshwater fishes in India

This study presents a comprehensive assessment of the present status of Indian freshwater fishes that have been recorded and barcoded till date. Survey of different checklists and Fish Base data showed that 890 freshwater fish species have been recorded in India so far, which belongs to 20 orders. Cypriniformes represents highest number of species recorded from India followed by Siluriformes and Perciformes. A total of 21% freshwater fish species in India are endemic, 72% are native and 24 species have been introduced while the statuses of 30 species remain questionable.

The *COI* barcode data have been deposited in both NCBI and BOLD databases. A large number of the deposited *COI* data in NCBI does not contain proper geographical information as compared to BOLD. BOLD shares a tightly integrated data exchange pipeline with NCBI (GenBank) that allows for automatic submission of data to GenBank.

Of the 20 orders of freshwater fishes recorded, *COI* barcodes have been generated for species of 15 orders. Cypriniformes order with highest number of species recorded in India (459) also represents highest number of species barcoded followed by Siluriformes and Perciformes. All the recorded families of Cypriniformes in India contain representative barcodes while Siluriformes and Perciformes approximately contain representative barcodes of 75% of the recorded families. The orders Carcharhiniformes, Myliobatiformes, Pristiformes and Scorpaeniformes do not contain barcodes from any representative species.

5.2 Sequence composition of *COI* barcode and taxonomic rank assignment.

The compositional analysis revealed that there was a tendency towards low G content. The 2nd, 3rd codon position and the complete barcode region showed an AT bias while 1st codon showed a GC bias. Both AT/ GC content and AT/GC skew at 3rd codon position showed strong positive correlation with their respective counterpart calculated for the entire genome. While, values at 1st and 2nd codon positions, showed negative correlation with the overall values. Strand asymmetry of overall barcode region showed a negative AT and GC skew. Nucleotide substitutions were found to be more prominent in the 3rd codon position than in 1st and 2nd codon position. G was least preferred at the third codon position. Most of the preferred codon were seen to have A at the third codon position with few having C at the third codon position.

Among the 1307 sequences belonging to 160 species and three orders (Cypriniformes, Siluriformes and Perciformes) of Indian freshwater fishes, sequence conservation was found to decrease as we moved higher in taxa level for both nucleotide and amino acid sequence. Among the variable sites, information content of each site measured in terms of R_{seq} value increased from the order to species level. Sequence conservation and variable site information of the nucleotide and amino acid together reflected potential of the *COI* gene in categorizing the species to their respective higher taxon.

The percentage of conserved sites increased from the order to species level for both nucleotide and amino acid sequence. Changes in the nucleotide or amino acid sequence that resulted in a variable site can occur in any taxonomic level. For example, a change at a particular position between two species of the same genus would impart changes in the subsequent higher taxonomic orders. Thus, to find out in which hierarchical level the change has actually occurred, the sequence information content for each site was analyzed at the order, family and genus level. At each taxon level the site with R_{seq} value equal to the highest possible value of R_{seq} (2 for nucleotides and 4.32 for amino acids) represented the site conserved at that level. For those that represented lower R_{seq} values, lower taxon level was searched to

locate the level where the saturation took place. Moreover, the R_{seq} value also described the amount of sequence information represented at each site i.e. the number of nucleotides or amino acid occurring at a particular site.

Of the total 510 nucleotide base pairs included in this study, 67% sites were found to be variable between the three orders. Further within the orders, 46% sites were found to be variable in Siluriformes order while 52.5% and 56.07% sites were found to be variable in Perciformes and Cypriniformes respectively. All the three orders had a significant proportion of sites (90%) with R_{seq} value lying in the range of 0.5 – 1.99 and 5% sites in the range of 0 - 0.49. This indicated that at order level, at least 50% of the sites had conserved nucleotide base pairs and among the variable sites 90% sites had possibility of occurrence of two nucleotides. Furthermore, in 74% sites, amino acids were variable between the different orders studied. But within the orders 52% variable amino acid sites were present in Cypriniformes and 49% and 29% sites respectively for Siluriformes and Perciformes. Most of the sites had possibility of occurrence of two amino acids at a given position. Thus, approximately 50% sites were conserved and remaining sites carried character traits to distinguish and identify species into their respective orders. R_{seq} values of nucleotides of the three orders were found to be strongly positively correlated while the R_{seq} values of amino acid were found to be weakly correlated (Table 5.1).

Table 5.1 Correlation of Nucleotide and Amino acid R_{seq} values between different orders of Indian Freshwater fishes.

	Cypriniformes	**Siluriformes**	**Perciformes**
Cypriniformes	**NA**	**0.88**	**0.82**
Siluriformes	**0.2**	**NA**	**0.83**
Perciformes	**0.22**	**0.31**	NA

Cells in grey shades represent nucleotide correlation; cells in yellow shade represent nucleotide correlation.

At family level, percentage of conserved and nearly conserved sites (sites with R_{seq} value 1.5 -1.99) rose to around 70% – 80% and at genus level it further increased to 80% - 99%. Moreover, a major share of the sites had a bias towards fewer numbers of nucleotide. Sites having variable amino acid sequence varied from as low as 3% in Siluridae to as high as 51% in Cyprinidae. However in Cyprinidae, of the 88 variable sites, 79 sites had R_{seq} value close to 4.31 indicating that the variation was mostly due presence of point mutations between species or population. At genus level, the percentage of variable sites further decreased to a lower range (0% – 11%). Thus, at family, approximately 60% – 70% sequence information was carried by nucleotide and amino acid sequence cumulatively. While at genus level sequence information content rose to an approximate value of 80%- 90%.

Thus, our results revealed that 510 bp of nucleotide sequence of *COI* barcode region along with the translated 170 amino acid sequence had the potential to identify species into higher taxonomic ranks. Potential of *COI* gene in higher taxon assignment and phylogeny was observed in some previous studies, which attempted to classify species to higher taxonomic rank (Wilson et al. 2011). Rach et al. (2008) used character based DNA barcoding method to classify 833 odonate specimen into 54 species and 22 genera. In the first officially reported DNA barcoding study, Hebert et al. (2003a) reported 96% success in assigning the studied specimen to kingdom Animalia and consecutively 100% success in class and order assignment using amino acid sequence.

Our results in concordance to previous studies reveal that, a pattern variation exists at each level of taxonomic hierarchy that endows the *COI* gene with inherent potential to discriminate species as well as higher taxa. Moreover, the bias towards binary pattern of variation of both nucleotides and amino acids observed in this study was also reported in other taxonomic groups such as aves (Stoeckle 2012). This suggests that, use of *COI* barcode in higher taxon assignment can be extended to other taxonomic groups as well. Higher taxon assignment with DNA barcodes will thus be possible with novel methods that will explore this information content latent in the gene.

5.3 Species level identification using distance based DNA barcode method.

Species identification through DNA barcoding is based upon the principle that interspecific divergence sufficiently outscores intraspecific divergence and the biological species can be clearly demarcated by a threshold value, which corresponds to the divergence between the nearest neighbors within a group (Hebert et al. 2003a). However, despite extensive application of DNA barcoding throughout the last decade, no universal standard threshold has been defined for interspecies demarcation. A prime reason being that mitochondrial DNA (mtDNA) rates of evolution vary between and within species and between different groups of species resulting in broad overlaps of intra and interspecific distances (Rubinoff et al. 2006). Most of the DNA barcoding studies have rather used a threshold that was specifically estimated for the dataset under study (Bucklin et al. 2011, Hebert and Humble 2011, Lijtmaer et al. 2011). Recent studies have preferred the use of difference between minimum congeneric and maximum conspecific divergence to define the barcoding gap (April et al. 2011, Bhattacharjee et al. 2012) and has found it to be more efficient over the use of mean of intra and interspecific sequence variability (Meier et al. 2008). In this study, 136 species showed cohesive clustering between conspecies in the NJ tree and have been considered as true species. However, few species like *Badis badis, Schizothorax progastus, Channa gachua, Puntius sarana, Macrognathus aral, Puntius chelynoides, Tor malabaricus, Channa striata, Epalzeorhynchos bicolor, Acanthocobitis botia and Mastacembelus armatus* formed subclusters under single node. These straightforward cases exhibit a mean divergence of 0.45% (S.E= 0.2) which is close to such studies elsewhere like 0.73% for North American freshwater fishes (April et al. 2011), 0.6% for Cuban freshwater fishes (Lara et al. 2010) and 0.39% for Australian marine fishes (Ward et al. 2005).

The summary of indistinguishable species and species with unidentified candidate species (UCS) are given in Table 5.2. Besides the straightforward cases that correspond to 82% of the studied species (Group 1), some conspecific sequences exhibited interspecific divergence and vice versa, indicating mismatch of nomenclature and DNA barcode.

Some conspecific sequences (Group 2) have clustered separately in NJ trees and have shown divergence above the threshold value. These groups may comprise of either some erroneously identified species or latent species. This high divergence may be an indicator of unidentified candidate species within named species. Previous DNA barcoding studies in other taxonomic groups have successfully identified cryptic species diversity within single known species (Puckridge et al. 2013, Smith et al. 2006, Winterbottom et al. 2014, Yassin et al. 2008), for example ten undescribed species embedded within a single known species of skipper butterfly was delineated using DNA barcodes (Hebert et al. 2004a). In some problematic cases, the number of representative sequences in the dataset was too few to be interpreted. *Lates calcarifer,* comprising of 94 sequences, formed 2 different clusters with 4 sequences clustering separately. The similarity search result of the 4 sequences using BLASTN have shown 99% (E-value = 0.00) match with *Pampus argenteus* while the remaining 90 sequences showed 100% match with *Lates calcarifer* sequences of other countries. This clearly have shown that the 4 sequences marked as *Lates calcarifer* is mislabeled, while the remaining 90 sequences represent true *Lates calcarifer* species.

Interestingly, *Heteropneustes microps* (type locality: Dambuwa, Sri Lanka) is distinguished from its nominal congeners *Heteropneustes fossilis* only by its caudal and anal fins being confluent (vs. separate) (Günther 1864). It has long been considered as synonym species of *Heteropneustes fossilis* (Ferraris 2007). Pethiyagoda (1991) stated that *Heteropneustes microps* is a result of anomalous fin regeneration in *Heteropneustes fossilis*, injury being one of the possible causes and considered *Heteropneustes microps* a junior synonym of *Heteropneustes fossilis.* However, there are also reports of these two species being sympatric thus reducing chances of hybridization between the two species. In our study, one individual of *Heteropneustes fossilis* has clustered away from rest of the 13 individuals of *Heteropneustes fossilis* with divergence in congeneric range and has clustered with *Heteropneustes microps* with low divergence, which creates the contention that *Heteropneustes microps* is a distinct species and the doubtful sequence of *Heteropneustes fossilis* is mislabeled.

Table 5.2 Number of indistinguishable species and the number of species with UCS (Unidentified Candidate Species represented by lineages that diverge by over 2%).

Order barcoded	Species	Specimen	Number of UCS	Species with overlapping barcodes
Beloniformes	1	4	0	0
Clupeiformes	3	8	0	0
Cypriniformes	82	613	6	5
Cyprinodontiformes	5	19	0	1
Mugiliformes	1	1	0	0
Osteoglossiformes	2	11	0	0
Perciformes	21	202	2	0
Siluriformes	55	509	4	2
Synbranchiformes	4	14	0	1
Tetraodontiformes	1	3	0	0
Total	175	1384	12	9

Sequences of *Barilius bendelisis and Barilius barna* have formed 6 dispersed clusters; the mean conspecific divergence within each cluster being 0.8% and the sequences representative of each cluster maintaining the interspecies threshold. The genus Barilius is comprised of at least 20 indigenous species from India (Froese and Pauly 2013), among them only 5 species have been barcoded so far. Therefore, the sequences of *Barilius bendelisis* and *Barilius barna* altogether may represent 6 species (Figure 5.1) which may either belong to the already described species or may represent latent species.

In Group 3, a total of 15 nominal species have exhibited fairly low interspecies divergence, caused by the inclusion of either some synonymous or misidentified species. There are some established cases of synonymy like, *Tor macrolepis and Tor mosal mahanadicus* as a synonym of *Tor putitora* (Laskar et al., 2013). Moreover, inadequate taxonomic details often result in identifying species located in different geographical location as two different species. One such case of synonymy has been addressed recently where *Mytus horai* was reported as a junior synonym of *Mystus vittatus* (Bhattacharjee et al., 2012). Similarly, *Labeo rajasthanicusis* known only from its type locality, Jaisamand lake in Rajasthan and has never been reported after its initial documentation (Talwar and Jhingran, 1991). In our study, cohesive clustering of all 3 sequences of *Labeo rajasthanicus* with *Labeo dussumieri* suggests that the former might be a junior synonym of the latter. *Channa marulius* with conspecific divergence of 10% has formed 2 distinct clusters, one of which has clustered closely with *Channa striata* with a divergence of 0.15% (S.E = 0.02). Whereas, the divergence between the remaining sequences of *Channa marulius* and *Channa striata* was above the congeneric threshold. Such high intraspecific divergence in these *Channa marulius* have also been reported earlier (Benziger et al. 2011) indicating that a few *Channa striata* specimens have been erroneously identified as *Channa marulius*, while the remaining 7 sequences represented true *Channa marulius* sequence.

Thus confusion regarding 17 of the 32 problematic species has been resolved, thereby leading to the categorization of 87.4% of the studied species as true and valid in parity with the existing checklist.

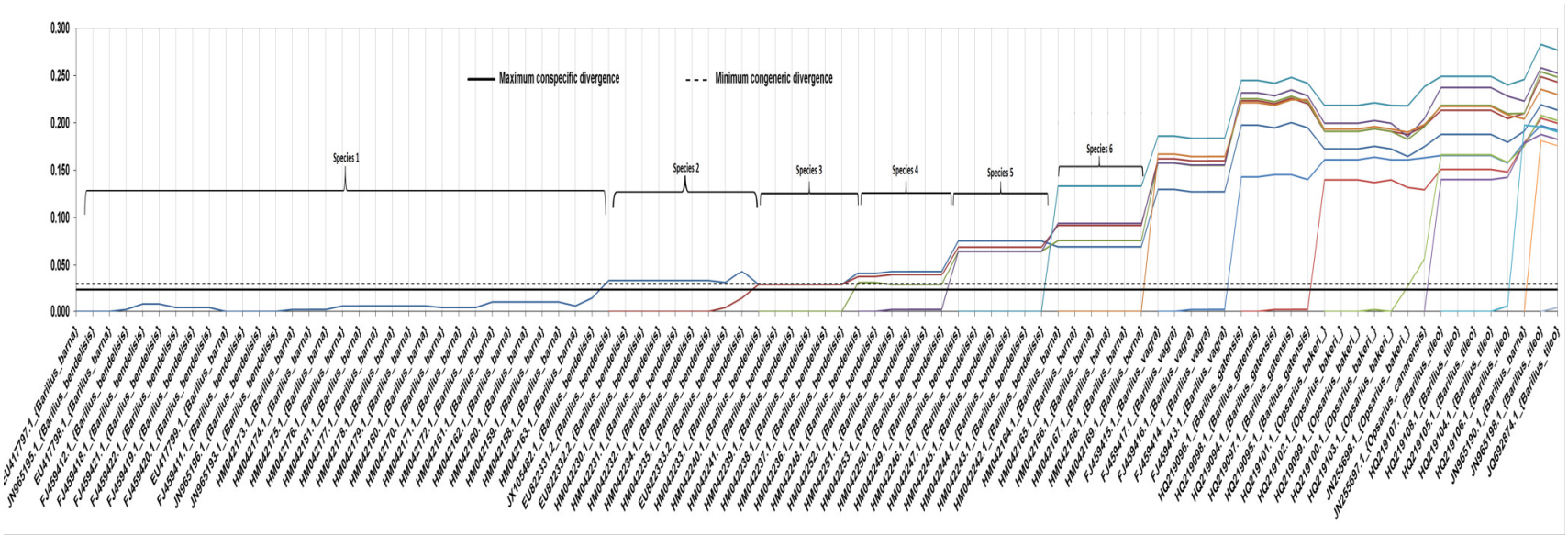

Figure 5.1 Divergence summary (1:1, computed with K2P model) of 90 *COI* barcode sequences of the genus Barilius and Opsarius. The Y-axis represents K2P divergence value while the species are represented along the X-axis. The maximum conspecific divergence (2.14%, black dotted line) and minimum congeneric divergence (2.3%, black solid line) represent the threshold level of conspecific and congeneric divergence respectively. Sequences that represented the species such as *Barilius vagra*, *Barilius gatensis, Barilius tileo* (except sequences of accession number JN965198, JQ692874) *Opsarius bakeri and Opsarius canarensis* showed divergence below maximum conspecific value with their conspecific sequences and above minimum congeneric value with respect to all other congeneric sequences and are considered as true species . Based on the thresholds it is found that conspecific sequences of *Barilius bendelisis and Barilius barna* do not reveal conspecific divergence and altogether represents 6 species (shown as species 1-6 in the figure) and thereby masks some latent species.

5.4 Character based assessment and development of Barcode Motif

A short continuous subset of barcode, rich in informative sites, can act as an effective alternative to the concatenated data matrix of diagnostic characters used in the customary character based methods. Identification of high informative segment within the full-length barcode is important to select a subset of minibarcode. To shrink the barcode, the number of selected features has to be minimized under some constraints while ensuring that the information content of the segment is high enough to delimit species. Furthermore, the length of the segment should be sufficient to generate requisite number of barcodes to describe all the extant species of the studied group. Transversion have been known to, play a dominant role over transitions in the evolution of species (Mitrofanov et al. 2002). Thus, to locate the high informative site we focused on transversion rich sites for generating a short species-specific barcode motif. Further, using the short fragments, a motif based identification system of DNA barcodes was developed.

A stretch of 171bp of transversion rich segment spanning from $BP_{261} – BP_{432}$ has been found to be present in all the three studied orders of fishes (Cypriniformes, Siluriformes and Perciformes). This segment includes the 119bp segment proposed by Bhattacharjee and Ghosh (2014) which successfully delimited species based on their K2P distance. After designing suitable primers, a 154bp nucleotide segment within the transversion hotspot region has been selected as a potential candidate. NJ trees constructed from the full length (513bp) and the minibarcode region (154bp) have shown a comparable clustering pattern of the 160 species. Some sequences that showed deviation from the expected pattern of clustering in the full-length barcode NJ tree, also showed similar trend in the minibarcode tree. These included instances of erroneous identification, mislabeled or misnomer sequences or cases of crypticism, that have been discussed in previous studies (Benziger et al. 2011, Bhattacharjee et al. 2012, Laskar et al. 2013). This further shows that the proposed minibarcode region is competent with the full-length barcode in delimiting species.

In this study, the barcode motifs were developed based on the fundamental concept of character-based system, which states that the members of a given taxonomic

group share attributes that are absent from comparable groups (Sarkar et al. 2002b). The barcode motifs consist of a consensus stretch of nucleotides conserved within a species and the intraspecies variable sites are marked by degenerate nucleotides.

An important parameter in designing species-specific barcode motifs is the number of intraspecies variable sites in each species. Most of the intraspecies variable sites were found to be 2 fold degenerate with few showing 3 fold degeneracy. Moreover, including degenerate sites to develop motif led to the chances that more than one species may share a common nucleotide at a particular position. Assuming that a motif developed for a particular species shares a_1, a_2, a_3 ... a_n common sites with another species. Then the probability of sharing common subsets of the motifs by more than one species is given as (a_1* a_2* a_3……. a_n) /X. This implies, with increase in number of variable sites, probability of getting overlapping barcodes increases exponentially. Thus, a short length barcode segment with high interspecies but low intraspecies variable site is crucial for designing species-specific barcode motifs.

On the other hand, decreasing the total length of barcode significantly, results in reducing the available scanning window of interspecies variation. This may further result in reducing the number of character differences between two closely related species. For instance, two species-specific motifs vary in a single position, with the variable position in one motif being represented by a degenerate nucleotide. If the nucleotide in the variable position of the second species is a subset of the degenerate nucleotide of the first species, then the two motifs may erroneously identify members of both species, resulting in a false positive match. The 154bp segment has a better coverage of the transversion dominant segment than the 119bp segment proposed by Bhattacharjee and Ghosh (2014). Subsequently, the window for interspecies variable sites decreases in the 119bp though no significant change is observed in the number of intraspecies variable sites. Moreover, interspecies variation in the 154bp minibarcode region was higher than that in the full-length barcode. While, intraspecies variation was slightly higher in the full-length barcode as compared to the minibarcode region. Thus, the 154bp length fragment proves to be an ideal choice for designing the species-specific barcode motif.

A crosschecking program, which matches the dataset of all the species with the developed motifs has confirmed that the motifs were unique and matched only with sequences of the concerned species. The efficiency of the motifs in assigning species of different geographic location was checked by comparing the motifs developed from Indian sequences with global *COI* data. For this, Cypriniformes was chosen, as it represents the most diverse order of fishes and is natively distributed on all continents except South America, Australia, and Antarctica (Mayden et al. 2009). Most of the Cypriniformes species (18 out of 24) with sequences from different geographic location were identified correctly using the 154bp barcode motifs. This shows that our algorithm searched for unique, species-specific sequences, but also considered intraspecific variation among haplotypes of each species (where different haplotypes were available). Most of the species that were not identified by their respective motifs had few representative sequences from India, resulting in a poor coverage of the intraspecies diversity in the developed motif.

As observed in some previous studies, the main limitation of the character based identification approach has been the fact that it is based upon a priori knowledge of sequences (Zou et al. 2011). Hence, lack of exact matches, undiscovered haplotypes or geographic variants can lead to failure in annealing to the barcode motifs. This limitation was minimized by reducing the barcode length and including species with adequate number of barcodes. Thus, employing DNA barcode "motifs" to identify previously defined groups of organism will greatly enhance the applicability of DNA barcodes in biodiversity assessment.

5.5 Conclusion

This study reflects the current quantitative and qualitative status of DNA barcoding of Indian freshwater fishes. Our survey has revealed that DNA barcoding of freshwater fish resources of India is far from being comprehensive with only 20% of the recorded species being barcoded until now. With the available DNA barcode data, 88% (approximately) of the species have been identified in parity with existing checklist. Thus, the species level demarcation based on the K2P divergence and NJ based phylogenetic clustering of *COI* sequences is worthy. However, this study has detected some cases of erroneous identification, and the presence of some latent species, which have resulted in an incoherent reference barcode library of freshwater fishes of Indian subcontinent. Therefore, there is a need to revisit the specimens whose barcode data are erroneous and further taxonomic inquiry is recommended for species whose statuses are found to be doubtful. The study will also provide direction to future studies by highlighting the fish groups, which need to be barcoded.

The study also revealed that amount of conserved sequence and information content of variable sequence, together contributes sufficient signal to the *COI* barcode region to assign species to higher taxonomic level. Further, the observed bias towards binary pattern of variation of both nucleotides and amino acids indicates that higher taxon assignment using *COI* will also be possible for other animal groups.

Further, the study shows that the short 154bp fragment of *COI* barcode carries sufficient information in the form of characters that can delimit species. The study has also established that transversion biased region of gene are more effective in developing barcodes for species diagnosis. The use of a continuous stretch of character attributes provides a substantial improvement over the conventional character based system. The short DNA barcodes can be incorporated into a data matrix that contains information in the form of characters and can act as a diagnostic feature for identifying species. This assay is anticipated to be a starting point for developing more sophisticated methods of designing barcode tags. Such tags would pave way for real-time application of DNA barcoding in species identification.