Chapter 4. RESULTS

Overview

This study is a first of its kind attempt to a comprehensive assessment of DNA barcoding of Indian freshwater fishes. Based on the objectives of the study, the work has been divided into 5 chapters as described below.

First, in "**Chapter 4.1**", "Assessment of freshwater fishes of India" all the available sources were surveyed to study progress in the status of both recorded and barcoded freshwater fishes of India.

In "**Chapter 4.2**", "Compositional analysis of COI DNA Barcodes of Indian Freshwater fishes" a detailed compositional analysis of the DNA barcode region was done to understand the pattern of variation within the gene. Further, the amount of sequence information inherent in the *COI* barcode region was calculated to reveal its potential in higher taxonomic rank assignment.

In the third chapter, "**Chapter 4.3**", "DNA barcode based taxonomic rank assignment using distance method" species identification was done based on *COI* DNA barcode through distance based methods. The study elucidated the actual species status of the studied species and helped to flag the species whose statuses have been doubtful.

In "**Chapter 4.4**", "Development of Character profiles for different hierarchical levels of a taxon" character based DNA barcoding was employed for species delimitation and identification to higher taxonomic rank. *COI* barcode sequence based character profiles were developed for different taxonomic ranks of Indian freshwater fishes.

Finally, in "**Chapter 4.5**", "Development of species specific barcode motif" a new method of "reading barcodes" was developed using a continuous short segment within the conventional barcode. This segment was described as the "barcode motif" and its effectiveness in delimiting species of Indian freshwater fishes and global Cypriniformes species was verified.

Chapter 4.1 Assessment of freshwater fishes of India.

4.1.1 Assessment of the recorded status

A compilation from different checklists and Fish Base data revealed the presence of 890 freshwater fish species found in India, which belonged to 20 orders. The order Cypriniformes alone represented 52% of the freshwater fish species recorded from India followed by, Siluriformes and Perciformes which accounted for 23% and 13% species respectively. Remaining 17 orders accounted for only 12% of the total recorded species. Among them Synbranchiformes and Clupeiformes orders consisted of 22 and 14 species respectively while Beloniformes and Cyprinodontiformes each consisted of 11 species. Orders Elopiformes, Gonorynchiformes, Pleuronectiformes, Scorpaeniformes had single representative species.

Among the total numbers of species recorded 187 species were found to be endemic, 646 species were native and 24 species were introduced (Table 4.1). Thus, a total of 21% freshwater fish species were endemic that contained 131 species of Cypriniformes, 35 species of Siluriformes and 7 species each from Perciformes and Synbranchiformes. The orders Beloniformes, Clupeiformes, Tetraodontiformes each included 2 endemic species. 72% of the freshwater species found in India were native with Cypriniformes, Siluriformes and Perciformes contributing major share of 308, 161 and 99 species respectively. Baring Salmoniformes all the orders of freshwater fishes found in India had at least one native species. The order Salmoniformes was introduced in India and contained 3 representative species (*Oncorhynchus mykiss, Salmo trutta, and Salvelinus fontinalis*). Nine species from Cypriniformes order, 5 species from Cyprinodontiformes, 4 species from Perciformes and 3 species from Siluriformes were introduced. The status of remaining 33 species remained questionable with 3 of them recognized as cases of misidentification (*Channa micropeltes, Clarias batrachus, and Euchiloglanis kishinouyei*). The details of all the species along with their occurrence status are provided in Appendix 1.

Table 4.1 Summary of occurrence status of species belonging to different orders of Indian freshwater fishes

Order	Endemic	Introduced	Misidentification	Native	Questionable	Total
Anguilliformes	0	0	0	6	0	6
Beloniformes	2	0	0	9	0	11
Carcharhiniformes	0	0	0	4	0	4
Clupeiformes	2	0	0	12	0	14
Cypriniformes	131	9	0	308	11	459
Cyprinodontiformes	1	5	0	4	1	11
Elopiformes	0	0	0	1	0	1
Gonorynchiformes	0	0	0	1	0	1
Mugiliformes	0	0	0	6	1	7
Myliobatiformes	0	0	0	6	1	7
Osteoglossiformes	0	0	0	2	0	2
Perciformes	7	4	1	99	6	117
Pleuronectiformes	0	0	0	1	0	1
Pristiformes	0	0	0	2	1	3
Salmoniformes	0	3	0	0	0	3
Scorpaeniformes	0	0	0	1	0	1
Siluriformes	35	3	2	161	6	207
Synbranchiformes	7	0	0	13	2	22
Syngnathiformes	0	0	0	7	1	8
Tetraodontiformes	2	0	0	3	0	5
Total	187	24	3	646	30	890

4.1.2 Assessment of the barcoded status

For each recorded species of Indian freshwater fishes both NCBI and BOLD were surveyed for the presence of representative *COI* barcode sequences. The detailed account of recorded and their corresponding barcoded status is presented in Appendix 1. Of the 20 orders of freshwater fishes recorded, *COI* DNA barcodes have been generated for species of 15 orders (Table 4.2, Figure 4.1). Cypriniformes order with highest number of species recorded in India (459) also represents highest number of species barcoded (114 i.e. 25% of the recorded species) followed by Siluriformes (88 of 207) and Perciformes (49 of 117) (Figure 4.1 c). The orders Elopiformes, Gonorynchiformes, Pleuronectiformes, Scorpaeniformes each has one recorded native species for which, the *COI* barcode data has been deposited. Similarly, all the species recorded for Mugiliformes and Salmoniformes have been barcoded. However, no species have been barcoded until date from the orders Carcharhiniformes, Myliobatiformes, Pristiformes and Scorpaeniformes.

All the 5 recorded families (Balitoridae, Cobitidae, Cyprinidae, Nemacheilidae, and Psilorhynchidae) of the order Cypriniformes, have *COI* barcodes of representative species. However, only 61% of the recorded genus and 25% of the recorded species have been barcoded. Cobitidae and Cyprinidae are among the most well represented families with barcodes from about 33% of the recorded species. Of the 14 recorded species of Balitoridae, 2 species have been barcoded and of the 9 recorded species of Psilorhynchidae only one has been barcoded. However, Nemacheilidae is the most poorly barcoded order with only 2 (*Acanthocobitis botia, Schistura beavani*) of the 103 recorded species being barcoded.

Out of the 15 recorded families of Siluriformes order in India, 12 families have been barcoded (Amblycipitidae, Ariidae, Bagridae, Clariidae, Erethistidae, Heteropneustidae, Olyridae, Pangasiidae, Plotosidae, Schilbeidae, Siluridae, and Sisoridae). Species from remaining 3 families Akysidae, Chacidae, Loricariidae have not been barcoded from India. *Akysis manipurensis* is the only recorded species of Akysidae family from India but it has not yet been barcoded. Loricariidae has one native species *Mystus seengtee* and two introduced species *Pterygoplichthys anisitsi* and *Pterygoplichthys multiradiatus*, none of which has been barcoded. Of the 54 recorded genus of Indian Siluriformes order 62% (34) have barcodes from representative species. All the recorded species of the genus Ailia, Arius, Bagarius, Erethistes, Eutropiichthys, Gogangra, Neotropius, Osteogeneiosus, Pangasius, Plotosus, Pseudeutropius and Wallago have been barcoded.

In the order Perciformes, 15 of the 20 families have been barcoded. Species from the families, Blenniidae, Datnioididae, Kurtidae, Leiognathidae, and Sciaenidae have not yet been barcoded. All the recorded species of the families, Anabantidae, Cichlidae, Kuhliidae, Scatophagidae, Terapontidae, and Toxotidae have been barcoded. 36% of the genera (23 out of 63) have been barcoded so far. All species from the genera Anabas, Chanda, Ctenops, Eleutheronema, Etroplus, Giuris, Glossogobius, Kuhlia, Ophiocara, Oreochromis, Polynemus, Pseudosphromenus, Scatophagus, Terapon, Toxotes have representative barcodes.

In the order, Synbranchiformes 6 out of 23 known species have been barcoded. Among the barcoded species, *Macrognathus aculeatus, Macrognathus aral, Macrognathus pancalus, Mastacembelus armatus* belong to Mastacembelidae family and the remaining two species (*Monopterus albus, Monopterus cuchia*) belong to the Synbranchidae family. About 55% of the recorded species have been barcoded in the order Cyprinodontiformes (6 out of 11). Among them *Aplocheilus panchax* belongs to the family Aplocheilidae, while the remaining species, *Gambusia affinis, Gambusia holbrooki, Poecilia reticulata, Xiphophorus hellerii, Xiphophorus maculatus* belongs to the Poeciliidae family. In the order Beloniformes, 6 of the 11 recorded species have been barcoded (*Oryzias dancena, Oryzias melastigma, Xenentodon cancila, Hyporhamphus limbatus, and Hyporhamphus xanthopterus*). Thus baring Zenarchopteridae, all the other 3 families (Adrianichthyidae, Belonidae, and Hemiramphidae) of the 4 recorded families of Beloniformes have been barcoded.

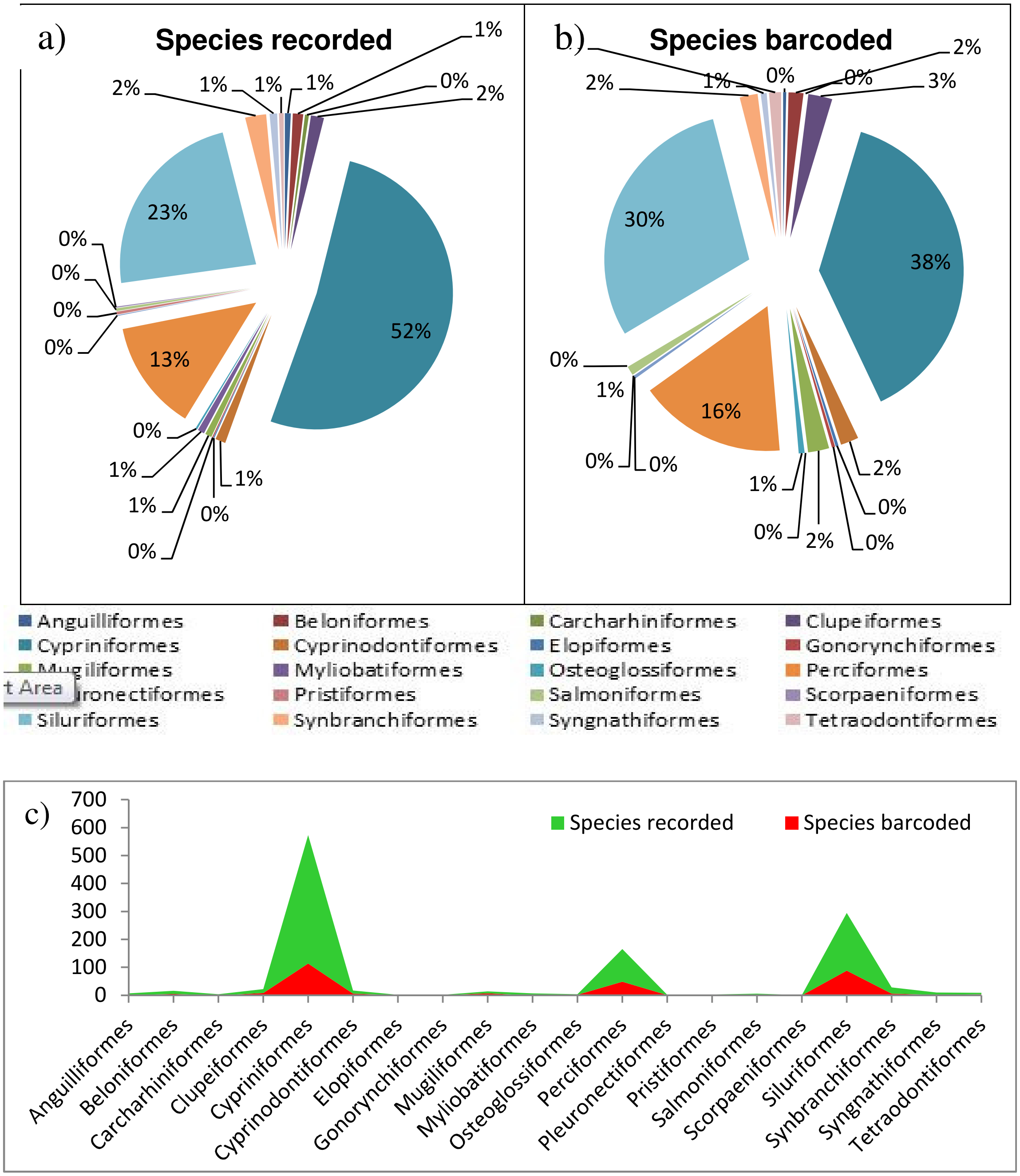

Figure 4.1 Diagram showing a) species recorded b) species barcoded c) species recorded vs species barcoded in 20 recorded orders of Indian freshwater fishes.

Different orders are represented by different colors in (a) and (b) as described in the figure legend. Percentage of species recorded and barcoded from each order out of the total species recorded and barcoded are indicated in the respective pie diagram (a) and (b). While in (c) a comparison between number of species barcoded and number of species recorded in each recorded orders is shown.

Table 4.2 Species recorded versus Species barcoded in different orders of Indian freshwater fishes.

#No	Orders	No of species recorded	No of species barcoded
1	Anguilliformes	6	1
2	Beloniformes	11	5
3	Carcharhiniformes	4	0
4	Clupeiformes	14	8
5	Cypriniformes	459	114
6	Cyprinodontiformes	11	6
7	Elopiformes	1	1
8	Gonorynchiformes	1	1
9	Mugiliformes	7	7
10	Myliobatiformes	7	0
11	Osteoglossiformes	2	2
12	Perciformes	117	49
13	Pleuronectiformes	1	1
14	Pristiformes	3	0
15	Salmoniformes	3	3
16	Scorpaeniformes	1	0
17	Siluriformes	207	88
18	Synbranchiformes	22	6
19	Syngnathiformes	8	2
20	Tetraodontiformes	5	4

4.1.3 Species authentication using DNA barcoding

Establishing food authentication methods is an important task for fisheries research laboratories and food control authorities. Local name of species can often be confusing as a single name may indicate more than one authentic species on the other hand two or more than one name may stand for a single species. One of the most important applications of DNA barcoding is its ability to identify species from specimen of any life stage and from fragmented specimen where descriptive morphological features are not complete or absent. Thus, in this study DNA Barcoding was used for identification of such samples.

Here, 15 freshwater fish specimens were collected from market. The specimens belonged to two main locations Barak river system and Hooghly river system. The samples were purchased from market based on their local names. Genomic DNA was extracted from the tissue of the specimen. The target barcode region of *COI* gene was amplified and sequenced. Figure 4.2 a, b shows a representative image of the process. Figure 4.2 (a) shows the gel image of PCR amplicons while 4.2 (b) shows a chromatogram of the raw sequence data. The barcode sequences were checked against the reference barcode database to find their species identity (Table 4.3). Taxonomic identification was resolved for the specimens for whom morphological features were intact and the *COI* barcode sequences were deposited in GenBank (Table 4.4).

4.1.3.1 Detecting erroneous identification of fillet.

Identification of fish in the form of fillet or in cut form is most difficult task as most of the key morphological features are lost in the process. However, closely related species may bear significant difference in economic value. Therefore proper labeling of fishes are important from economic and conservation point of view. Here, DNA barcoding method was employed to identify some of common fishes available in the market and popularly consumed by the Bengali community of eastern India. Since morphological features of the specimen were not found to be intact, we had to depend solely on BLAST and BOLD search for species identification of these specimens.

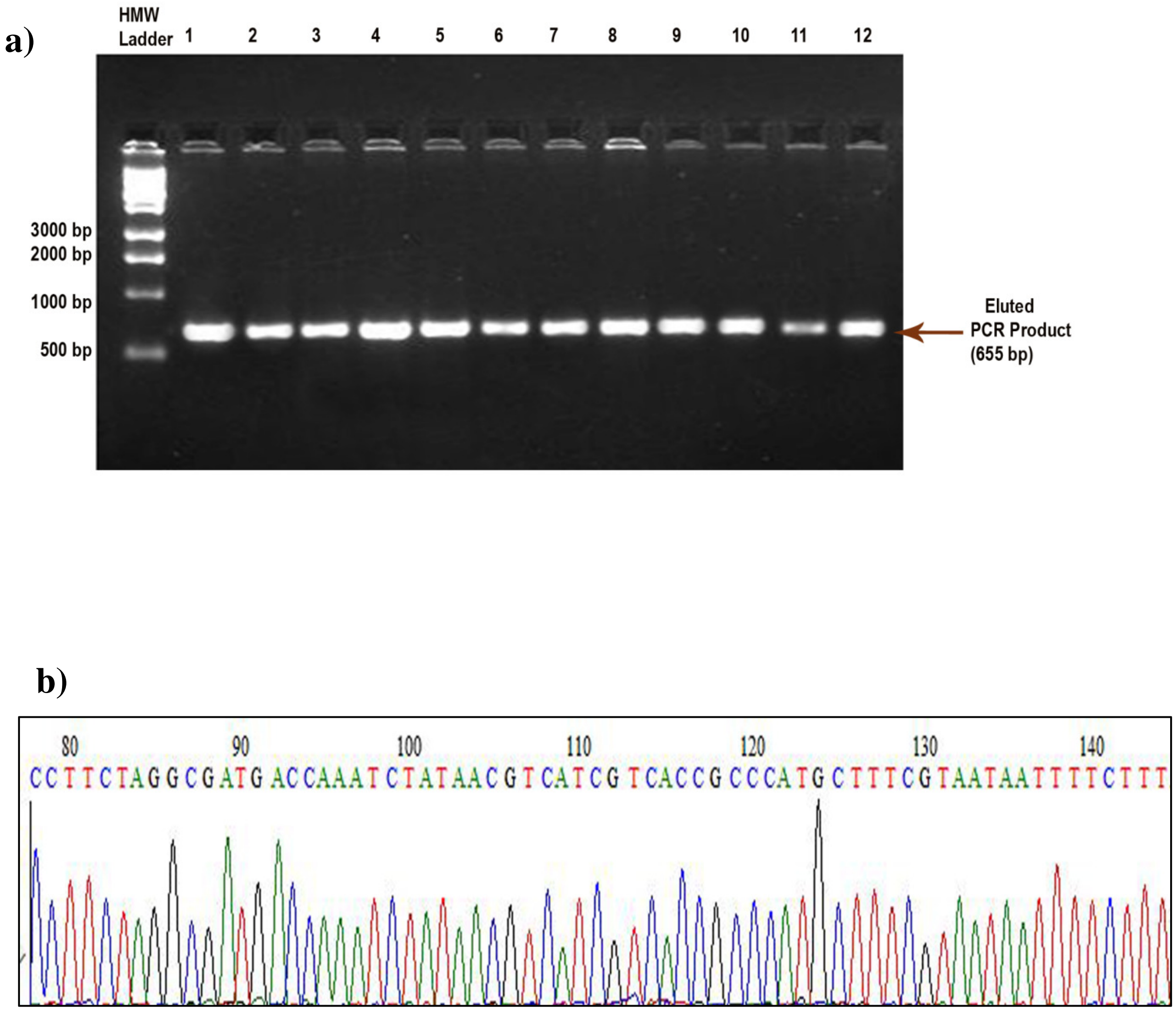

Figure 4.2 Representative image of DNA sequencing process showing:

a) **Gel image of PCR amplicons of *COI* barcode region of fish samples.**

b) **DNA sequencing Chromatogram data.**

Specimen with the code SGMCAC-MOF2 was purchased under the name Gagala. The BLAST search showed 99% similarity (100% query coverage and 0.0 E-value) with both *Arius maculatus*. The BOLD search further showed similarity with *Arius maculatus* ranging from 99.36 – 99.53%. The specimen was thus identified as *Arius maculatus.*

Specimen with the code SGMCAC-MOF3 was purchased under the name Tengara. The BLAST search showed 100-99% identity (99.5- 98% query coverage and 0.0 E-value) with *Mystus vittatus.* BOLD identification search result was unable to make a species level match and suggested that the queried specimen is likely to be one among *Mystus vittatus, Mystus horai and Mystus montanus.* In BOLD search, the query sequence showed 99.5 - 100% similarity with six *Mystus vittatus* sequences and 99.83 % similarity with *Mystus horai*. Another specimen SGMCAC-MOF9 was purchased under the same name of Tengara. BLAST result showed 99% similarity with *Mystus gulio*. BOLD identification search result clearly stated that the submitted sequence matched to *Mystus gulio*. It further mentioned this identification is solid unless there is a very closely allied congeneric species that has not yet been analyzed and such cases are rare. This therefore revealed that the two specimens purchased under same common name from market actually revealed two different established species. Further literature search revealed that *Batasio tengana*, *Mystus bleekeri* along with *Mystus vittatus* all have the common name tengra. *Mystus gulio* however is known as nuna-tengra in Bangladesh.

Specimen SGMCAC-MOF5 was purchased as Magur. Similarity search result with both BOLD and BLAST revealed 100% similarity with *Clarias batrachus*. Specimen SGMCAC-MOF6 was purchased as Kajori and its *COI* sequence showed 100% similarity with *Ailia coila* in both BLAST and BOLD identification search. In BLAST search two sequences of *Ailia. coila* with accession number FJ459444 and FJ459442 showed 100% similarity with the query sequence while 14 other sequences showed 99% similarity.

Table 4.3 Similarity search table of specimen barcoded in the study using BLAST search and BOLD identification engine.

Sample ID	Common name	BLAST			BOLD	
		Species	Query coverage	Identity	Species	Identity
SGMCAC-MOF2	Gagala	*Arius maculatus*	100%	99%	*Arius maculatus*	99.53
SGMCAC-MOF3	Tengra	*Mystus vittatus*	98%	99%	*Mystus vittatus*	100%
					Mystus horai	100%
SGMCAC-MOF5	Magur	*Clarias batrachus*	100%	100%	*Clarias*	100%
SGMCAC-MOF6	Kajori	*Ailia coila*	98%	100%	*Ailia coila*	100%
SGMCAC-MOF7	Kam magur	*Plotosus canius*	99%	96%	*Plotosus nkunga*	99%
		Plotosus nkunga	97%	88%		
SGMCAC-MOF8	Pabda	*Ompok*	100%	98%	*Ompok*	100%
SGMCAC-MOF9	Tengra	*Mystus gulio*	100%	99%	*Mystus gulio*	100%
SGMC-MOF18	Mrigal	*Cirrhinus*	100%	100%	*Cirrhinus*	100%
SGMC-MOF28	Bata	*Labeo bata*	99%	100%	*Labeo cf. boga*	100%
SGMC-MOF29	Bata	*Labeo bata*	99%	100%	*Labeo cf. boga*	100%
SGMC-MOF30	Bata	*Labeo bata*	99%	100%	*Labeo cf. boga*	100%
SGMC-MOF33	Bata	*Labeo bata*	99%	100%	*Labeo cf. boga*	100%
SGMC-MOF34	Kharish	*Labeo bata*	99%	100%	*Labeo cf. boga*	100%
SGMC-MOF35	Kharish	*Labeo bata*	99%	100%	*Labeo cf. boga*	100%

Table 4.4 Details of specimen sequenced as submitted in GenBank.

Sample ID	Order	Family	Species	Accession Number	Lat_Lon
SGMCAC-MOF2	Siluriformes	Ariidae	*Arius maculatus*	KJ959637	22.54N 88.31E
SGMCAC-MOF3	Siluriformes	Bagridae	*Mystus vittatus*	KJ959638	22.54N 88.31E
SGMCAC-MOF5	Siluriformes	Clariidae	*Clarias batrachus*	KJ959639	22.54N 88.31E
SGMCAC-MOF6	Siluriformes	Schilbeidae	*Ailia coila*	KJ959640	22.54N 88.31E
SGMCAC-MOF7	Siluriformes	Plotosidae	*Plotosus sp*	KJ959641	22.54N 88.31E
SGMCAC-MOF8	Siluriformes	Siluridae	*Ompok bimaculatus*	KJ959642	22.54N 88.31E
SGMCAC-MOF9	Siluriformes	Bagridae	*Mystus gulio*	KJ959643	22.54N 88.31E
SGMC-MOF18	Cypriniformes	Cyprinidae	*Cirrhinus sp*	KJ959644	24.76N 92.83E
SGMC-MOF28	Cypriniformes	Cyprinidae	*Labeo bata*	KJ959645	24.76N 92.83E
SGMC-MOF29	Cypriniformes	Cyprinidae	*Labeo bata*	KJ959646	24.76N 92.83E
SGMC-MOF30	Cypriniformes	Cyprinidae	*Labeo bata*	KJ959647	24.76N 92.83E
SGMC-MOF33	Cypriniformes	Cyprinidae	*Labeo bata*	KJ959648	24.76N 92.83E
SGMC-MOF34	Cypriniformes	Cyprinidae	*Labeo bata*	KJ959649	24.76N 92.83E
SGMC-MOF35	Cypriniformes	Cyprinidae	*Labeo bata*	KJ959650	24.76N 92.83E

The specimen SGMCAC-MOF7 was purchased under the common name Kam magur. Similarity search using BOLD, confirmed species status as *Plotosus nkunga*. The query sequence showed 99% similarity with two database sequence of *Plotosus nkunga* and 87% similarity with 4 other sequences of *Plotosus nkunga*. However BLAST similarity search revealed a slightly different result. The query sequence showed 96% similarity with 3 sequences of *Plotosus canius* over 99% query coverage. Further, it showed only 88% similarity with two database sequence of *Plotosus nkunga* over 97% coverage of query length. Thus, overall the species status of the specimen could not be confirmed while classification up to genus was established as Plotosus.

Specimen SGMCAC-MOF8 purchased as pabda, matched to *Ompok bimaculatus* in BOLD identification search with 97-100% similarity with database sequences. Similarly in BLAST search 97-98% similarity with *Ompok bimaculatus* reference sequences. Specimen SGMC-MOF18 was bought under the common name Mrigal. A species level match could not be made with BOLD identification search, and it was suggested that the species could be either *Cirrhinus mrigala or Cirrhinus cirrhosis*. However, in BLAST search the query sequence showed 100% identity over 100% query coverage with 3 database sequences of *Cirrhinus cirrhosus* and 100% identity over 99% query length with *Cirrhinus mrigala.* Thus, considering the two search engines, the species status of the specimen could not be confirmed while genus was confirmed to be Cirrhinus.

4.1.3.2 Detecting erroneous identification of juvenile market specimen.

Specimens SGMC-MOF28, SGMC-MOF29, SGMC-MOF30, SGMC-MOF33 were purchased as juvenile Bata fish while the specimen SGMC-MOF34 and SGMC-MOF35 were purchased as juvenile Kharish. All the specimens were purchased from the same trader; the first four specimens were purchased as whole fish with morphological features intact while the last two were purchased in the form of fillet. However all the sequences showed similar BLAST and BOLD similarity search results. In BLAST search all the sequences showed 100% similarity with database sequence of *Labeo bata* while in BOLD, species status was not confirmed though the query sequences showed 100% similarity with two sequence of *Labeo boga*. Further, in BOLD identification engine, the query sequences showed 99.84-100% similarity with 18 sequences of *Labeo bata*. The morphometric study of these specimens showed high similarity with described features of *Labeo bata* and significant distinction from *Labeo boga*. The specimen possessed a pair of small maxillary barbells hidden inside the labial fold and no cartilaginous support to the lips. The dorsal originated midway between the snout tip and the anterior base of anal. Pelvic originated slightly nearer to the snout tip than to the caudal base (Figure 4.3).

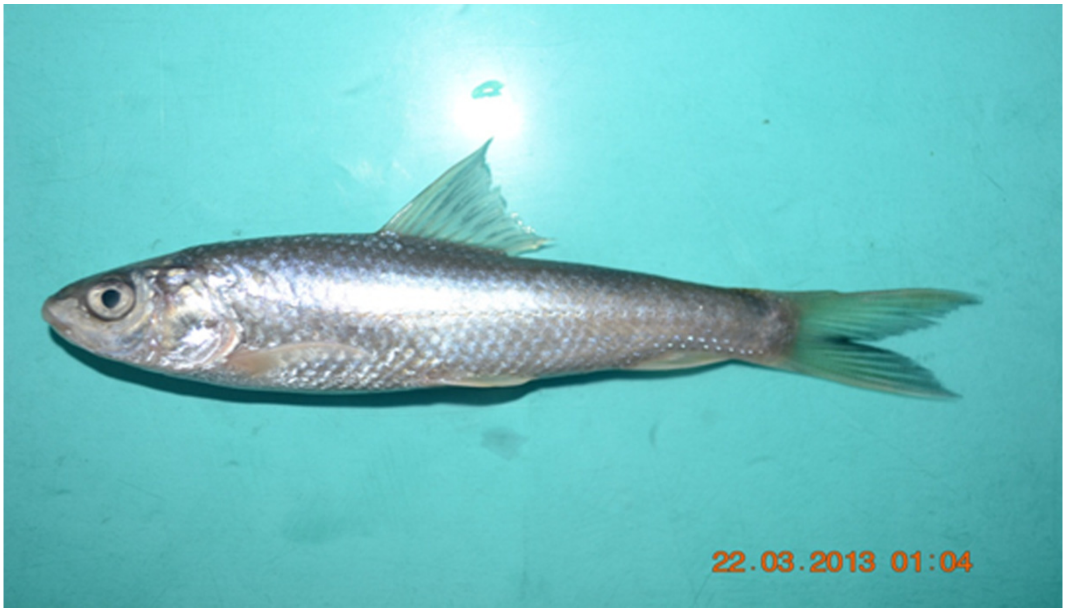

Figure 4.3 Representative image of collected specimen of *Labeo bata*.

Chapter 4.2 Compositional analysis of *COI* DNA barcodes.

4.2.1 Nucleotide composition analysis

Nucleotide composition at all codon position of 10 orders of Indian Freshwater fishes were analyzed as shown in Table 4.5. The base frequencies for each sequence and for total barcode length were calculated by MEGA 5.1. The analysis revealed that nucleotide composition at third codon position showed significant variation while for first and second codon position nucleotide composition was almost uniform across different orders (Figure 4.4).

The nucleotide composition of the 10 orders studied here showed similar pattern across full-length barcode and across 1st and 2nd codon position. At 3rd codon position, all the orders showed variation in all the four nucleotide position. There was a tendency towards low G content and at the 3rd codon position the G content was significantly low for all the orders. The total GC content was lower than AT content. At 2nd and 3rd codon position, there was a bias towards AT over GC, however at 1st codon position GC content was dominant over AT (Figure 4.5).

Comparing the correlation of average GC content with GC content at each codon position, in 10 orders of Indian freshwater fishes, GC content at the 3rd codon position was strongly positively correlated (r = 0.986) to the overall GC content of the barcode region (Figure 4.6). At second codon position, the GC content was weakly negatively correlated (r = -0.067) to overall GC content and at first codon position, it was positively correlated (r = 0.315) to the overall GC content. To obtain a closer perception of GC content distribution, the correlation between GC content at each codon position and average GC content in species of different orders of Indian freshwater fishes was plotted as shown in Figure 4.7.

Table 4.5 Nucleotide compositional analysis of 10 orders of Indian freshwater fishes.

Order	T(U)	C	A	G	T-1	C-1	A-1	G-1	T-2	C-2	A-2	G-2	T-3	C-3	A-3	G-3
Siluriformes	29.1	27.3	27	16.6	17.6	24.2	28.5	29.7	42.9	30.1	14.7	12.3	26.7	27.5	37.9	7.9
Cypriniformes	29.1	26.8	26.8	16.8	18.7	24.1	27	30.2	42.4	30.5	14.7	12.4	27.7	25.7	38.7	7.8
Perciformes	29.1	30	23.8	16.7	18.4	23.3	29.4	28.9	43.5	29.9	14.5	12.2	26.2	37	27.6	9.1
Synbranchiformes	29.1	27	27.6	15	18.5	24.2	29.8	27.4	43	30.5	14.6	11.9	29.6	26.3	38.4	5.6
Beloniformes	29.1	28.7	24.2	16.2	17.8	24.4	28.9	28.9	42.6	30.4	14.7	12.3	32.1	31.4	29.1	7.4
Clupeiformes	29.1	26.2	26.4	17.5	18.6	23.8	27.1	30.5	43.2	29.6	14.8	12.4	27.8	25.1	37.5	9.6
Osteoglossiformes	29.1	25.7	28.4	15.8	19.1	22.7	28.3	29.9	43	29.8	14.9	12.4	28.3	24.4	42	5.3
Cyprinodontiformes	29.1	28.1	25.1	15.6	17.1	25	29.4	28.5	42.4	30.6	14.7	12.4	34.1	28.7	31.2	6
Tetraodontiformes	29.1	31.6	23.6	17.7	17.6	24.1	28.2	30	42.9	30	14.7	12.4	20.7	40.7	27.8	10.8
Mugiliformes	29.1	26.7	23.9	16.9	18.8	23.5	28.8	28.8	42.9	30	14.7	12.4	35.9	26.5	28.2	9.4

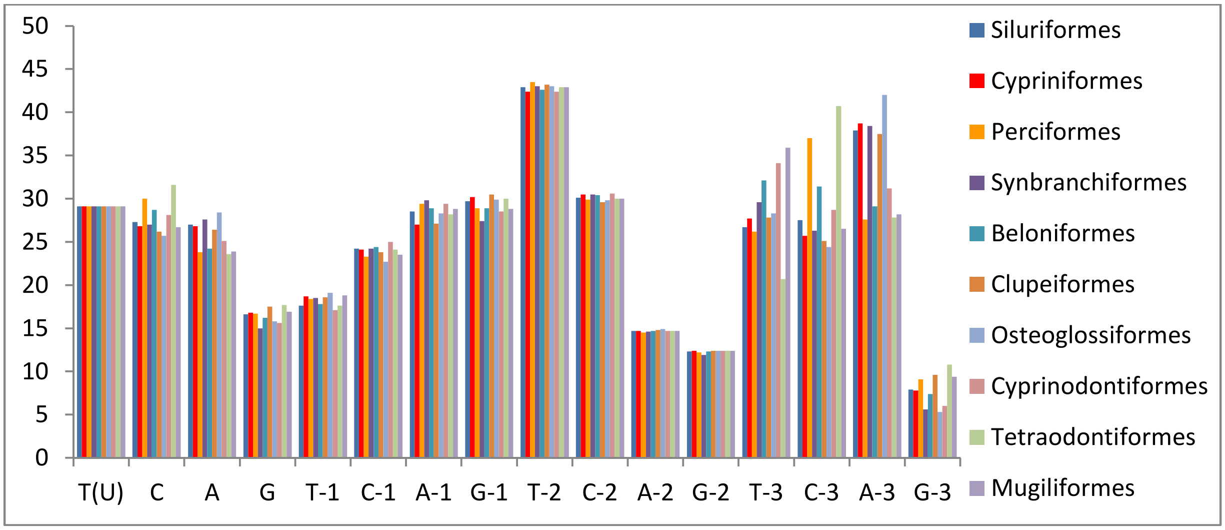

Figure 4.4 Nucleotide compositions in 10 orders of Indian freshwater fishes. Different colored bars represent different orders as given in the figure legend.

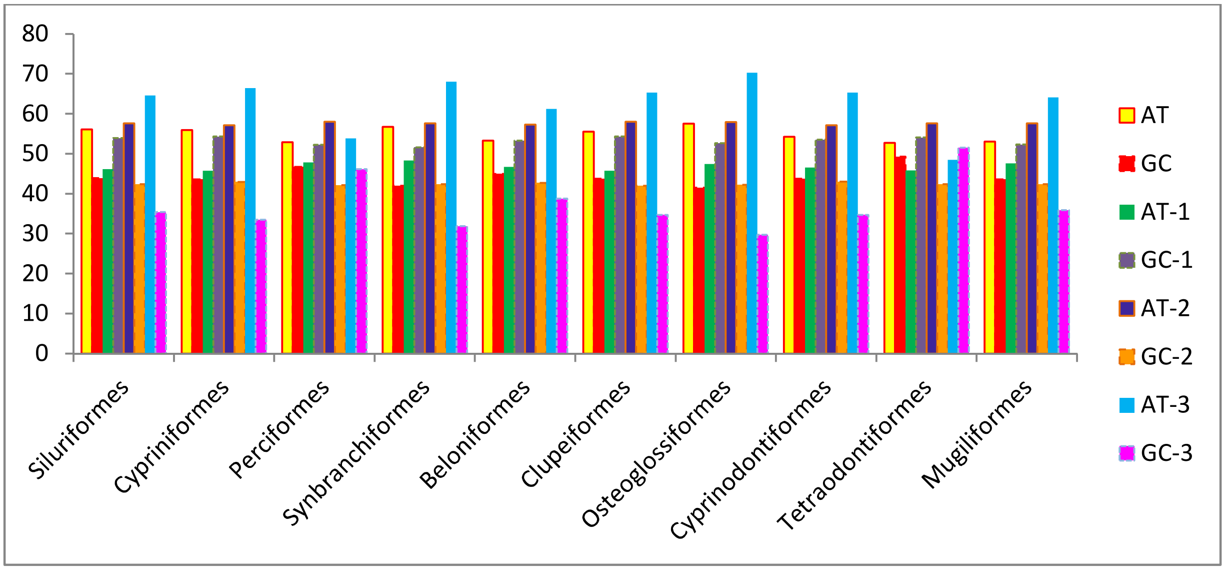

Figure 4.5 AT-GC Bias in 10 orders of Indian freshwater fishes.

Different colored bars represent AT, GC composition at different codon positions in each order as given in the figure legend.

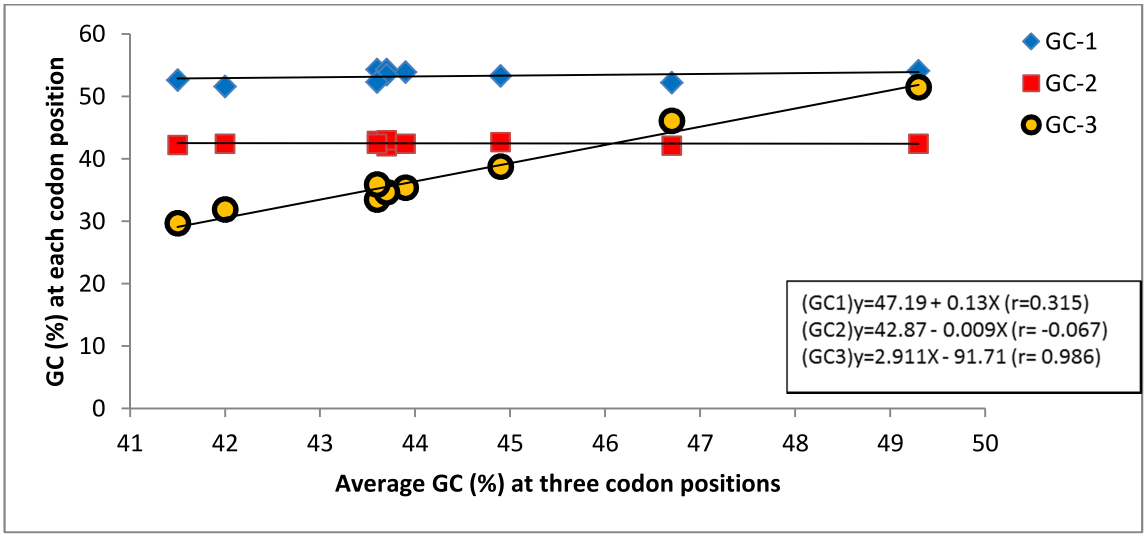

Figure 4.6 Correlation of average GC content with GC content at each codon position, in 10 orders of Indian freshwater fishes.

Each point represents correlation of average GC content in an order at a given codon position with average GC content. Each codon position is marked with a different marker as given in the figure legend.

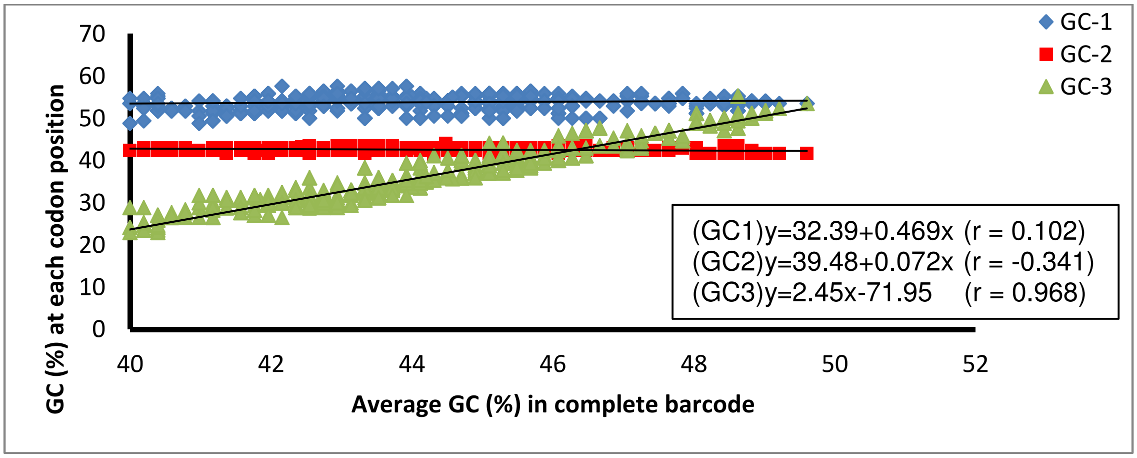

Figure 4.7 Correlation between GC content at each codon position and average GC content in 1307 species of the orders Cypriniformes, Siluriformes and Perciformes.

Each point represents correlation of GC content in a species at a given codon position with average GC content. Each codon position is marked with a different marker as given in the figure legend.

A similar trend was observed in Figure 4.7, with a strong positive correlation (r = 0.968) between third codon and average GC content. Further, the second codon position showed negative correlation (r = -0.368) while first codon position showed positive correlation (r = 0.102). The results showed the expected pattern of high variation at the third codon position and lowest variation at the second codon position.

In addition to variations in GC content, mitochondrial genomes also vary in their patterns of strand asymmetry (usually measured as GC skew and AT skew). Figure 4.8 shows the plot of AT and GC skew for different orders of Indian freshwater fishes at each codon position in the barcode and for the total barcode region. Strand asymmetry in the total barcode region showed a different pattern than in each codon position. Complete barcode region of all the studied species (Figure 4.8 a) showed a negative GC skew. Further, most of the sequences also showed a negative AT skew with few exceptions. At first codon position, (Figure 4.8 b) both AT and GC skew showed positive values while at second codon position, (Figure 4.8 c) AT and GC skew showed negative values. Third codon position (Figure 4.8 d), showed a positive AT skew and negative GC skew.

GC and AT skew at the third codon position showed strong positive correlation (r^{GC3} = 0.88 and r^{AT3} = 0.97) with average GC and AT skew of the complete barcode (Figure 4.9, Figure 4.10). However, both AT and GC skews at the first and second codon position showed weak correlation to the overall trends (r^{GC1} = 0.43 and r^{AT1} = 0.22, r^{GC2} = -0.08 and r^{AT2} = 0.34). At 1st codon position both AT and GC skew was distributed in the positive coordinate. At 2nd codon position both AT and GC skew was distributed in the negative coordinate. Further, at 2nd codon position, GC skew was found to be negatively correlated with overall GC skew while AT skew was weakly positively correlated with overall AT skew.

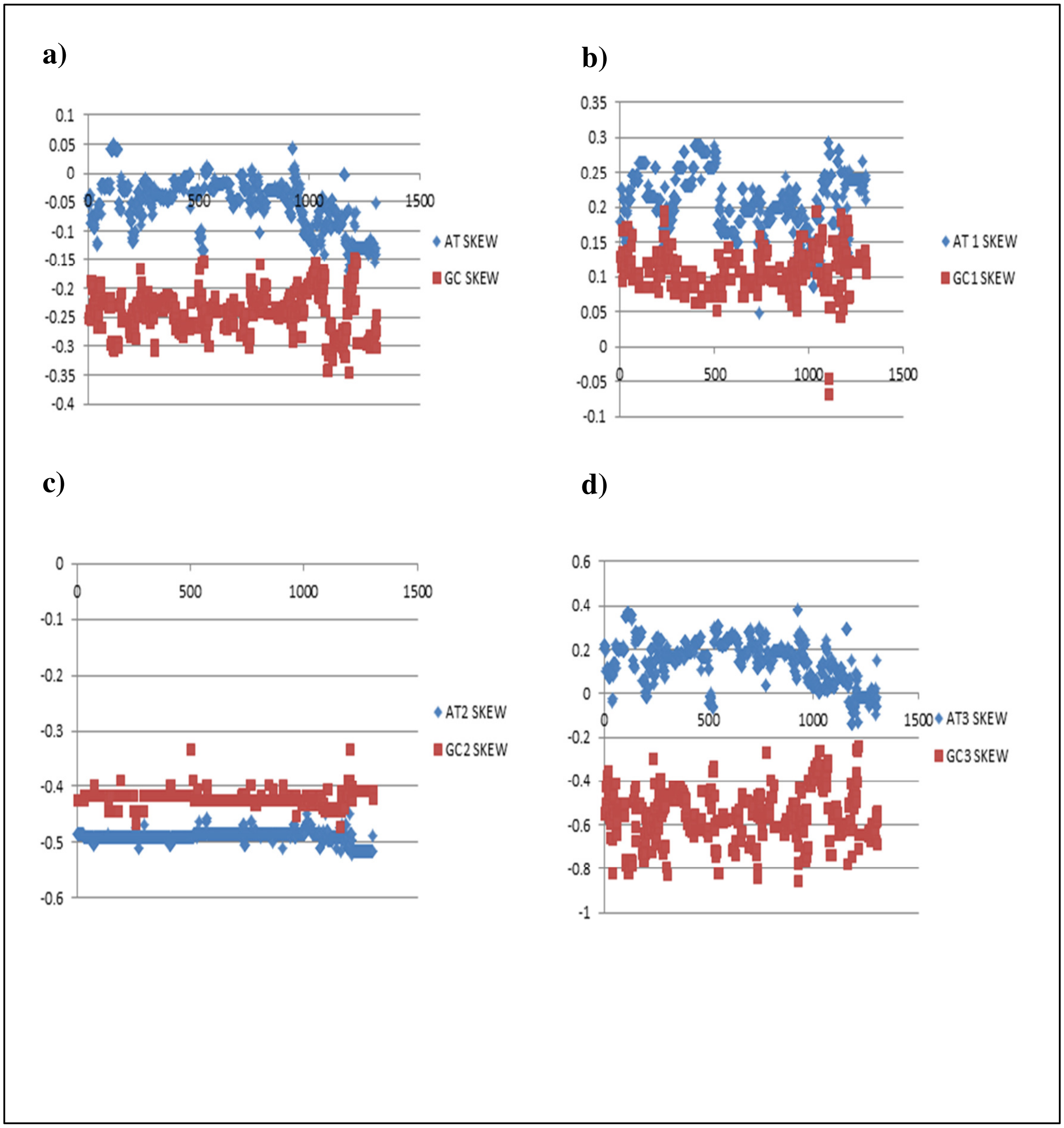

Figure 4.8 Plot of AT-GC SKEW of 1307 species of Indian freshwater fishes a) Overall b) 1st codon position c) 2nd codon position d) 3rd codon position.

AT skews at different codon position are represented by blue diamonds while GC skews are represented by red squares.

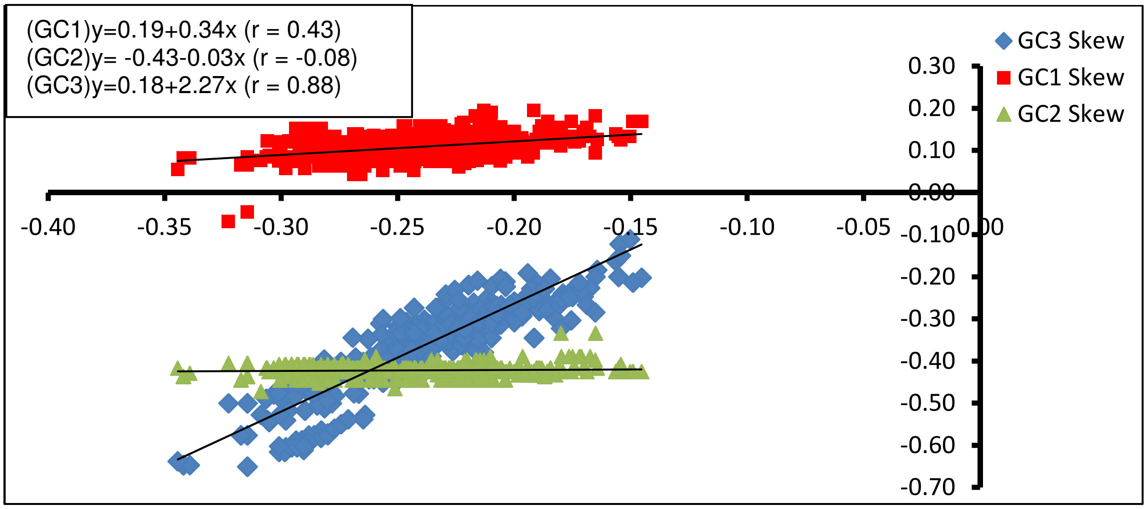

Figure 4.9 Correlation between GC skew at each codon position and average GC skew in species of different orders of Indian freshwater fishes.

Each point represents correlation between GC skew at each codon position in a species at a given codon position and average GC skew of that species. The codon positions are marked by different markers as given in the figure legend.

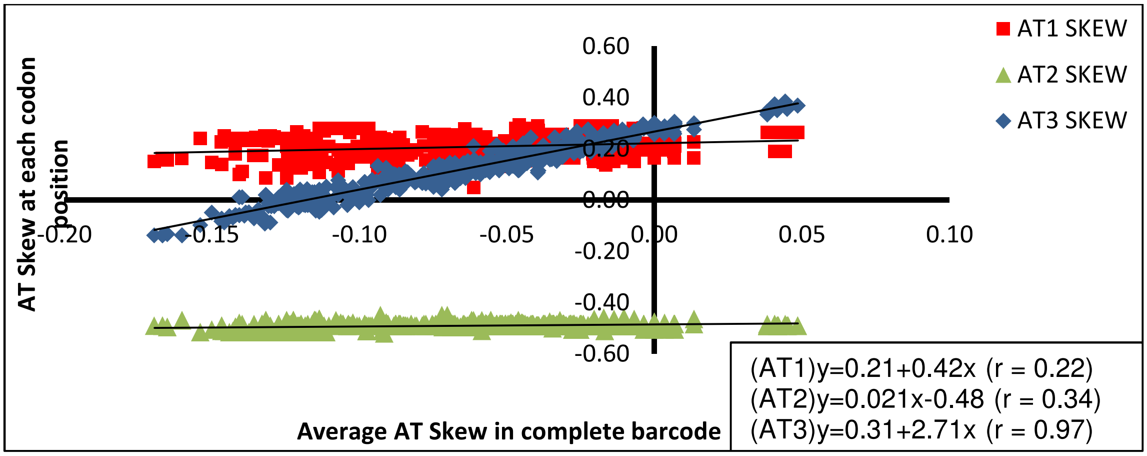

Figure 4.10 Correlation between AT skew at each codon position and average AT skew in species of different orders of Indian freshwater fishes.

Each point represents correlation between AT skew at each codon position in a species at a given codon position and average AT skew of that species. The codon positions are marked by different markers as given in the figure legend.

4.2.2 Nucleotide pair frequencies

The frequency of occurrence of nucleotide pairs is calculated as represented in Table 4.6. When two nucleotide sequences are compared, the frequencies of 10 or 16 different types of nucleotide pairs can be computed. As shown in the table 4.6, Identical pairs accounts for most of the nucleotide pair frequencies. Transition was found to be dominant over transversion with transitional sites occurring 1.2 times more than transversional sites when compared for entire barcode length. Similarly transition to transversion ratio was found to be 1.5 and 1.1 at 2nd and 3rd codon positions. However, at 1st codon position transition was found to be significantly high with R (transition/transversion) equal to 4. Among other nucleotide pairs TC and CT held the significant share in overall and 3rd codon position followed by TA, AT and CA, AC nucleotide pairs. GC and GT nucleotide pairs represented the lowest fraction of nucleotide pair.

Substitutions of nucleotides were found to be more prominent in 3rd codon position than in 1st and 2nd codon position. Second codon position exhibited least number of substitutions for all nucleotide pairs. 1st codon position showed highest number of substitution of the nucleotide pair CT and TC. For remaining nucleotide pairs substitution in the 1st codon was significantly low.

Table 4.6 Nucleotide pair frequencies of 1383 sequences of 10 orders of Indian freshwater fishes across 510bp region of *COI* gene.

	Ii	si	sv	R	TC	TA	TG	CT	CA	CG	AT	AC	AG	GT	GC	GA
Avg	411	54	44	1.2	18	10	2	19	10	2	8	8	8	2	2	9
1st	159	9	2	4.0	3	0	0	3	0	0	0	0	1	0	0	1
2nd	168	1	1	1.5	0	0	0	0	0	0	0	0	0	0	0	0
3rd	85	44	41	1.1	15	10	2	16	10	2	8	7	6	2	1	8

All frequencies are averages (rounded) over all taxa. ii = Identical Pairs; si = Transitional Pairs; sv = Transversional Pairs; R = si/sv.

4.2.3 Codon usage

There are 64 (43) possible codons that code for 20 amino acids (and stop signals) so one amino acid may be encoded by several codons (e.g., serine is encoded by six codons in nuclear genes). It is therefore interesting to know the codon usage for each amino acid. The numbers of the 64 codons used in a gene is computed for all examined sequences. In addition to the codon frequencies, the relative synonymous codon usage (RSCU) statistic is also computed. Many amino acids are coded by more than one codon; thus multiple codons for a given amino acid are synonymous. However, many genes display a non-random usage of synonymous codons for specific amino acids. A measure of the extent of this non-randomness is given by the Relative Synonymous Codon Usage (RSCU) (Sharp et al. 1986).

The RSCU for a particular codon (i) is given by RSCUi = Xi / åå Xi/n.

where Xi is the number of times the i^{th} codon has been used for a given amino acid, and n is the number of synonymous codons for that amino acid.

The codon usage is calculated (Table 4.7) using vertebrate mitochondrial codon table as the template for 170 codons of 1383 sequences used in the study. The codon usage pattern was found to be similar to vertebrate mitochondrial codon table and much variation was not seen. As found in other vertebrates, two more stop codons are found in fishes, AGA and AGG in addition to the usual UAG and UAA. However, our sequences did not contain any stop codons and UGA coded for tryptophan. Further, AUA coded for methionine instead of Isoleucine. Differences in the frequency of occurrence of synonymous codons in coding DNA were seen. Among the six codons coding for Leucine preference was seen for CUA and CUU. In the studied strand cysteine was not coded by any codon. Most of the preferred codon were seen to have A at the third codon position with few having C at the third codon position. Only two amino acids, Arginine and Tyrosine showed preference for codon with U at the third codon position. However, G was least preferred at the third codon position.

Table 4.7. Codon Usage for 1383 sequences of Indian freshwater fishes across 510bp region of *COI* gene.

Codon	Count	RSCU	Codon	Count	RSCU	Codon	Count	RSCU	Codon	Count	RSCU
UUU(F)	4.2	0.83	UCU(S)	2.4	1.37	UAU(Y)	2	1.32	UGU(C)	0	2
UUC(F)	5.9	1.17	UCC(S)	2.8	1.56	UAC(Y)	1	0.68	UGC(C)	0	0
UUA(L)	3.8	0.9	UCA(S)	4	2.27	UAA(*)	0	0	UGA(W)	3.6	1.79
UUG(L)	0.6	0.14	UCG(S)	0.2	0.14	UAG(*)	0	0	UGG(W)	0.4	0.21
CUU(L)	6.3	1.48	CCU(P)	2.3	0.75	CAU(H)	1.1	0.75	CGU(R)	0.3	0.66
CUC(L)	3.8	0.9	CCC(P)	4.2	1.4	CAC(H)	1.9	1.25	CGC(R)	0.1	0.12
CUA(L)	8.5	2.02	CCA(P)	4.9	1.63	CAA(Q)	2.6	1.81	CGA(R)	1.5	3.04
CUG(L)	2.3	0.55	CCG(P)	0.6	0.22	CAG(Q)	0.3	0.19	CGG(R)	0.1	0.18
AUU(I)	9.5	1.35	ACU(T)	2.4	0.79	AAU(N)	3.7	0.81	AGU(S)	0.1	0.06
AUC(I)	4.6	0.65	ACC(T)	2.7	0.88	AAC(N)	5.3	1.19	AGC(S)	1.1	0.6
AUA(M)	6.8	1.39	ACA(T)	6.8	2.2	AAA(K)	0.9	1.79	AGA(*)	0	0
AUG(M)	3	0.61	ACG(T)	0.4	0.13	AAG(K)	0.1	0.21	AGG(*)	0	0
GUU(V)	3.8	1.13	GCU(A)	4.2	0.95	GAU(D)	1.9	0.75	GGU(G)	2.1	0.62
GUC(V)	2.4	0.71	GCC(A)	6.8	1.54	GAC(D)	3.1	1.25	GGC(G)	2.1	0.62
GUA(V)	5.9	1.79	GCA(A)	6	1.35	GAA(E)	0.9	1.74	GGA(G)	6	1.76
GUG(V)	1.2	0.36	GCG(A)	0.7	0.15	GAG(E)	0.1	0.26	GGG(G)	3.4	1

4.2.4 Sequence information in the *COI* barcode using R_{seq} value

A total of 1307 sequences belonging to 160 species of Indian freshwater fishes were analyzed to reveal the variation pattern of sequence conservation of *COI* gene across various taxa level. Sequences were trimmed to eliminate any indels or missing nucleotides and finally a consensus of 510 bp sequences was retrieved. To understand the position of this 510 bp in the *COI* gene, a BLAST search was carried out against bovine *COI* sequence available in database (Accession number: HQ860420) for which the complete sequence and the structure of *COI* have been studied. The search revealed the first nucleotide of our consensus sequence lied in 98th position of bovine *COI* (accession number: HQ860420.1) sequence and showed a fair amount of similarity all along, with the last sequence lying close to 608th position. ORF Finder located the start codon at 35th position of the bovine sequence. Thus, the first nucleotide of our sequences lied at the 1st codon position.

Sequence analysis revealed that conservation of sequences shows a hierarchical pattern. Figure 4.11 reveals the pattern of variation of sequence conservation across taxa. As we move higher in taxonomic hierarchy (i.e. from species to order) nucleotide conservation decreases. The three orders studied here shows 47% - 71% conservation in their nucleotide composition followed by 77% - 98% conservation within their families. Genus showed variation in nucleotide conservation from 77%- 99%. Families having only one or few genus showed high range of conservation (>90%) which is actually a reflection of the conservation in the underlying genus.

A total of 237 sites were found to be variable within Siluriformes order while 268 and 286 sites were found to be variable in Perciformes and Cypriniformes respectively out of the total of 510 nucleotide base pairs. The degenerate nature of genetic code was prominently seen with most of the changes (64.72 %) occurring in third codon position followed by first codon position (24.7%) and second codon position (10.51%) accounting for the lowest among the lot.

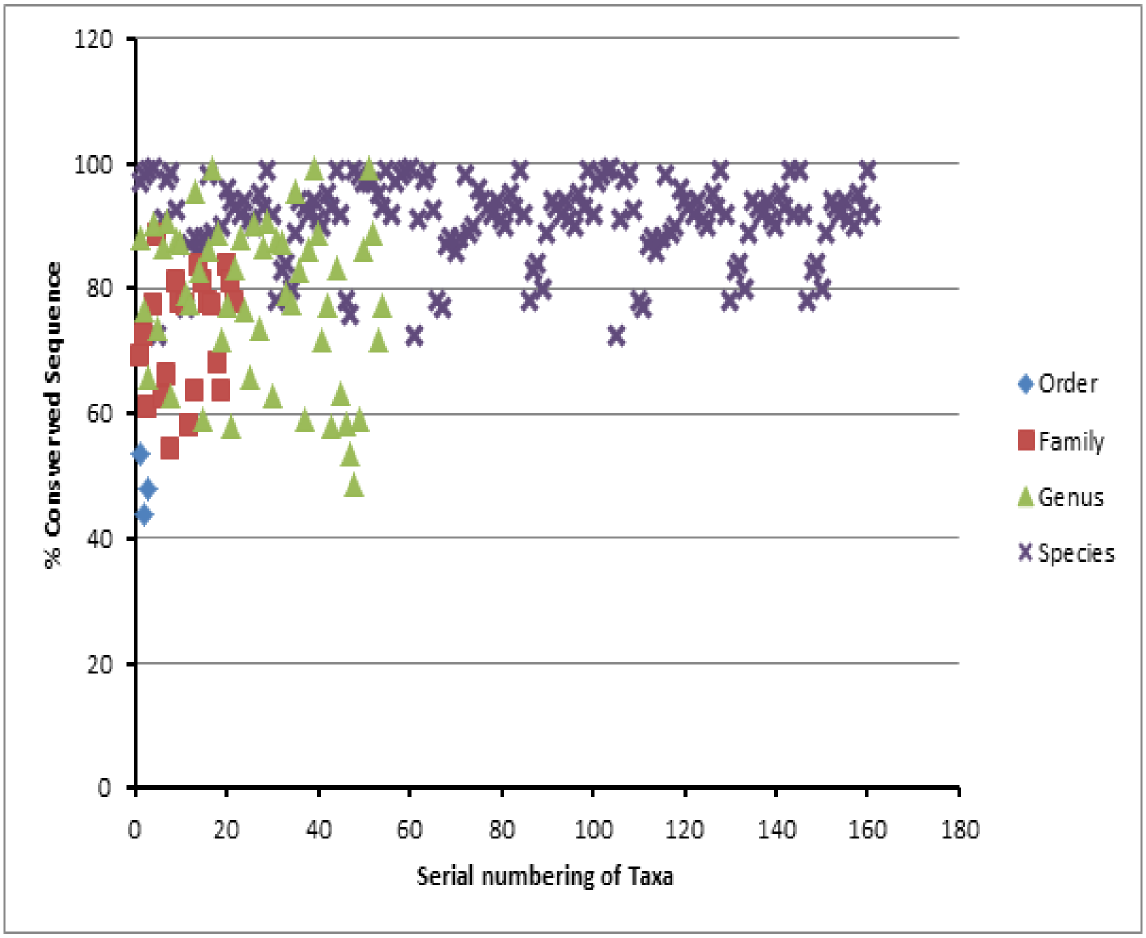

Figure 4.11 Pattern of distribution of sequence conservation across various taxonomic ranks of Indian freshwater fishes.

The horizontal axis represents serial arrangement of different taxa. Different markers mark different taxonomic ranks viz: orders are marked by diamonds, families by squares, genera by triangles and species by crosses. Vertical axis represents percentage of conserved sequence in each taxon. Clustering of the markers in the scatter diagram represents the variation pattern of conserved sequences in the representative taxa**.**

The degree of sequence conservation was studied for each of these variable sites in *COI* gene of all the three orders to understand the amount of sequence information available in each taxon that could facilitate their identification. In the order Cypriniformes, among the 120 sites approaching saturation with single nucleotides, 66 sites belonged to second codon position ,while 47 sites belonged to first codon position and only 2 sites were in third codon position. 81 sites with Rseq value lying in the range of 1- 1.49 contained either two bases with equal probability or one being dominant over the other. Further 85 sites showed random possibility of having any of the nucleotide bases. Then the families under Cypriniformes order were analyzed, which had representative barcodes of more than one genus. Cyprinidae with 30 representative genus and Balitoridae with 4 genus were checked for sequence information in the sites which showed variation when all members of Cypriniformes order were analyzed. Among these sites, 135 sites were conserved in Balitoridae family and only two sites were conserved in Cyprinidae family. A large number of sites (122 in Balitoridae and 203 in Cyprinidae) had possibility of having two nucleotides, however lesser number of sites showed possibility of having random nucleotides. Further, genera with more than one representative species were analyzed for sequence information. At this level, more than 90% - 60% of the order level variable sites were found to be conserved and a small amount (0% – 5%) of sites had three or more nucleotides.

Similar analyses were carried out for Siluriformes and Perciformes orders. In Siluriformes, four families with more than one representative genus were further analyzed, followed by analysis of five genera with more than one species. A similar pattern of sequence information pattern was reflected in each taxon. At order level, (34%) sites in Siluriformes and (32%) in Perciformes showed random variation of nucleotides. However, the number of randomly variable sites showed a steady decrease as we went to lower taxa in both the orders (0-10%).

Amino acid sequence showed a high range of conservation for all the three orders even at higher taxa level. At order level, Cypriniformes had 90 variable sites out of 170 total amino acids, Siluriformes showed 49 while Perciformes had 50 variable amino acid positions. The information content for the variable amino acid position at each taxa level of all the members of these orders are shown in (Table 4.8). The maximum R_{seq} for amino acid residues is $\log_2 20 = 4.322$ and positions showing this value reflected complete conservation of the amino acid. The value of R_{seq} of amino acid sequence varied between 3.12- 4.32 with all the orders having 90% of the sites in the range of 4.02 - 4.32 R_{seq} value, reflecting high conservation of sequence. In Cypriniformes order, of 82 sites close to saturation, all were conserved in the family Balitoridae. However in family Cyprinidae, all but 2 sites showed R_{seq} in the range of 4.02 - 4.31. Further analysis of Cyprinidae family revealed that most of these 79 sites are conserved at genus level with an average of 5% sites having a possibility of bearing two nucleotides at one site. Most of the sites that did not attain conservation at genus level were found to be conserved at species level, indicating a small amount approximate 5% change at species level.

Similarly, in the order Siluriformes, of the 49 variable amino acid sites, 42 showed R_{seq} in range of 4.02 - 4.31. Most of the families of Siluriformes order showed 90% sequence conservation. At genus level sequence conservation further increased and only few changes were seen at species level. Similar results were obtained for the order Perciformes with 82% sites at order level, lying in the range of 4.02 - 4.31 R_{seq} value. At family level in Perciformes order, only the family Channidae was represented with barcodes of multiple species under the genus Channa. In this genus however, 17 of the 50 variable sites lied in the range of 3.72 - 4.01 R_{seq} value showing possibility of more than two nucleotides in many members at each of these sites. *Channa barca,* represented by 5 individual barcode sequences, showed variation from the remaining members of the genus at 18 sites with most of these substitutions being radical in nature. Thus, these changes in Channidae genera were species specific.

Table 4.8 Frequency distribution of R$_{seq}$ value of variable nucleotides and amino acids in *COI* gene of different taxonomic ranks of Indian freshwater fishes.

TAXA			Nucleotides R$_{seq}$ Values					Amino acid R$_{seq}$ Values				
ORDER	FAMILY	GENUS	0-0.49	0.5-0.99	1-1.49	1.5 - 1.99	2 - 2.49	3.12-3.41	3.42-3.71	3.72-4.01	4.02 - 4.31	4.32 - 4.61
Cypriniformes			23	62	81	120	0	0	3	5	82	0
	Baltoridae		6	23	118	4	135	0	3	6	79	2
		Acanthocobitis	0	0	20	2	264	1	1	1	0	87
		Nemacheilus	0	0	2	0	284	0	0	0	0	90
		Schistura	0	2	68	0	216	0	0	0	0	90
		Balitora	0	0	1	0	285	0	0	0	0	90
	Cyprinidae		20	61	79	124	2	1	0	3	1	85
		Labeo	3	8	84	27	164	0	0	0	7	83
		Tor	0	1	12	44	229	1	0	0	14	75
		Puntius	18	40	92	53	83	2	2	8	7	71
		Barilius	2	23	92	77	92	7	3	19	7	54
		Neolissichilus	0	1	14	7	264	1	0	0	2	87
		Garra	0	8	72	26	180	0	3	2	0	85
		Osteobrama	0	4	83	1	198	0	6	1	0	83
Siluriformes			26	55	83	73	0	0	2	5	42	0
	Schilbeidae		5	13	78	42	99	0	1	1	3	44
		Eutropiichthys	0	0	59	2	176	0	0	2	0	47
	Siluridae		5	16	82	37	97	1	1	0	1	46
		Ompok	5	2	94	18	118	0	2	0	1	46
	Bagridae		23	46	96	34	38	1	2	6	8	32
		Sperata	0	0	50	0	187	0	1	0	0	48
		Mystus	13	44	101	31	48	1	1	4	8	35
	Sisoridae		2	25	64	109	37	1	1	0	20	27
		Glyptothorax	0	5	76	53	103	1	1	0	6	41
Perciformes			24	61	95	86	1	4	3	1	41	1
		Channa	16	41	105	53	53	5	0	18	8	18
		Lates	0	0	5	36	227	0	0	1	4	45

Chapter 4.3 DNA barcode based taxonomic rank assignment using distance method.

4.3.1 Species level identification

A total of 1383 mitochondrial *COI* barcode sequences of 175 species belonging to 10 orders, 34 families, 77 genera (Appendix 2) and constituting almost 20% of Indian freshwater fishes were retrieved. Among them, 172 barcode sequences, representing 70 different species, were generated from North-East India. No insertion, deletion or stop codons were observed in any sequence. The absence of stop codon as well as coherent partial amino acid codes confirmed them to be a partial fragment of mitochondrial *COI* gene. In the dataset, for most species, multiple specimens were used to document intraspecific variability (with an average of 5 specimens per species). However, 30 species were represented by single specimen only.

The sequence analysis revealed a hierarchical increase in K2P mean divergence across all the taxon from within species (1.6%, S.E = 0.1) to within genus (9.925%, S.E = 2.7), within family (15.66%, S.E = 1.9) and within orders (25.32%, S.E = 2.3) and is presented in Table 4.9.

Table 4.9 Summary of genetic divergences (K2P model) for each taxonomic level of comparison.

Comparison within	Taxa(n)	Mean	Min	Max	S.E
Species	175	1.14	0	18.87	0.001
Genus	77	7.16	0	21.42	0.011
Family	34	15.66	11.51	32.23	0.019
Order	10	25.32	20.42	45.41	0.023

Conspecific divergence between sequences of same species varied in the range 0% - 18% .While, congeneric divergence, between sequences of different species under same genus, varied from 0% - 21%. This overlap in the distribution of conspecific and congeneric means caused hindrance in defining the threshold value for species boundary. To resolve the problem, the NJ tree (Appendix 3) was explored and 3 general cases were found, as shown in Figure 3.1.

1) Sequences, with same species name, exhibiting cohesive clustering by the conspecies and distinct clustering by the congeners with high bootstrap support (90-100%).

2) Sequences, with same species name, not exhibiting cohesive clustering by the conspecies.

3) Sequences, with a different species name, exhibiting cohesive clustering.

Among the 3 cases, only the first group abided by the first principle of DNA barcoding (same named species should cluster cohesively and distinctly from the rest) and represented 82% of the total 175 species. These sequences were considered to represent true species.

In this set, the highest conspecific mean divergence was shown by *Mastacembelus armatus* (2.3%, S.E = 0.1) and the lowest congeneric divergence was shown between *Tor khudree* and *Tor mosal* (2.9%, S.E = 0.2). These values were used to define the species boundary in this study as shown in Figure 4.12. All the remaining species of this group showed conspecific divergence lower than 2.3% and congeneric divergence higher than 2.9%.

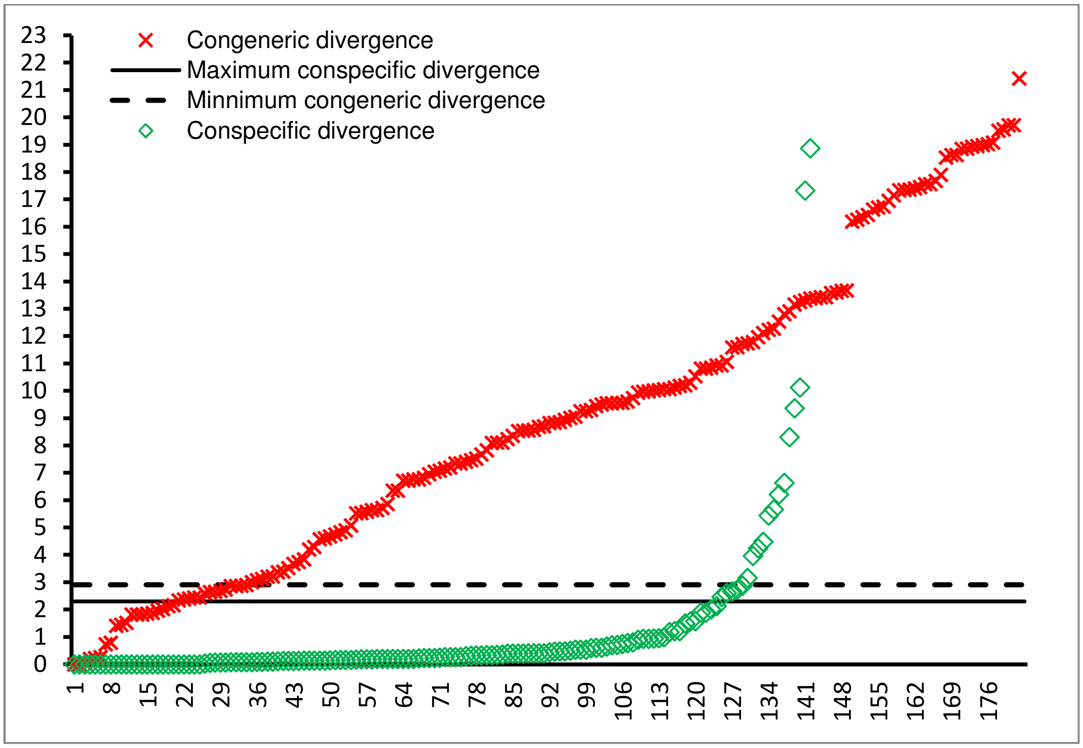

Figure 4.12 Distribution of conspecific and congeneric K2P mean divergence among 175 species of Indian freshwater fishes (arranged in ascending order).

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. Data series marked by 'cubes' represent conspecific divergence of 142 species, which were represented by more than one sequence. 77.7% of the total 175 species showed divergence below 1.14% and represented true species. Sequences with divergence between 1.14% and 2.3% represented recently diverged species and geographically isolated population of same species (e.g. *Badis badis, Schizothorax progastus, Channa gachua, Puntius sarana, Macrognathusaral, Puntius chelynoides, Tor malabaricus, Channa striata, Epalzeorhynchos bicolor, Acanthocobitis botia, Mastacembelus armatus*). Sequences that lie above congeneric threshold line were suspected of having cryptic species diversity within single named species and are discussed in Table 4.10.

Table 4.10 Same named species with divergence above minimum congeneric value and formed two or more distinct sub-clusters.

Species name	Dist.	S.E	# sequence	No of clusters	Bootstrap
Lates calcarifer	2.6	0.002	94	2	100, 100
Heteropneustes fossilis	2.66	0.003	14	2	100, 100
Pterygoplichthys pardalis	2.73	0.005	3	2 subclusters in single node	100, 100
Channa orientalis	2.87	0.004	11	2 subclusters in single node	100, 99
Puntius filamentosus	3.16	0.005	4	2	100, 100
Barilius bendelisis	3.96	0.005	40	dispersed clusters	78, 100
Bariliu barna	4.24	0.005	29	dispersed clusters	78, 100
Osteobrama cotio cotio	4.48	0.007	4	2	100, 100
Labeo bata	5.43	0.007	17	2	100, 100
Devario devario	5.66	0.01	2	2	100, 101
Clupisoma garua	6.21	0.009	3	2	100, 102
Clarias batrachus	6.63	0.05	18	2	100, 100
Garra hughi	8.3	0.01	5	2	98, 82
Barilius tileo	9.36	0.012	6	2	100, 100
Channa marulius	10.13	0.011	6	2	100, 100
Glossogobius giuris	17.32	0.021	2	2	100, 100
Puntius conchonius	18.87	0.02	2	1	100, 100

The remaining two groups, that constituted 18% of the studied species, were considered problematic.

Sequences, with same species name, formed 2 or more distinct sub-clusters and showed divergence above maximum conspecific value. Here 17 species viz., *Puntius conchonius* (n = 2), *Osteobrama cotio cotio* (n = 4), *Pterygoplichthys pardalis* (n = 3), *Devario devario* (n = 2), *Clupisoma garua* (n = 3), *Garra hughi* (n = 5), *Channa marulius* (n = 6), *Glossogobius giuris* (n = 2), *Barilius tileo* (n = 6), *Lates calcarifer* (n = 94), *Heteropneustes fossilis* (n = 14), *Clarias batrachus* (n = 18), *Labeo bata* (n = 17), *Channa orientalis* (n = 11), *Puntius filamentosus* (n = 4), *Barilius bendelisis* (n = 40), *Barilius barna* (n = 29) formed separate or dispersed cluster and are mentioned in Table 4.10

Clarias batrachus, represented by 18 sequences (mean = 6.6%, S.E = 0.6), formed 2 distinct clusters with 6 sequences (FJ459456-59, JQ667517-18) being separated from the remaining 12 sequences of the species by 11.5% mean distance (Figure 4.13(i) a). However, divergences within the individual clusters were 1.2% (S.E = 0.2) and 0.6% (S.E = 0.2) respectively. *Heteropneustes fossilis* showed conspecific divergence of 2.6% (S.E = 0.3) and one sequence of *Heteropneustes fossilis* (HQ009491) clustered away from the rest of the 13 sequences with mean K2P divergence between them being 10.5% (S.E = 0.3). This single sequence of *Heteropneustes fossilis* clustered with *Heteropneustes microps* with a congeneric distance of 0.2% (S.E = 0.1) (Figure 4.13(i) b). With the exclusion of the single aberrant sequence, the remaining sequences of *Heteropneustes fossilis* formed close cluster with a conspecific divergence of 0. 9% (S.E = 0.1). *Lates calcarifer*, represented by 94 sequences and conspecific divergence of 2.6% (S.E = 0.2), formed 2 different clusters with 4 sequences (HQ219138-HQ219141) clustering separately from the remaining 90 sequences of the same species.

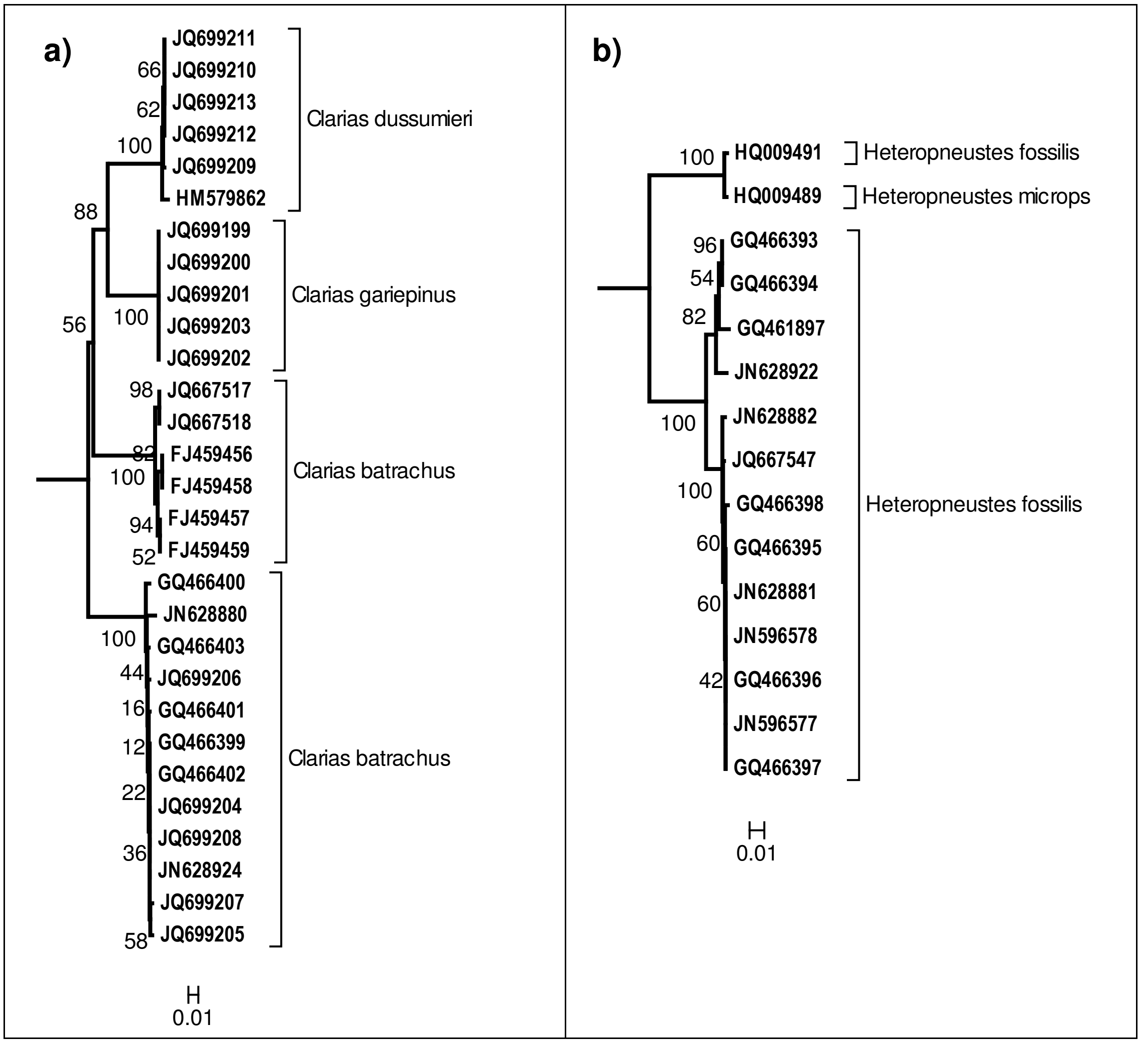

Figure 4.13 (i) Sections of Neighbor–Joining (from Appendix 3) tree showing the problematic groups of (a) *Clarias batrachus* **(b)** Heteropneustes genus.

The NJ tree was constructed based on K2P model with 1000 bootstrap support. The problematic groups presented in the above figure are: In (a) *Clarias batrachus* have clustered into two distinct clusters. In (b) one sequence of *H. fossilis,* HQ009491 forms cluster with *a* sequence (HQ009489) of *H.microps*.

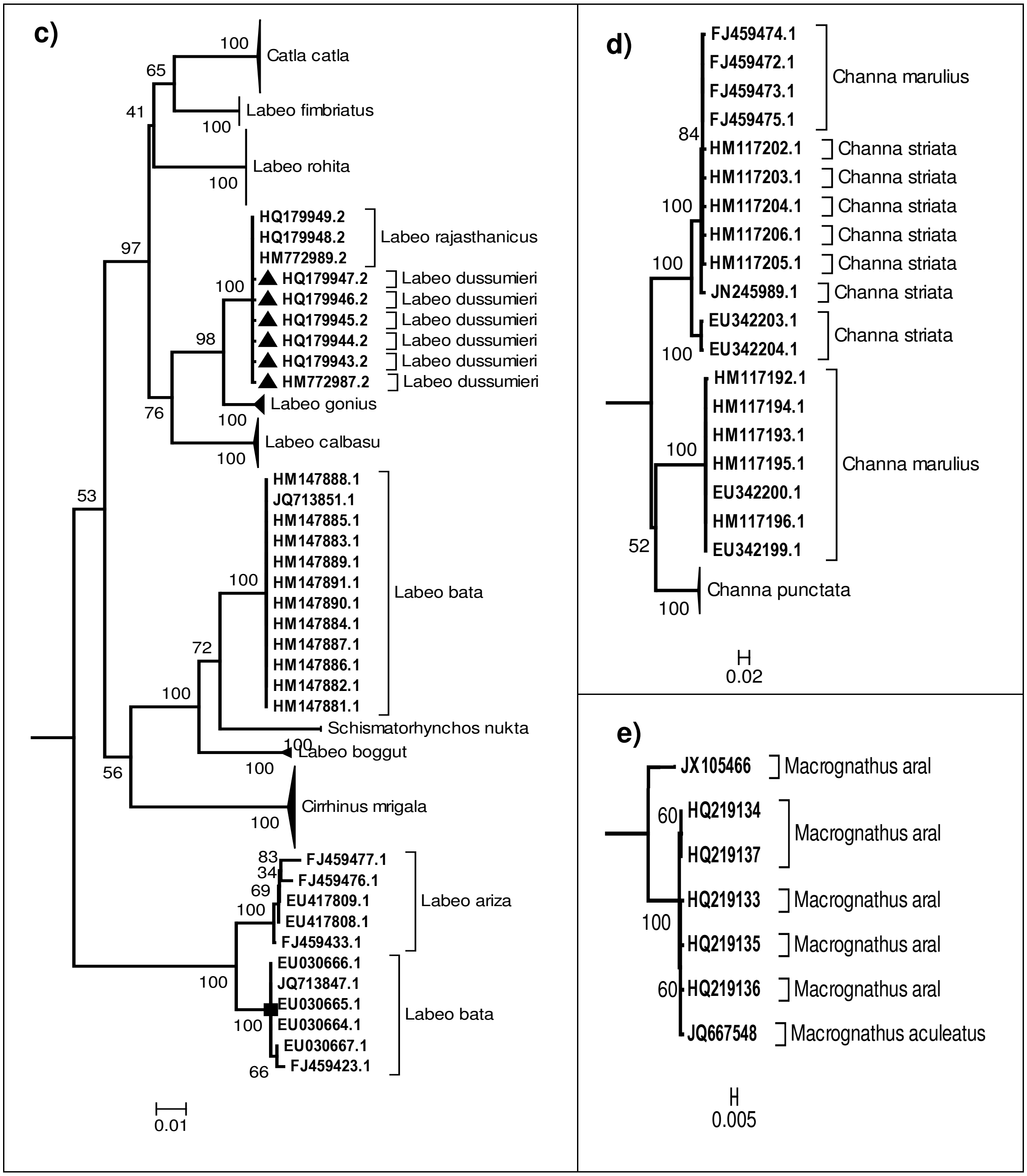

Figure 4.13 (ii) Sections of Neighbor–Joining tree (from Appendix 3) showing the problematic groups of (c) Labeo genus (d) Channa genus (e) Macrognathus genus.

The problematic cases are: In (c) *L. dussumieri and L. rajasthanicus* forms single cluster while few sequences of *L. bata* clusters with *L. ariza*. In (d) some sequences of *C. marulius* clusters with *C. striata*. In (e) JQ667548 sequence of *M. aculeatus* clusters *with M. aral.*

Similarly, *Labeo bata* with a conspecific mean of 5.4% (S.E = 0.7) formed 2 distinct clusters with one cluster of 6 sequences (EU30664-67, JQ713847, FJ459423), clustering away with mean divergence between the 2 clusters being 11.5% (S.E = 0.6) and mean distance within the clusters being 0% and 2% respectively (Figure 4.13(ii) c).

In Barilius genus, 69 sequences of the 2 species *Barilius bendelisis and Barilius barna* were indistinguishable in the NJ and ML tree as their respective representative sequences did not form unique clusters. Rather, they collectively formed 6 clusters and were designated as Barilius cluster1 (comprising accession number FJ459411, FJ459412, FJ459418-22, JN965193, JN965195, JN965196, JX1054820-65 of *Barilius bendelisis* and accession number EU417797-99, HM042258-63, HM042170-81 of *Barilius barna*), Barilius cluster2 (accession number EU822331-33, HM042230-35 of *Barilius bendelisis*), Barilius cluster3 (accession number HM042236-41 of *Barilius bendelisis*), Barilius cluster4 (accession number HM042248-53 of *Barilius bendelisis*), Barilius cluster5 (accession number HM042242-47 of *Barilius bendelisis*), Barilius cluster6 (accession number HM042164-69 of *Barilius barna*) in the NJ (Figure 4.14). The sequences within each of the 6 cohesive clusters of Barilius showed mean K2P divergence below the value of maximum conspecific divergence while the mean divergence between the sequences of separate clusters were above the value of minimum congeneric divergence (Table 4.11).

In another such instance, *Channa orientalis*, with mean conspecific distance of 2.8% (S.E = 0.4), diverged into 3 clusters (FJ459480 - FJ459484; JX105470-74 and JQ667514) with average distance between the 3 clusters being 3.5% (Figure 4.13(ii) d). For the remaining 11 species, the number of representative sequences was not sufficient to draw any conclusion.

Table 4.11 Mean divergence within and between clusters of Barilius genus.

Clusters of species	**Mean divergence within and between clusters of Barilius genus**										
Barilius cluster1	1.3										
Barilius cluster2	4.8	0.7									
Barilius cluster3	4.7	3.2	0.0								
Barilius cluster4	6.0	4.4	3.0	0.2							
Barilius cluster5	9.9	8.2	6.9	6.9	0.0						
Barilius cluster6	9.2	11.9	10.8	12.4	16.3	0.0					
Barilius vagra	13.4	16.3	16.1	16.3	18.8	17.6	0.1				
Barilius gatensis	20.4	22.8	22.4	23.1	23.9	22.1	15.1	0.2			
Opsarius bakeri	17.6	19.6	19.3	20.0	21.6	19.2	16.5	13.1	0.9		
Opsarius canarensis	18.2	20.2	19.7	20.6	23.9	19.8	16.8	12.2	5.0	n/c	
Barilius tileo	20.5	23.4	23.7	24.9	26.2	22.6	18.5	16.8	17.7	16.1	12.7

Diagonals represent mean divergence (in bold) within the group and the remaining cells represent pair-wise divergence between the groups.

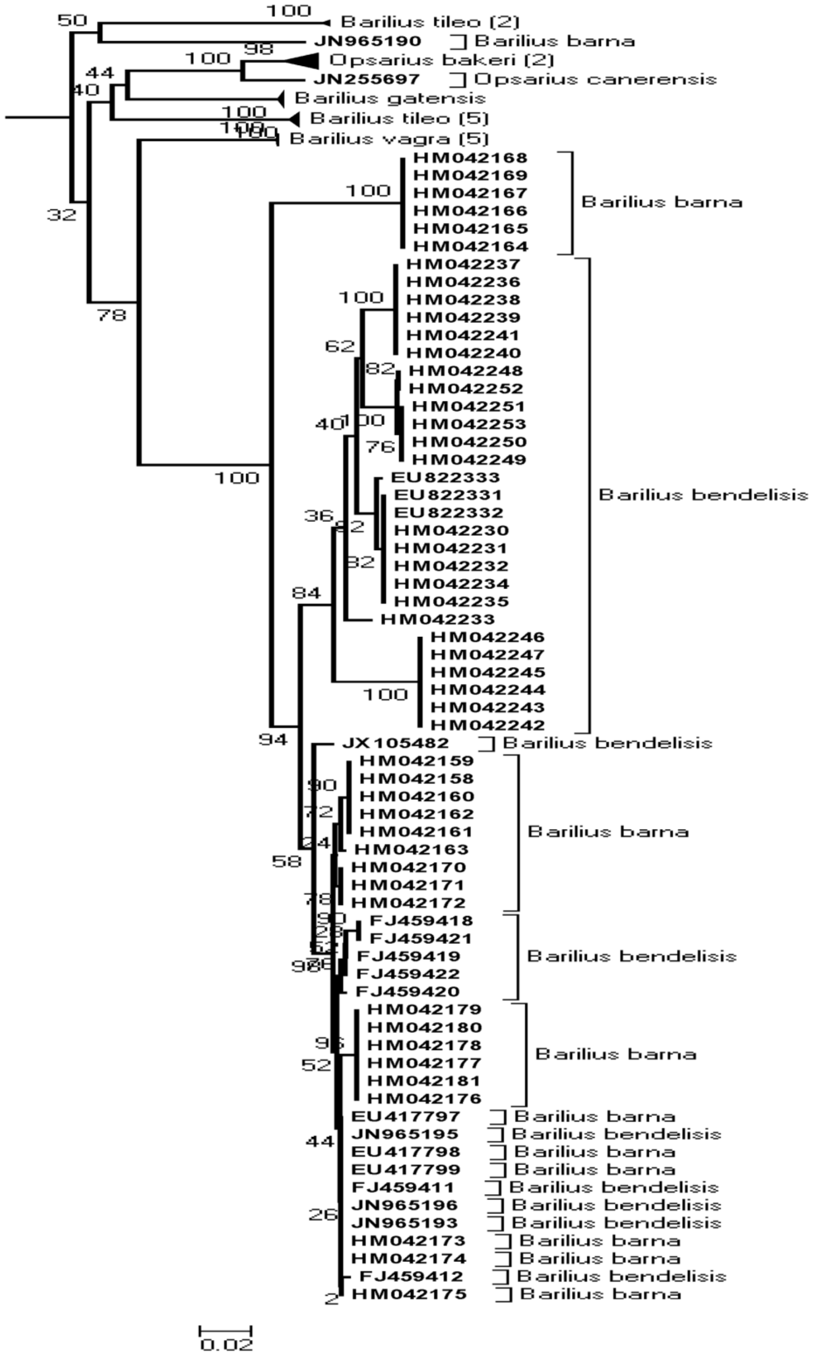

Figure 4.14 Neighbor joining tree showing clustering of species of Barilius genus. The numbers at the nodes are bootstrap values based on 1000 replications. Species exhibiting deep intraspecific divergence are marked by square and those with narrow interspecific distance are marked by triangle; those showing both types of discrepancy are marked with pyramid.

In Group 3, sequences, with different species name, (that are expected to cluster separately) clustered cohesively and showed divergence below maximum conspecific divergence *(Labeo dussumieri versus Labeo rajasthanicus* (Figure 4.13(ii) c) , *Poecilia sphenops versus Poecilia velifera, Aspidoparia morar versus Aspidoparia jaya, Mystus vittatus versus Mystus horai, Mystus tengara versus Neotropius atherinoides, Macrognathus aral versus Macrognathus aculeatus* (Figure 4.13(ii) e) and are detailed in Table 4.12. *Poecilia sphenops versus Poecilia velifera and Labeo dussumieri versus Labeo rajasthanicus* exhibited a divergence of 0%. Moreover, pairs with interspecific divergence between 1.14% and 2.3% were not clustered together, but formed dispersed clusters supported with low bootstrap value e.g. *Tor tor* together with *Tor putitora, Tor macrolepis, Tor mussullah, Tor mosal mahanadicus; Poecilia latipinna* with *Poecilia sphenops and Poecilia velifera*.

Table 4.12 Different named species that clustered cohesively and showed divergence below maximum conspecific divergence.

# No	Species 1	Species 2	% distance	S.E
1	*Labeo dussumieri*	*Labeo rajasthanicus*	0	0
2	*Poecilia sphenops*	*Poecilia velifera*	0	0
3	*Tor mosal mahanadicus*	*Tor macrolepis*	0.036	0
4	*Aspidoparia morar*	*Aspidoparia jaya*	0.199	0.001
5	*Tor putitora*	*Tor macrolepis*	0.224	0
6	*Tor mosal mahanadicus*	*Tor putitora*	0.26	0.001
7	*Mystus vittatus*	*Mystus horai*	0.427	0.001
8	*Mystus tengara*	*Neotropius atherinoides*	0.736	0.002
9	*Macrognathus aral*	*Macrognathus aculeatus*	0.792	0.001

4.3.2 Special case of the unusual genetic diversity in *Clarias batrachus.*

Distance analysis of *COI* barcode sequence based on K2P method of *Clarias batrachus* species from India showed large conspecific mean divergence (6.85 ± 0.76) % thus indicating either the presence of different haplotypes of *Clarias batrachus* in India or presence of some mislabeled sequence. To explore the possibility of presence of distinct haplotypes of *Clarias batrachus*, the sequences were grouped according to different geographic location (within India) from which they have been collected and genetic variation within and between the groups were analyzed (Table 4.13). Low conspecific mean genetic distance was observed for *Clarias* batrachus species of a particular geographic location. Mean genetic distance of *Clarias batrachus* from Alibagh coast, Mumbai; West Bengal and Lala, Assam were found to be 0.5%, 0.31% and 1.5% respectively, while remaining 10 individuals for which no location has been specified, shows an overall mean of 0.15%. The mean distance between populations of Lala, Assam and Mathabhanga, West Bengal were found to be as high as 13.63%. Many other interpopulation divergences also showed similar high range of conspecific divergence that lies in the range of congeneric divergence. However, not all population exhibited high divergence e.g. species from Lala, Assam and Alibagh coast, Maharashtra shared narrow divergence of 0.934%, thereby indicating that geographic isolation can be partly but not solely responsible for the observed high genetic divergence within Indian *Clarias batrachus* species.

Altogether, six sequences of Indian *Clarias batrachus* cluster separately from all other sequences of *Clarias batrachus* and diverge from the other Indian *Clarias batrachus* sequences by (13.21 ± 1.25) %. Moreover, with exclusion of these six sequences, conspecific divergence of remaining Indian *Clarias batrachus* sequences lowers to (1.21 ± 0.25) %. This indicates the presence of cryptic species diversity within Indian *Clarias batrachus* and that the six sequences might represent a different species. In previous taxonomic classification, many species showed conflict with *Clarias batrachus* and some are considered to be synonyms (Ferraris, 2007) like *Clarias assamensis, Clarias* *punctatus, Clarias marpus, Clarias magur, Clarias fuscus* etc (Day, 1958) . Among these species, many were traditionally found in India and some were introduced later (Talwar and Jhingran, 1991). Thus, barcode analysis indicates that discrepancy in conspecific mean divergence of Indian *Clarias batrachus* may be due to the presence of one or more of these species wrongly identified as *Clarias batrachus*. *Clarias batrachus* sequences from Thailand and Philippines showed low conspecific divergence of 0 % and (0.121 ± .08) % and formed single cohesive cluster in Neighbor joining tree. Further, sequences of the two countries form two distinct clusters (Figure 4.15) with divergence between them being 2.5%. This indicates that sequences of *Clarias batrachus* from these two countries represent unique haplotype specific for each country. This shows the presence of high genetic diversity of *Clarias batrachus* species across the world. Species of *Clarias batrachus* shows high range of divergence between different countries, which is easily traceable by *COI* barcodes.

Table 4.13 Genetic divergence of different species of Clarias genus across various geographical location calculated using K2P model.

Species	Geographic Location	Mean Dist.	S.E
Clarias batrachus	Thailand	0.12	0. 08
Clarias batrachus	Philippines	0	0
Clarias batrachus	Vietnam	N/A	N/A
Clarias batrachus	India (all sequences)	6.85	0.07
Clarias batrachus	India (not available)	5.66	0.62
Clarias batrachus	India (Maharashtra)	0.19	0.1
Clarias batrachus	India(West Bengal)	0.31	0.18
Clarias batrachus	India (Assam)	1.51	0.47
Clarias dussumieri	India (not available)	0.3	0.19
Clarias gariepinus	India (not available)	0.31	0.46

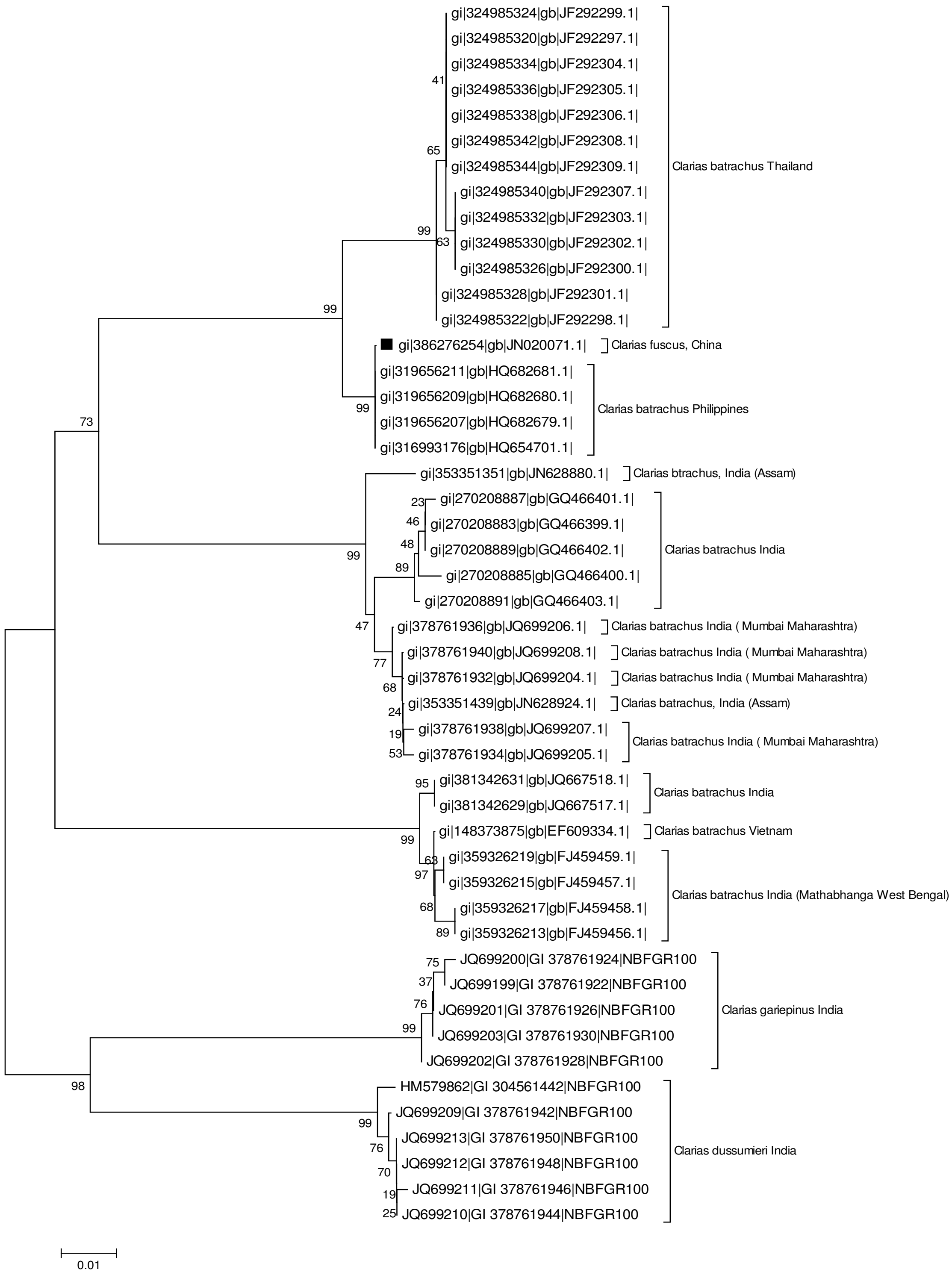
gi|324985324|gb|JF292299.1|
gi|324985320|gb|JF292297.1|
gi|324985334|gb|JF292304.1|
41
gi|324985336|gb|JF292305.1|
gi|324985338|gb|JF292306.1|
gi|324985342|gb|JF292308.1|
65
gi|324985344|gb|JF292309.1|
Clarias batrachus Thailand
gi|324985340|gb|JF292307.1|
gi|324985332|gb|JF292303.1|
99
63
gi|324985330|gb|JF292302.1|
gi|324985326|gb|JF292300.1|
gi|324985328|gb|JF292301.1|
99
gi|324985322|gb|JF292298.1|
■ gi|386276254|gb|JN020071.1|
Clarias fuscus, China
gi|319656211|gb|HQ682681.1|
gi|319656209|gb|HQ682680.1|
99
Clarias batrachus Philippines
73
gi|319656207|gb|HQ682679.1|
gi|316993176|gb|HQ654701.1|
gi|353351351|gb|JN628880.1|
Clarias btrachus, India (Assam)
23
gi|270208887|gb|GQ466401.1|
46
gi|270208883|gb|GQ466399.1|
48
gi|270208889|gb|GQ466402.1|
Clarias batrachus India
99
89
gi|270208885|gb|GQ466400.1|
gi|270208891|gb|GQ466403.1|
47
gi|378761936|gb|JQ699206.1|
Clarias batrachus India (Mumbai Maharashtra)
gi|378761940|gb|JQ699208.1|
Clarias batrachus India (Mumbai Maharashtra)
77
68
gi|378761932|gb|JQ699204.1|
Clarias batrachus India (Mumbai Maharashtra)
24
gi|353351439|gb|JN628924.1|
Clarias batrachus, India (Assam)
19
gi|378761938|gb|JQ699207.1|
Clarias batrachus India (Mumbai Maharashtra)
53
gi|378761934|gb|JQ699205.1|
95
gi|381342631|gb|JQ667518.1|
Clarias batrachus India
gi|381342629|gb|JQ667517.1|
gi|148373875|gb|EF609334.1|
Clarias batrachus Vietnam
99
63
gi|359326219|gb|FJ459459.1|
97
gi|359326215|gb|FJ459457.1|
Clarias batrachus India (Mathabhanga West Bengal)
68
gi|359326217|gb|FJ459458.1|
89
gi|359326213|gb|FJ459456.1|
75
JQ699200|GI 378761924|NBFGR100
37
JQ699199|GI 378761922|NBFGR100
76
JQ699201|GI 378761926|NBFGR100
Clarias gariepinus India
99
JQ699203|GI 378761930|NBFGR100
JQ699202|GI 378761928|NBFGR100
98
HM579862|GI 304561442|NBFGR100
JQ699209|GI 378761942|NBFGR100
99
JQ699213|GI 378761950|NBFGR100
76
Clarias dussumieri India
70
JQ699212|GI 378761948|NBFGR100
19
JQ699211|GI 378761946|NBFGR100
25
JQ699210|GI 378761944|NBFGR100
0.01

Figure 4.15 NJ tree showing genetic divergence of *Clarias batrachus* from different geographic location. *Regional information within a country (if available) is included in parentheses.

4.3.3 Genus level identification

Representative species from 81 genera of Indian freshwater fishes have been barcoded until now. Of them 50 genera are represented by single species and thus monophyly in neighbor joining tree cannot be inferred for these genera. The remaining 31 genera were represented by more than one species ranging from a minimum of 2 to maximum of 16 species per genus (Table 4.14). Of these, 16 genera did not form monophyletic cluster with their representative species.

Mean divergence of the congeners were in the range of 0% - 21.5%. However, species of different genus also showed divergence as low as 0.7% as in case of *Mystus tengara* and *Neotropius atherinodes*. Further, low range of divergence was seen in Neolissochilus and Tor genus (4%-5%), followed by Catla and Labeo genus (6.5%-8%). Many more species of different genus showed mean divergence above 8%, thus, showing wide range of overlap between congeneric and confamilial mean divergence.

In the NJ tree, most genuses showed cohesive clusters. However, many clusters of congeneric species also included species from different genus. All species of Labeo genus clustered within a single root node along with *Catla catla, Cirrhinus mrigala and Schismatorhynchos nukta*. Glyptothorax, Barilius, Clarias, Channa are among the genus with large number of representative species which cluster into cohesive unit. In Mystus genus, all species baring *Mystus montanus* cluster together. Puntius and Tor are the two most diverse genera with their species clustering under different nodes.

Table 4.14 Summary of genera forming distinct cluster in NJ tree.

Sl. no.	Genus	No of species	No of specimen	Single cluster
1	Schistura	2	10	No
2	Aspidoparia	2	1	Yes
3	Barbonymus	2	10	No
4	Barilius	5	84	No
5	Cirrhinus	2	1	No
6	Devario	3	8	Yes
7	Epalzeorhynchos	2	4	Yes
8	Esomus	2	8	No
9	Garra	3	18	No
10	Labeo	9	93	No
11	Neolissochilus	2	20	Yes
12	Opsarius	2	7	No
13	Osteobrama	2	14	No
14	Puntius	16	55	No
15	Rasbora	2	6	No
16	Schizothorax	2	13	No
17	Tor	7	113	Yes
18	Amblyceps	3	4	Yes
19	Clarias	3	28	Yes
20	Gagata	3	13	No
21	Glyptothorax	11	204	Yes
22	Heteropneustes	2	15	Yes
23	Horabagrus	2	3	Yes
24	Mystus	9	39	No
25	Ompok	4	112	Yes
26	Sperata	2	11	Yes
27	Colisafasciata	2	7	Yes
28	Channa	10	85	No
29	Xiphophorus	2	11	Yes
30	Poecilia	4	8	Yes
31	Macrognathus	3	13	No

4.3.4 Family level identification

28 families of Indian freshwater fishes were included in the study. Among them 11 families (Balitoridae, Cobitidae, Cyprinidae, Schilbeidae, Sisoridae, Bagridae, Siluridae, Osphronemidae, Latidae, Cichlidae, Channidae) did not form a monophyletic cluster in NJ tree. The details of the clade formation of the remaining 17 species are given in Table 4.15. Examining the mean divergence for all the members of each family, no threshold was derived that could define the family boundaries. All members of the Cyprinidae family formed cohesive clusters except *Raimas bola*. However, members of Balitoridae family (*Homaloptera montana, Nemacheilus montana*, *Schistura beevani, Schistura corica, Acanthocobitis botia, Salmostoma bacaila*) clustered under a single node within the Cyprinidae node. Members of Channidae family formed monophyletic cluster. *Horabagrus brachysoma* and *Horabagrus nigricollaris* clustered separately from the monophyletic cluster of Bagridae. Remaining families of the order Siluriformes like Siluridae, Sisoridae and Schilbeidae etc formed dispersed clusters.

Table 4.15 Summary of families forming distinct cluster in NJ tree.

#No	Family	Distinct Cluster in NJ tree
1	Notopteridae	multiple genera.
2	Poeciliidae	multiple genera
3	Mastacembelidae	multiple genera
4	Amblycipitidae	single representative genus.
5	Clariidae	single representative genus.
6	Heteropneustidae	single representative genus.
7	Pangasiidae	single representative sequence.
8	Clupeidae	single representative sequence.
9	Erethistidae	single representative species.
10	Loricariidae	single representative species.
11	Gobiidae	single representative species.
12	Nandidae	single representative species.
13	Engraulidae	single representative species.
14	Pristigasteridae	single representative species.
15	Belonidae	single representative species.
16	Callichthyidae	clusters with Bagridae
17	Mugilidae	Single sequence clusters with Chanidae

4.3.5 Order level identification

Most of the families within an order were hierarchically nested within a single node that represented the Order (Table 4.16). Overall, mean divergence within order was 25.32 and varied in the range of 20.42% to 40.41%. With few exceptions, most members of the order Siluriformes and Cypriniformes, clustered within a single clade. In Cypriniformes order, Cobitidae family and *Raimas bola* clustered away from remaining members of the order. While, *Mystus montanus* clustered along with members of Cypriniformes order and away from remaining members of Siluriformes order. *Canthophrys gongota and Lepidocephalichthys guntea* clustered close to Siluriformes order and distinct from Cypriniformes order. *Monopterus cuchia* of the order Synbranchiformes and *Setipinna phasa* of Clupeiformes clustered close to members of Siluriformes. *Xenentodon cancila* of Beloniformes order clustered within Siluriformes clade. Most of the members of Perciformes order except for *Betta spelndens* formed distinct clade. *Chitala chitala, Notopterus notopterus and Osteoglossum bicirrhosum* of Osteoglossiformes order clustered together. Species of the order Clupeiformes, *Setipinna phasa, Nematalosa nasus and Pellona ditchela* did not form any cohesive unit.

Table 4.16 Summary of orders forming distinct cluster in NJ tree.

# No.	Order	Distinct Cluster in NJ tree
1	Cypriniformes	Single clade with few exceptions
2	Siluriformes	Single clade with few exceptions
3	Perciformes	except *B. spelndens* other species clustered within single
4	Clupeiformes	Species did not form any cohesive unit
5	Osteoglossiformes	All 3 species formed cohesive cluster
6	Mugiliformes	Single species *R.corsula* clustered with Perciformes
7	Beloniformes	Single species *X.cancila* clustered with Siluriformes
8	Cyprinodontiformes	All 4 species formed cohesive cluster
9	Synbranchiformes	All 4 species formed cohesive cluster
10	Tetraodontiformes	Not formed

Chapter 4.4 Development of character profiles for different hierarchical levels of a taxon.

Character profiles for 1307 sequences belonging to three major orders of Indian freshwater fishes were developed. The profiles were developed for different hierarchical levels viz. species, genus and family. CAOS was used to derive diagnostic character traits for full-length *COI* barcodes.

4.4.1 Character profiles for species

The character states for 137 Indian freshwater fish species at 273 nucleotide positions of the *COI* gene region were developed. Species with more than two representative sequences were considered. Table 4.17 shows the representative character states at 17 nucleotide positions of the *COI* gene region for 49 species of Cyprinidae family. The particular nucleotide positions were chosen as representative sequence position of the entire character state of each species. The positions with more than two nucleotides at single site were considered to be uninformative.

Baring four pairs of species, all of the 137 species, revealed a unique character state combination across the entire 513bp length of *COI*. In Figure 4.16 the 4 points that are marked in red represents species pairs that share character states. Conspecific sequences of *Tor putitora* showed variation in 8 positions which were also shared by *Tor macrolepis*. Thus, the two species did not form unique character state combination. *Scizothorax progastus and Scizothorax richardsonii* showed variation in two positions 128th and 350th. Conspecific sequences of *Mystus vittatus* showed variation in 14 positions, which was shared by *Mystus horai*. The two species together shared private character attributes that varied from the remaining species thus forming an entity and supporting findings of other studies whereby the two species have been considered as synonymous species. Similarly, *Mystus tengara* and *Neotropius atherinoides* differed by 14 diagnostic characters of which 2 positions were shared between the two species.

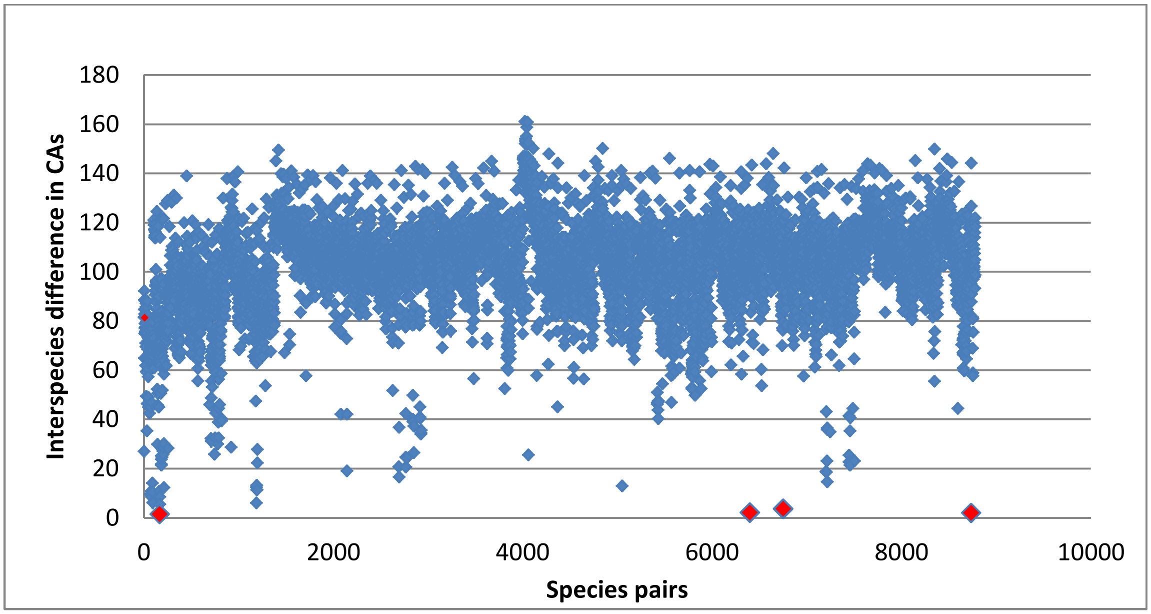

Figure 4.16 Pairwise interspecies difference in character attributes between 137 species. Each point represents difference in character states between species pairs. Points that are marked in red represent species pairs that share character states. X axis represents species pairs and Y axis represents number of character difference.

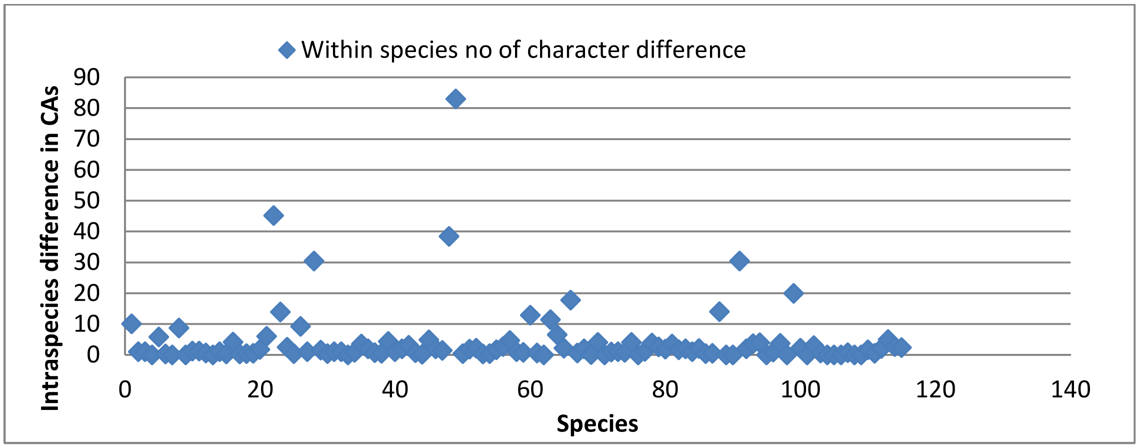

Figure 4.17 Average intraspecies differences in character attributes within 116 species.

Each point represents average character differences within a species. X axis represents species and Y axis represents number of character difference.

Number of character difference between conspecific sequences is shown in Figure 4.17. 21 species did not have sufficient representative barcode sequences to estimate the character states efficiently. Hence character attributes for 116 species have been used in the figure. 19 species *Puntius chalakkudiensis, Puntius chelynoides, Tor khudree, Lepidocephalichthys guntea, Aspidoparia morar, Balitora brucei, Raimas bola, Rita rita, Sperata aor, Scizothorax richardsonii, Mystus oculatus, Horabagrus brachysoma, Barbonymus altus, Puntius tetrazona, Sisor rabdophorus, Betta splendens, Colisa laila* exhibited only pure diagnostic attributes with other species while conspecific sequences showed total conservation across entire length.

Few species *Labeo rohita, Eutropiichthys vacha, Channa aurantimaculata, Carassius auratus, Bagarius bagarius, Glyptothorax garhwali, Opsarius bakeri, Channa barca, Channa stewartii, Clarias gariepinus, Glyptothorax brevipinnis, Glyptothorax granulus, Osteobrama belangeri, Tor mussullah, Danio rerio, Channa bleheri, Scistura beavani, Barilius vagra, Homaloptera montana, Labeo fimbriatus, Sperata seenghala, Cyprinus carpio carpio, Eutropiichthys murius, Mystus malabaricus, Corydoras aeneus, Glyptothorax ventrolineatus, Mystus bleekeri, Puntius ticto, Ompok pabda* showed presence of only few pure diagnostic attributes with 2-3 variable sites between conspecific sequences. Two conspecific sequences of *Glossogobius gurius* (FJ59498, JQ713857) showed variations in 83 nucleotide positions. Similarly, 11 conspecific sequences of *Channa marulius* showed 90 divergent nucleotide positions and 5 conspecific sequences of *Garra hughi* showed 82 divergent positions. *Puntius denisonii and Clarias batrachus* each showed divergence of 79 and 71 sequences respectively. 13 conspecific sequences of *Labeo calbasu* showed divergence across 62 nucleotide positions. Three groups of species (ANGBF7332-12, ANGBF7369-12, ANGBF7370-12, ANGBF7371-12, ANGBF7372-12, ANGBF7373-12; CYTC3829-12, CYTC5334-12 and DBFN047-11, DBFN068-11, GBGCA2520-13, GBGCA2521-13, GBGCA2522-13) from three different geographical location showed variation in character states.

Table 4.17 Character profiles of *COI* barcode for species of Cyprinidae family at 17 nucleotide positions.

Species		2	3	5	6	8	11	14	17	20	21	23	25	26	29	32	33
C.auratus|12	C	A	C	T	C	A	T	T	C	A	A	T	A	C	T	A	A
C.catla|47	C	G	C	T	C	A	T	T	C	A	A	T	A	T	T	T	A
P.chelynoides|3	C	G	C	T	C	A	C	T	C	A	A	T	A	C	T	C	A
P.sophore|17	C	A	C	C	C	A	T	T	T	A	A	T	A	T	T	T	A
B.gonionotus|1	C	A	C	T	C	A	C	T	T	A	A	T	A	T	T	T	A
L.bata|32	C	A/G	C	T	C	A	C/T	C	C	A	A	T	A	C	T	T	A
L.gonius|4	C	G	C	T	C	A	T	C	C	A/G	A	T	A	T	T	T	A
C.mrigala|50	C	G	C	T	C	A	C	C	C	A	A	T	A	C	T	C	A
T.khudree|12	C	G	C	T	C	A	T	T	C	A	A	T	A	T	T	T	A
T.mussullah|8	C	G	C	T	C	G	T	T	C	A	A	T	A	T	T	T	A
T.putitora|51	C	G	C	T	C	A	T	T	C	A	A	T	A	T	T	T	A
T.tor|18	C	G	C	T	C	A	T	T	C	A	A	T	A	T	T	T	A
C.latius|4	C	A	C	C	C	A	T	T	C	A	A	T	A	T	T	T	A
D.devario|1	T	G	T	T	T	T	T	T	T	T	T	T	A	T	T	T	G
L.boggut|9	C	G	C	T	C	A	C	C	C	A	A	T	A	C/T	T	T	A
T.macrolepis|6	C	G	C	T	C	A	T	T	C	A	A	T	A	T	T	T	A
N.hexastichus|14	C	G	C	T	A/C	A	T	T	C	A	A	T	A	T	T	T	A
N.hexagonolepis|14	C	G	C	T	C	A	T	T	C	A	A	T	A	T	T	T	A
P.filamentosus|3	C	A	C	A	T	A	C	C	C	A	A	T	A	T	C	T	A
P.tambraparniei|1	C	A	C	A	T	A	C	C	C	A	A	T	A	T	C	T	A
P.chalakkudiensis|3	C	A	C	C	T	G	T	C	C	A	A	T	A	T	T	C	A
P.vittatus|7	C	A	C	T	C	A	T	T	T	A	A	C	A	C	T	A	A
P.sarana|15	C	G	C	A	T	A	T	T	C	A	A	T	A	T	C	T	A
R.rasbora|7	C	A	C/T	A	C	A	G	C	C	A	A	T	A	T	T	T	A
C.cirrhosus|1	C	G	C	T	C	A	C	C	C	A	A	T	A	C	T	C	A

B.vagra|5	C	G	C	T	C	T	G	C	C	A	A	C	A	T	T	T	A
H.molitrix|5	C	A	C	T	C	G	C/T	T	C	A	A	T	A	T	C/T	T	A
G.gotyla|8	C	A	C	C	T	A	T	C	T	A	A	C	A	C	C	T	A
E.danricus|7	C	T	C	A	T	A	T	C	C	A	A	C	A	T	T	T	A
B.ariza|4	C	A	C	T	C	A	T	C	C	A	A	T	A	C	T	T	A
A.morar|2	C	A	C	C	C	A	T	C	T	A	A	T	A	T	T	A	A
D.aequipinnatus|1	T	G	T	T	T	T	T	T	T	T	T	T	A	T	T	T	G
D.rerio|11	C	A	C	T	C	T	T	T	T	A	A	T	A	T	T	T	A
R.daniconius|5	C	A	C	A	C	A	A	C	C/T	A	A	T	A	T	T	C/T	A
L.rohita|41	C	G	C	T	C	A	T	T	C	A	A	T	A	T	T	T	A
L.fimbriatus|29	C	A/G	C	T	C	A	T	T	C	A	A	T	A	T	T	T	A
L.calbasu|13	C	G	C	T	C	A/C	T	C/T	C	A	A	T	A	T	C/T	T	A
R.bola|3	C	A	C	C	C	T	G	T	C	G	A	C	T	T	T	A	A
D.malabaricus|11	T	G	T	T	T	T	C	T	T	A/T	A/T	T	A	T	T	T	G
G.hughi|5	C	G/T	C	T	C/T	A	T	T	C	A	A	C/T	A	C/T	C/T	T	A
B.tileo|5	C	A	C	C	C	G	C	C	C	A	A	C	A	T	C	T	A
O.bakeri|5	C	A	C	C	C	G	C	C	C	A	A	C	A	T	T	T	A
B.gatensis|5	C	A	C	T	C	A	C	T	C	A	A	C	A	C	T	T	A
T.malabaricus|4	C	G	C	T	C	G	T	T	C	A	A	C	A	T	T	T	A
O.belangeri|9	C	G	C	T	C	A	T	C	C	A	A	T	A	T	C	C	A
P.ticto|10	C	A	C	C	C	A	T	T	T	A	A	T	A	T	T	A	A
O.cotio|1	C	A	C	C	T	A	G	C	C	A	A	T	T	T	T	T	A
H.nobilis|1	C	A	C	T	C	G	C	T	C	A	A	T	A	T	C	T	A
S.richardsonii|2	C	G	C	T	C	A	T	T	C	A	A	T	A	C	T	A	A
E.bicolor|1	C	A	C	C	T	G	T	C	C	A	A	T	A	T	T	T	A
B.altus|3	C	A	C	T	C	A	C	T	T	A	A	T	A	T	C	A	A
P.tetrazona|3	C	G	C	T	T	A	T	T	T	A	A	T	A	T	T	T	A
E.frenatum|2	C	A	C	T	C	A	T	T	C	A	A	T	A	C	T	T	A

4.4.2 Character profiles for genus

The character states for 64 Indian freshwater fish genera at 327 nucleotide positions of the *COI* gene region were constructed. All the 327 diagnostic positions of each genus contained conserved or two fold degenerate site. Non-significant positions with more than two nucleotides within a genus were not considered as diagnostic characters. Positions exhibiting more than one nucleotide within a genus were marked by degenerate nucleotides. Most of the genus immediately revealed a unique combination of character states at 327 positions. Figure 4.18 shows clear intergenus demarcation based on number of differences of characters between different genera studied. Within a genus an average of 40 diagnostic attributes were exhibited by the congeneric species. Within this range, few genus pairs that showed exceptions were Mystus and Neotropius, Tor and Neolissochilus, and Catla and Labeo. For the remaining 22 genus few diagnostic characters were identified. The genus Puntius did not show any unique diagnostic character when compared to the genera Barbonymous, Garra. However, it varied in 3 pure diagnostic character states with each of the genus Carassius, Catla.

The distance based on number of difference of nucleotide varied from 0 to 88 within genus (Figure 4.19). The genera Pseudeutropius, Pangasianodon, Batasio, Andinoacara, Hemichromis, Nandus had single representative species and were not used to calculate within genus character difference. Most of the genera showed character difference of about 30 nucleotides. Lepidocephalichthys, Aspidoparia, Balitora, Raiamas, Rita, Sisor, Betta showed complete conserved characters within the respective genera. The genera Carassius, Bagarius, Opsarius, Danio, Cyprinus, Homaloptera, Corydoras, Cirrhinus, Botia, Ailia, Clupisoma, Schizothorax showed few variations within the representative species. However, some genera like Schistura, Colisa, Garra, Epalzeorhynchos, Ompok, Clarias, Amblyceps, Barilius, Puntius, Mystus, Channa, Glossogobius, and Pseudotropius showed wide variation within the representative species. In Table 4.18, a representative character profile of *COI* barcode for genera of Cyprinidae family at 20 nucleotide positions is shown.

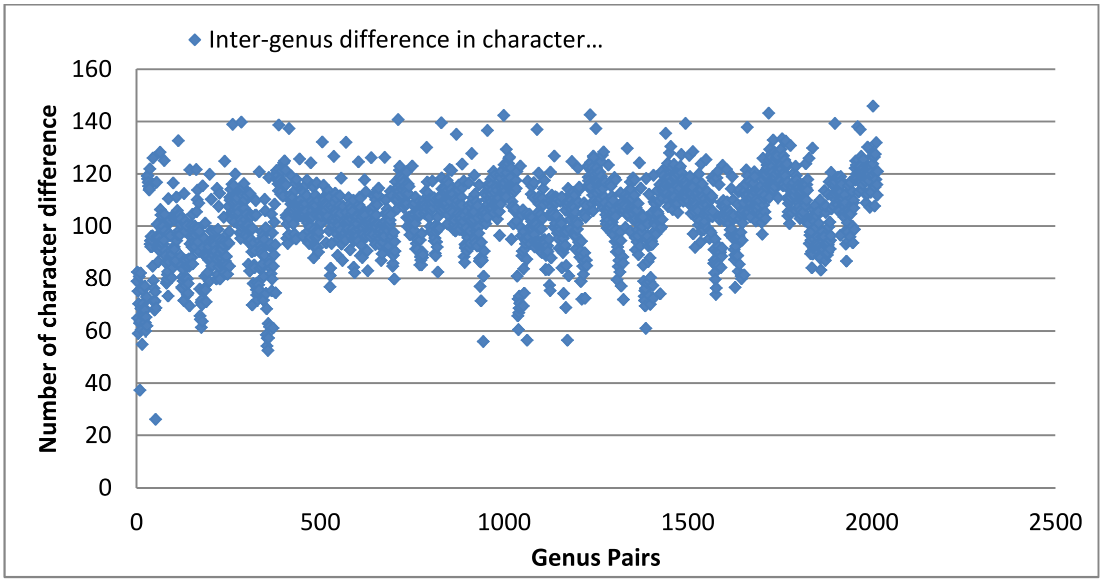

Figure 4.18 Pairwise intergenus difference in character attributes between 64 genera.

Each point represents difference in character states between genus pairs. X axis represents genus pairs and Y axis represents number of character difference.

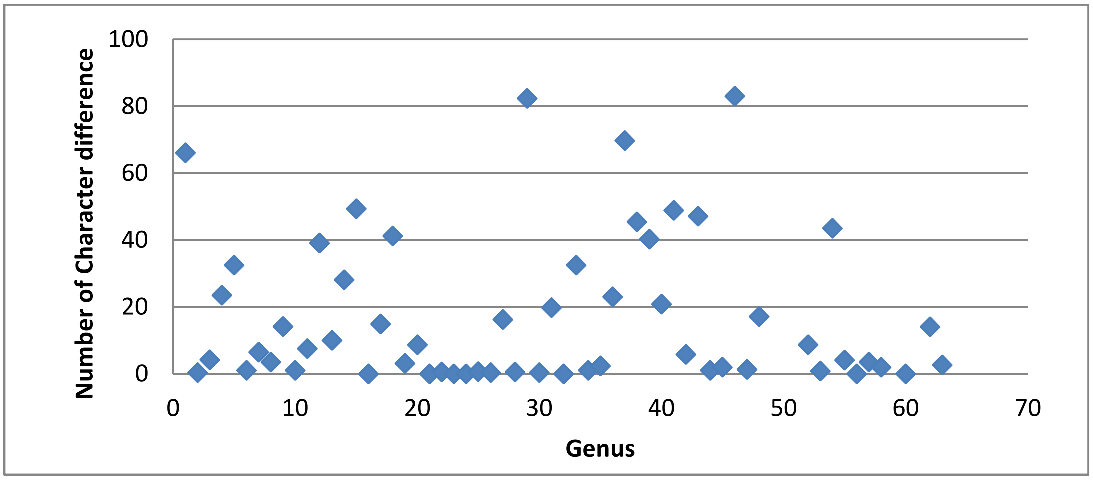

Figure 4.19 Average congeneric differences in character attributes within 64 genera.

Each point represents number of character state difference between congeneric species pairs. X axis represents genus and Y axis represents number of character difference.

Table 4.18 Character Profiles of *COI* barcode for genera of Cyprinidae family at 20 nucleotide positions.

Genus	1	2	3	5	6	7	10	12	14	16	18	19	20	21	23	24	25	27	28	30
Puntius	C	A/G	C	A/T	C/T	A/G	A/T	C/T	C/T	A	C/T	A	C/T	C/T	A/T	A	C/T	C/T	T	C/T
Carassius	C	A	C	T	C	A	T	T	C	A	T	A	C	T	A	A	T	T	C	C
Catla	C	G	C	T	C	A	T	T	C	A	T	A	T	T	T	A	T	A/G	T	C/T
Barbonymus	C	A	C	T	C	A	C	T	T	A	T	A	T	C/T	A/T	A	C	T	C/T	C
Labeo	C	A/G	C	T	C	A/C	C/T	C/T	C	A/G	T	A	C/T	C/T	T	A	C/T	A/T	T	C/T
Cirrhinus	C	G	C	T	C	A	C	C	C	A	T	A	C	T	C	A	C	C	T	T
Tor	C	G	C	T	C	A/G	T	T	C	A	C/T	A	T	T	T	A	C	C	T	T
Crossocheilus	C	A	C	C	C	A	T	T	C	A	T	A	T	T	T	A	C	T	T	T
Devario	T	G	T	T	T	T	C/T	T	T	A/T	T	A	T	T	T	G	A	T	T	C
Neolissochilus	C	G	C	T	A/C	A	T	T	C	A	T	A	T	T	T	A	C	C	T	T
Rasbora	C	A	C/T	A	C	A	A/G	C	C/T	A	T	A	T	T	C/T	A	T	A	T	T
Barilius	C	A/G	C	C/T	C	A/T	G/C	C/T	C	A	C	A	C/T	C/T	T	A	T	T	C/T	C
Hypophthalmichthys	C	A	C	T	C	G	C/T	T	C	A	T	A	T	C/T	T	A	T	T	T	C
Garra	C	A/T	C	C/T	C/T	A	T	C/T	C/T	A	C/T	A	C/T	C/T	T	A	C	T	C/T	C
Esomus	C	T	C	A	T	A	T	C	C	A	C	A	T	T	T	A	T	T	A	C/T
Bangana	C	A	C	T	C	A	T	C	C	A	T	A	C	T	T	A	T	T	T	C
Aspidoparia	C	A	C	C	C	A	T	C	T	A	T	A	T	T	A	A	C	C	T	C
Danio	C	A	C	T	C	T	T	T	T	A	C/T	A	T	T	T	A	T	T	T	C
Raiamas	C	A	C	C	C	T	G	T	C	G	C	T	T	T	A	A	C	T	A	A
Opsarius	C	A	C	C	C	G	C	C	C	A	C	A	T	T	T	A	C	T	T	C
Osteobrama	C	A/G	C	C/T	C/T	A	G/T	C	C	A	T	A/T	T	C/T	C/T	A	T	T	C	C
Cyprinus	C	G	C	T	C	A	T	T	C	A	T	A	T	C	A/T	A	C	C	T	C
Schizothorax	C	G	C	T	C	A	T	T	C	A	T	A	C	T	A	A	C	T	T	T
Epalzeorhynchos	C	A	C	C/T	C/T	A/G	T	C/T	C	A	T	A	C/T	T	A/T	A/T	G/C	T	T	C

4.4.3 Character profiles for family

The character states for 21 Indian freshwater fish families across 513 basepair for 327 positions were generated. Non-significant positions with more than two nucleotides within a family were not considered as diagnostic characters. All the diagnostic positions of each family contained conserved or two fold degenerate sites.

Figure 4.20 shows distribution of number of diagnostic attributes within each of the studied family. The characters are calculated for each pair of genus within a family. The figure shows that the diagnostic attributes for families Cyprinidae, Bagridae and Sisoridae varied from lowest range (0-20) to a highest of approximately 120. While, the families, Balitoridae, Osphronemidae varied from a lowest range of 40 and Schilbeidae from a lowest of 20 to the maximum range of 120 diagnostic variations per genus pairs. Cobitidae family with two representative species showed highest variation in diagnostic sites within species. The two species *Botia almorhae* and *Lepidocephalichthys guntea* varied in 103 diagnostic positions. Siluridae with 122 sequences from 6 representative species showed no variation in pure diagnostic characters.

In the family Cyprinidae, 564 sequences belonging to 25 genera and 53 species were considered for generating character states. The character profile for Cyprinidae family over 22 nucleotide positions is demonstrated in Table 4.19. Within the family overall 233 positions were found to be diagnostic. However, most members of Cyprinidae family showed a pairwise distribution of diagnostic characters within a range of 80 variations. Most of the members of same family shared many diagnostic sites while differed significantly from members of other families. However, Puntius genus belonging to Cyprinidae family of the order Cypriniformes showed lack of distinct pure diagnostic characters with respect to Mystus genus of the Bagridae family belonging to the order Siluriformes. Family specific diagnostic characters were not confirmed for the families Erethistidae, Loricariidae, Channidae, Amblycipitidae, Clariidae, Gobiidae, Heteropneustidae, Pangasiidae, Callichthyidae, Latidae, Erethistidae, Loricariidae as these families were represented by one or more species of single genus.

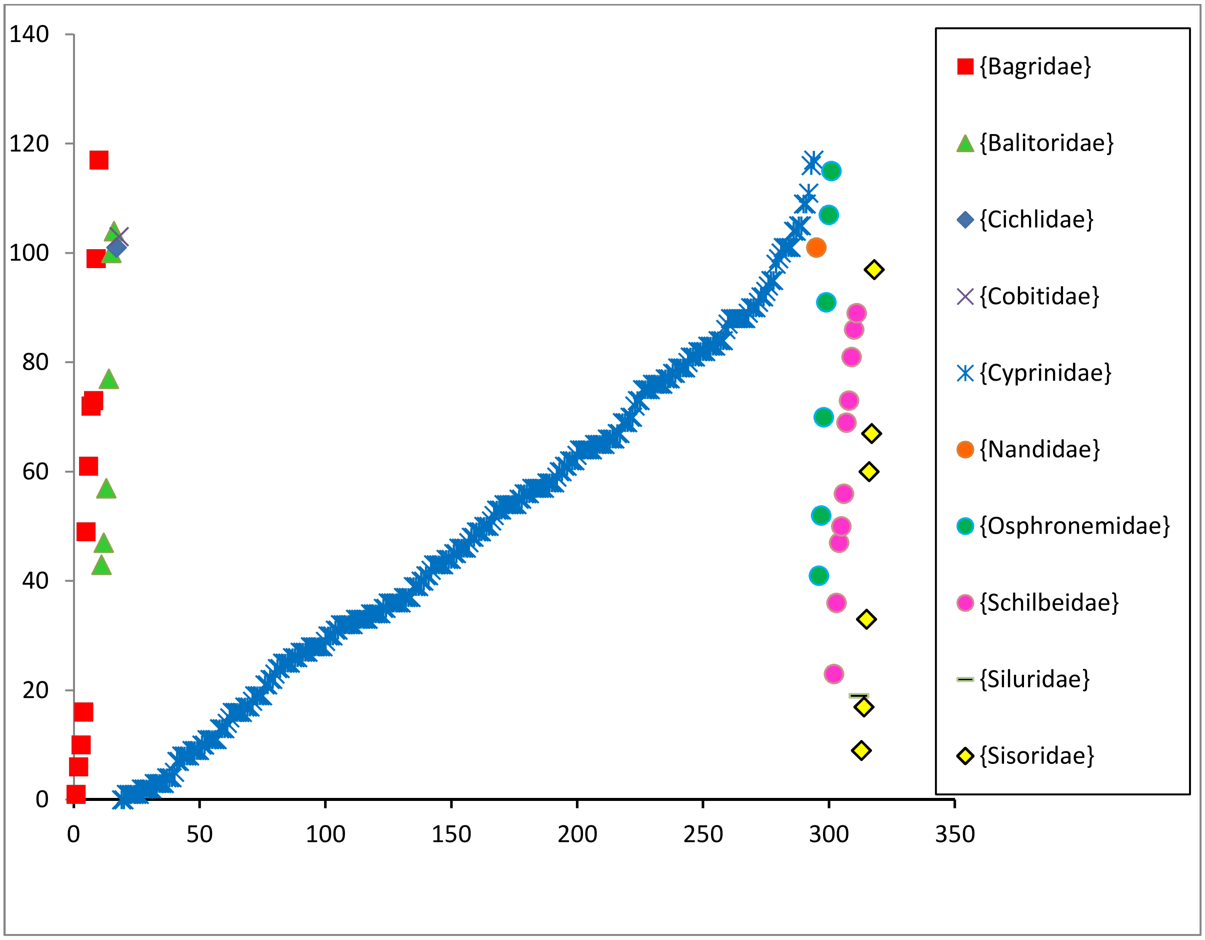

Figure 4.20 Distribution of number of diagnostic characters within families.

Different families are marked by different colored data points as described in the figure legend. Each point represents number of character difference between each pair of genus within a family. Families having barcodes for more than one representative genus and the genus having more than one representative species were used for the analysis.

Table 4.19 Character profiles of *COI* barcode for families of Cypriniformes order at 22 nucleotide positions.

Family	**1**	**2**	**3**	**4**	**5**	**6**	**8**	**9**	**10**	**11**	**12**	**14**	**15**	**17**	**18**	**20**	**21**	**23**	**25**	**26**
Cyprinidae|564	C/T	A/T	C/T	T	A/T	A/T	A/T	A/G	G	A/T	A/G	C/T	G	C/T	C	A/T	A/T	C/T	A/T	C/T
Cobitidae|10	C	A/C	C	T	A/T	C	A	G	G	A/T	G	T	G	C/T	C	A	A	T	A	C
Balitoridae|29	C	A/C	C	T	C/T	C	A/T	G	G	C/T	G	C/T	G	C/T	C	A	A	T	A	C/T
Channidae|84	C	G/C	G/C	A/T	A/T	C/T	G/C	G/C	G/C	A/T	A/G	C/T	G/C	C/T	A/C	A/G	A	C/T	A	T
Sisoridae|225	C	C/T	A/T	T	A/T	C/T	G/C	G	G	A/C	G	T	G	C/T	C	A/G	A	C/T	A	C/T
Bagridae|59	C	A/T	C/T	T	A/T	C/T	A/T	G	G	G/C	G	C/T	G	C/T	C	A/G	A	T	A	C/T
Schilbeidae|34	C	A/T	C	T	A/T	C/T	A/G	G	G	C/T	G	T	G	C/T	C	A/G	A	T	A	C/T
Siluridae|122	C	C	C	T	T	C	A	G	G	C	G	C/T	G	C	C	A	A	T	A	C/T
Osphronemidae|14	C	C/T	C	T	T	C/T	G/C	G	G	A/T	A/G	C/T	G	C/T	C	A/G	A	C/T	A	C/T
Amblycipitidae|4	C	C/T	C	T	C/T	C/T	A	G	G	A/C	G	C/T	G	C	C	A	A	T	A	T
Nandidae|5	C	C	C	T	C	C	G	G	G	G/C	G	T	G	C	C	A	A	C/T	A	T
Clariidae|29	C	C/T	C	T	T	C/T	A	G	G	A	G	C/T	G	C	C	G	A	C/T	A	T
Gobiidae|2	C	C	C	T	A/G	C	A	G	G	C	G	C/T	G	C	C	A	A	T	A	T
Heteropneustidae|15	C	T	C	T	A	C	A/G	G	G	T	G	T	G	C	C	A	A	T	A	T
Pangasiidae|1	C	C	C	T	T	C	A	G	G	C	G	C	G	C	C	A	A	T	A	T
Callichthyidae|5	C	T	C	T	T	C	G	G	G	C	G	C	G	T	C	A	A	T	A	C
Cyprininae|6	C	G	C	T	T	C	A	G	G	T	G	T	G	C	C	A	A	T	A	T
Latidae|89	C	A	C	T	C	C	G	G	G	G	G	C/T	G	C	C	A	A	C	A	T
Erethistidae|2	C	C	C	T	C	C	A	G	G	C	G	T	G	T	C	A	A	C	A	T
Cichlidae|2	C	A/T	C	T	C	C	T	G	G	A	G	C	G	C	C	A	A	T	A	T
Loricariidae|3	C	C	C	T	A/T	C	A	G	G	T	G	T	G	C	C	A	A	T	A	T

Chapter 4.5 Development of species specific barcode motif.

4.5.1 Retrieval of highly informative region

Overall 1307 sequences belonging to the 3 major orders (Cypriniformes, Siluriformes and Perciformes) of Indian freshwater fishes were used for the analysis of *COI* barcodes. Intraorder transition and transversion substitutions were calculated for each of the 654 base pair positions for the three orders separately. Figure 4.21 shows a plot of substitution pattern of the nucleotides across the full length barcodes. As seen in the figure, transitions and transversions follow a distinct pattern of distribution in all the three orders of fishes. Overall, transitions (represented by +1) are dominant over transversions (represented by -1). Transversion biased sites are found to be scattered randomly throughout the full-length barcode. The graph points out the presence of a continuous stretch of transversion dominant segment lying between base pair positions BP_{261} –BP_{432} (position calculated from BLAST SEARCH against full length *COI*), in all the three studied orders of fishes. The 171 base pair long transversion 'hotspot' was selected as a potential region for selecting a minibarcode.

4.5.2 Primer design and validation

The sequences flanking the "transversion hotspot" loci in the *COI* barcode were found to be conserved within a particular species and across species within a genus. A degenerate primer was designed using these neighboring upstream and downstream regions. The primer pair, forward: 5' - CNSGNATRAAYARYAWRAGYTTYTG - 3'(BP_{233} –BP_{257}) and reverse: 5'-RTTVANDRWNGTNGYDATRAARTTRATNG -3' (BP_{431} –BP_{459}) satisfied all the required primer properties. An *in-silico* primer test confirmed the potential of the proposed primers in amplifying sequences from all the three studied orders of fishes. Representative sequences of all the 160 species generated PCR product of 227bp length in the target region. Finally, a 154bp nucleotide segment lying between BP_{267} –BP_{421} was selected, thus leaving sufficient upstream and downstream sequence to avoid noisy peaks in the target region.

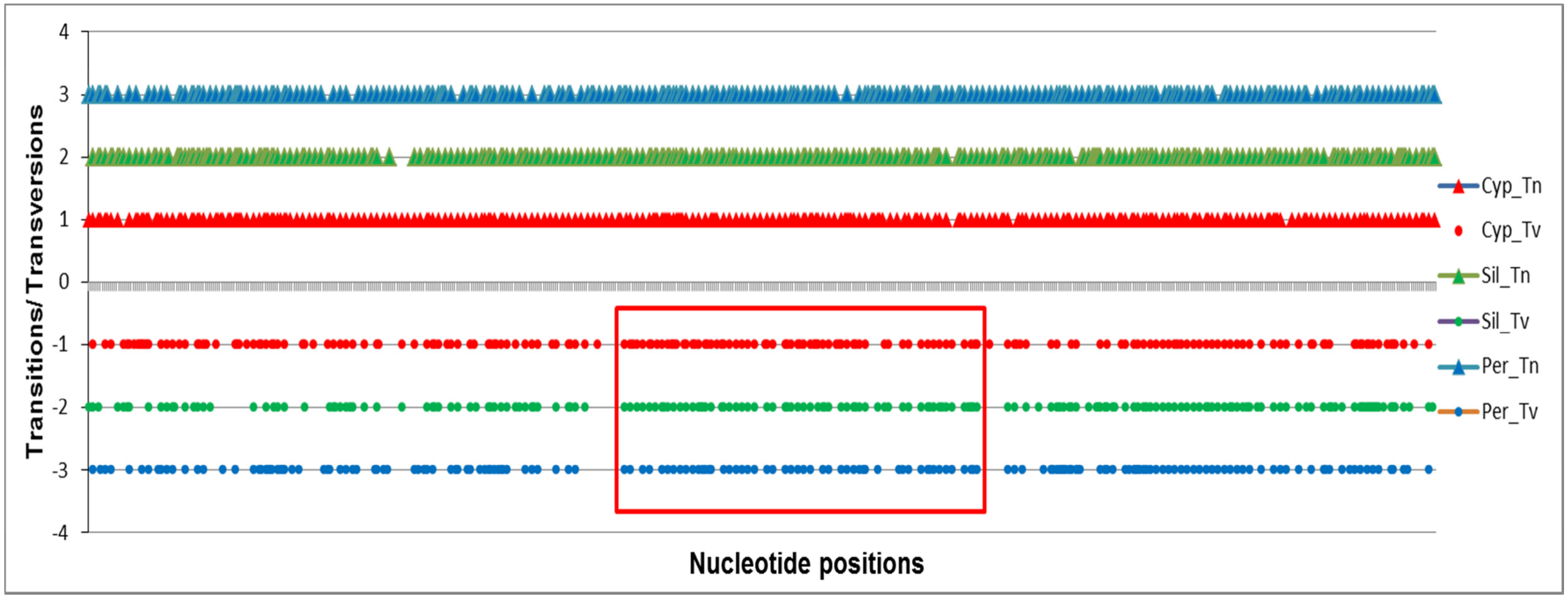

Figure 4.21 Distribution patterns of transitions and transversion across full-length barcode (654bp) of 1307 sequences.

Nucleotide positions showing transition substitution are represented by three different positive values for the three orders (1: Cypriniformes, 2: Siluriformes and 3: Perciformes). Positions with transversion substitution are represented by their corresponding negative values (-1: Cypriniformes, -2: Siluriformes and -3: Perciformes). Three different colors represent species of three orders, Red for Cypriniformes, Green for Siluriformes and Blue for Perciformes. Triangles represent transitions while circles represent transversions. The box highlights the 154bp transversion dominant segment.

PRIMER DETAILS

233 CCCGAATAAATAACATAAGCTTCTG 257

Primer sequence: CCCGAATAAATAACATAAGCTTCTG
Sequence length: 25
Base counts: G=3; A=10; T=6; C=6; other=0;
GC content (%): 36.00
Basic Tm (degrees C): 53
Salt adjusted Tm (degrees C): 48
Nearest neighbor Tm (degrees C): 60.86

431 CCATCAACTTTATTACAACAACTATTAAC 459

RevCom = GTTAATAGTTGTTGTAATAAAGTTGATGG

Primer sequence: GTTAATAGTTGTTGTAATAAAGTTGATGG
Sequence length: 29
Base counts: G=8; A=9; T=12; C=0; other=0;
GC content (%): 27.59
Basic Tm (degrees C): 53
Salt adjusted Tm (degrees C): 48
Nearest neighbor Tm (degrees C): 60.78

Sl. No.	Name	Sequence	Mer	A + T	G + C	Others	Basic T_m (°C)	Nearest neighbor T_m (°C)
1	Fwd_P	CCCGAATAAATAACATAAGCTTCTG	25	16	9	0	53	60.86
2	Rev_P	GTTAATAGTTGTTGTAATAAAGTTGATGG	29	21	8	0	53	60.78

4.5.3 Validation of the proposed segment for use in species delimitation

This 154bp segment of transversion hotspot region was retrieved for all the species of the three orders. A NJ tree (Appendix 4), based on the K2P divergences of the nucleotides of the minibarcode region, showed that in most cases conspecific sequences clustered together and were distinct from the remaining species. Comparison of the two NJ trees, constructed from full length (600bp) (Appendix 3) and minibarcode (154bp) (Appendix 4) *COI*, revealed similar clustering pattern for most of the sequences. Sequences of some species like *Channa orientalis, Garra hughi* showed deep-split in both the NJ trees. Sequences of some differently named species (*Mystus tengara* and *Neotropius atherinoides, Mystus vittatus and Mystus horai, Heteropneustes microps and Heteropneustes fossilis, Channa bleheri and Channa barca*) clustered together in both the trees. While, some sequences of same-named species, (*Clarias batrachus, Channa marulius,*) were clustered distinctly in both the trees. Thus, both full length and minibarcode segment yielded similar tree topology. Sequences, whose species status was not confirmed in the NJ tree, were not included for building the motifs.

4.5.4 Design of barcode-motifs

The 154bp transversion rich segment showed higher interspecies variation as compared to the complete barcode region (Figure 4.22). However, both the minibarcode and the full-length barcode showed similar trend of intraspecies variation with the full-length barcode exhibiting slightly higher range (Figure 4.23). Most of the intraspecies variable sites were found in the third codon position and showed two-fold degeneracy. Then species-specific 154bp barcode motifs were designed using nucleotide segment from BP_{267} –BP_{421} of *COI* barcode for 109 species using a C++ program (Appendix 5). The intraspecies variable sites were represented by degenerate nucleotides.

4.5.5 Motif specificity validation

The ability of the barcode motifs in assigning sequences to their respective species was cross checked using a program MOTIF MATCH that checked each motif against the full-length barcodes of all the species for an identical match. This program was employed in studying the specificity of the developed motifs by matching it against sequences of the 3 orders of Indian freshwater fishes. The program checked whether the motifs matched only with their respective species or showed false positive match with other species.

The results revealed that the nucleotide segment from BP_{267} – BP_{421} of all the sequences matched with their respective motifs only. The 109 barcode motifs showed an average of 20 variable sites between congeneric species. Motifs for all the species baring *Cyprinus carpio* and *Cyprinus carpio carpio* were non-identical. *Cyprinus carpio carpio* is a junior synonym of the former and thus an identical motif is justified. A second crosscheck was performed using the 'Find motif' program in MEGA 5.1. This search also revealed similar results and confirmed that the motifs aligned with their respective species.

Further, the 'Find motif' program in MEGA 5.1 revealed that the sequences aligned with a template COI sequence in the expected region, that is between BP_{267} – BP_{421} and the motifs were highly specific for each species. The motif showed alignment only in the expected region of their respective species.

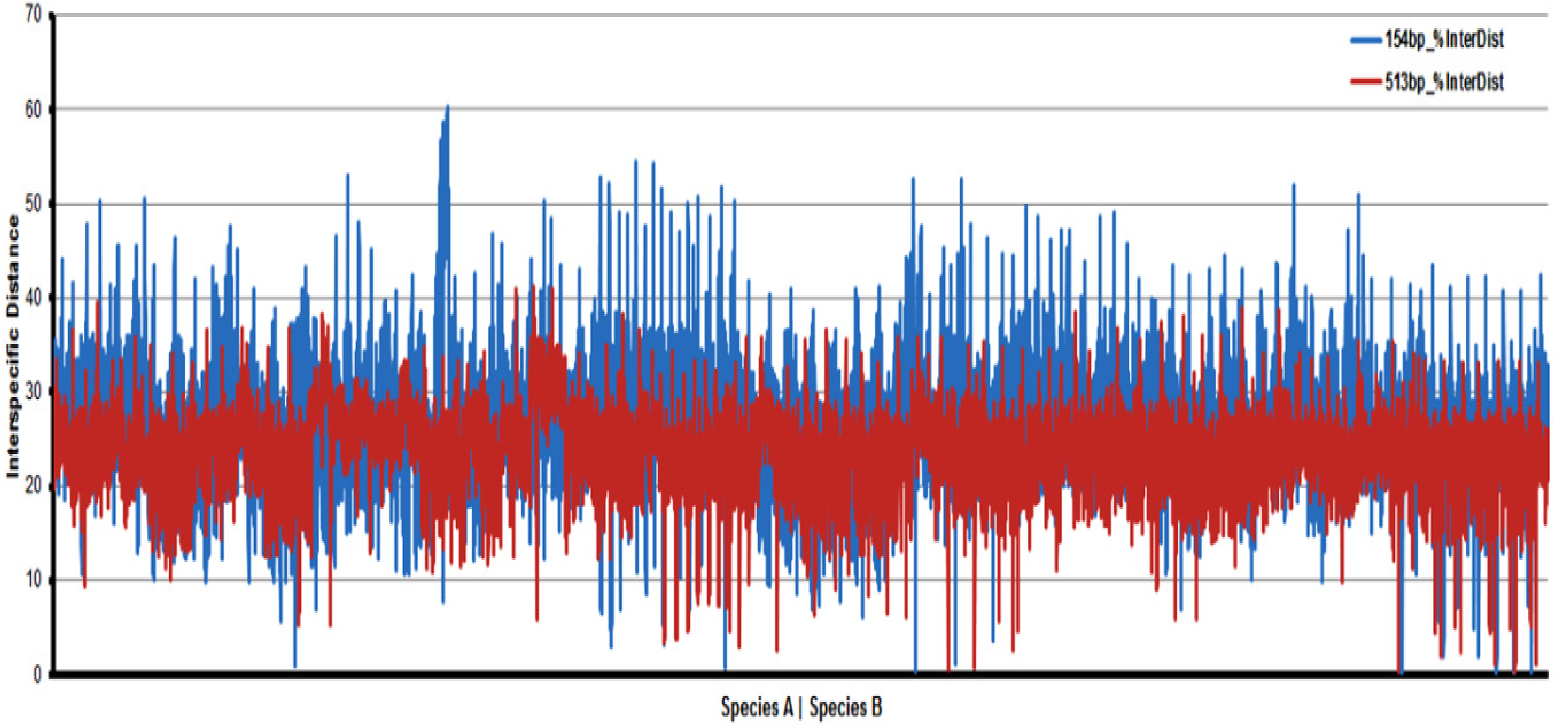

Figure 4.22 Comparison of interspecies distance between full length *COI* barcode (513bp) and minibarcode (154bp). Each bar represents interspecific distance between species pairs. The blue line represents interspecific distance between congeneric species based on 154bp transversion rich segment of *COI* barcode. The blue line represents interspecific distance between congeneric species based on full recoverable length (513bp) *COI* barcode.

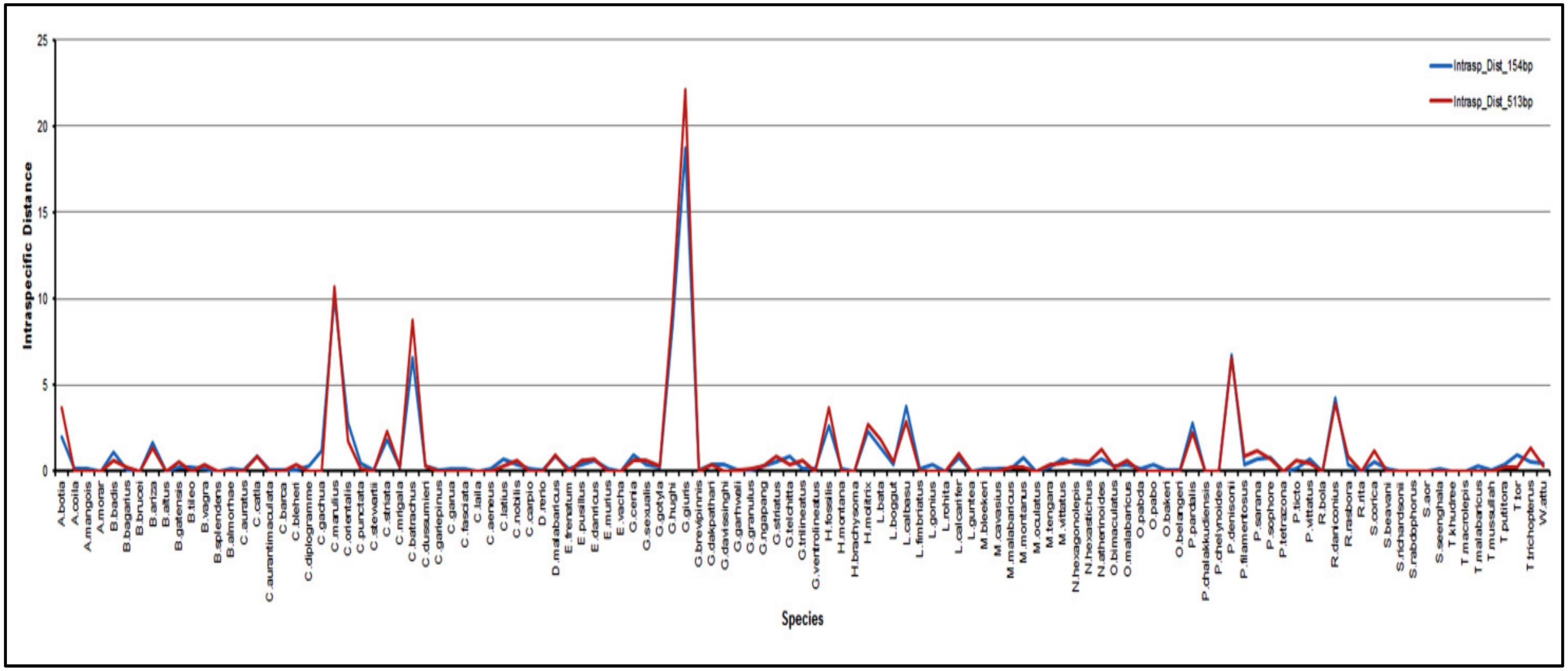

Figure 4.23 Comparison of intraspecies distance between full length *COI* barcode (513bp) and minibarcode (154bp).

Each line represents distance between conspecific sequences. The blue line represents intraspecific distance between conspecific sequences based on 154bp transversion rich segment of *COI* barcode. The blue line represents interspecific distance between conspecific sequences based on full recoverable length (513bp) *COI* barcode.

4.5.6 Validation of the barcode motifs with global Cypriniformes data

Further, to check whether the motifs developed from Indian fish-sequences were capable of correctly identifying fish-species from different geographical locations; the motifs were crosschecked against the global Cypriniformes data. 24 species of Cypriniformes order had representative barcodes in the database from more than one countries including India. For those species, 154bp motifs were developed with the Indian sequences using the MOTIF-BUILD program. The motifs were then checked against the full-length barcodes of 393 sequences of Cypriniformes species retrieved from the global database using the MOTIF-MATCH program.

The results revealed, (Table 4.20) 18 out of the 24 species showed straightforward match with their respective species. Among the remaining 6 species, few sequences within each species showed anomaly with respect to the barcode motif. *Aspidoparia morar* was represented by 4 *COI* barcode sequences with two sequences each from India and Japan. The sequences from Japan showed 28 variable sites with respect to the Indian sequences.

Similarly, motifs for species like *Devario aequipinnatus* and *Puntius conchonius* could not be accurately derived due lack of adequate number of correct barcodes. Sequences of the three species *Puntius ticto, Rasbora daniconius, Rasbora rasbora*, from different geographical locations, showed unexpectedly high deviations from the motif derived from Indian sequences. Indian *Puntius ticto and Rasbora daniconius* sequences varied in 30 and 12 nucleotide positions respectively from rest of the global conspecific sequence. *Rasbora rasbora,* submitted from Raffles Museum of Biodiversity Research (exact sequence location unclear), varied from the Indian conspecific sequences in 19 nucleotide positions though no change in amino acid was observed.

Table 4.20 Motif verification of Cypriniformes species developed from Indian sequences against representative barcodes from other parts of the World present in BOLD.

Species	Total Species	Species [Indian]	World location info	Motif Match Status
Aspidoparia morar	4	2	Japan	MISMATCH
Barbonymus altus	9	4	Laos,Vietnam,Cambodia, India4	MATCH
Carassius auratus	76	16	Australia 1,Canada 6,China 2,India 15,Japan 1,Korea 17,Philippines 5,Singapore 1,Turkey 2,USA 1	MATCH
Danio rerio	30	11	Canada 2,India 11,Singapore 2,UK, Singapore or New Zealand 9,Unspecified 6	MATCH
Devario aequipinnatus	7	2	India 2, unspecified UK/Singapore/New Zealand 5	MISMATCH
Devario malabaricus	19	12	India 12, unspecified UK/Singapore/New Zealand 7	MATCH
Garra gotyla	15	8	India 8, unspecified 5,UK/Singapore/New Zealand 2	MATCH
Labeo bata	41	40	India 40, japan 2	MATCH
Labeo rohita	63	40	India 40,Unspecified 17,Thailand 5,Laos 1	MATCH
Puntius denisonii	26	18	India 18, unspecified, UK/ Singapore/ New Zealand 6, Singapore 2	MATCH
Puntius filamentosus	11	3	India 3, unspecified, UK/Singapore/New Zealand 8	MATCH
Puntius tambraparniei	7	1	India 1, unspecified, UK/Singapore/8	MATCH
Puntius tetrazona	12	3	India 3, unspecified, UK/Singapore/New Zealand 5, Singapore 2	MATCH

Puntius ticto	16	11	India 11, unspecified, UK/Singapore/New Zealand 3, Thailand 3	MISMATCH
Puntius vittatus	7	7	India 7, unspecified, UK/Singapore/New Zealand 6	MATCH
Rasbora daniconius	9	5	India 5, London 1,japan 2,USA 1	MISMATCH
Rasbora rasbora	11	8	India 8, unspecified, UK/Singapore/New Zealand 3	MISMATCH
Bangana ariza	5	4	India 4,Nepal 1	MATCH
Cyprinus carpio carpio	15	7	India 7,Philippines 5,China 3	MATCH
Epalzeorhynchos bicolor	7	2	India 2, Singapore 2, unspecified, UK/Singapore/New Zealand 2, unspecified 2	MATCH
Epalzeorhynchos frenatum	4	2	India 2, unspecified, UK/Singapore/New Zealand 2	MATCH
Hypophthalmichthys molitrix	12	5	India 5,Brazil 3, China 4	MATCH
Puntius chalakkudiensis	9	5	India 5, unspecified, UK/Singapore/New Zealand 4	MATCH
Puntius conchonius	9	2	India 2, unspecified, UK/Singapore/New Zealand 7	MISMATCH