DNA barcodes of some threatened freshwater indigenous fishes in India

Subhasree Sengupta1,2 and Sumit Homechandhuri1,*

1Aquatic Bioresource Research Laboratory, Department of Zoology, University of Calcutta, 35 Ballygunge Circular Road, Kolkata 700 019, India
2Department of Zoology, Bethune College, 181, Bidhan Sarani, Kolkata 700 006, India

This study was directed to identify early life-history stages of eight culturable species currently under threat through DNA barcodes. A total of 22 sequences of cytochrome c oxidase I gene were generated from eight species using multiple specimens producing 655 nucleotide base pairs per taxon. The neighbour-joining tree showed three major clusters for three different orders. Similar species were clustered under the same nodes and dissimilar species were clustered under separate nodes signifying species specificity. This can be utilized for accurate identification of threatened fish species and in turn assist in conservation of fish stock either inside freshwater protected areas or in the river systems.

Keywords: Barcoding, freshwater fish, life-history stages, neighbour-joining tree.

INDIA is endowed with a rich fish genetic biodiversity (2200 fish species) and ranks ninth in terms of freshwater mega biodiversity1,2. The small, shallow rivers provide refuge to a large number of varied fish population and are also undergoing major habitat alterations due to anthropogenic stresses. The Mahananda–Tangon–Punarbhaha river system (MTPRS) is one such shallow interconnected river system in Malda district, West Bengal, India, harbouring a rich fish biodiversity throughout the year. Out of this rich fish biota, eight most economically important, culturable species were selected for the present study. They were Chitala chitala Hamilton, 1822; Neopterus notopterus Pallas, 1769; Ompok bimaculatus Hamilton, 1822; Eutrophichthys vacha Hamilton, 1822; Alia colla Hamilton, 1822; Neotropius atherinoides Bloch, 1794; Pangasius pangasius Hamilton, 1822 and Gudasia chapra Hamilton, 1822. The populations of these fish fauna are presently in decline3 and require efforts towards conservation through culture or introduction in fish sanctuary. Propagation of stock in controlled environment requires pure seed collected from nature and hence, precise identification of early life-history stages of these fishes is important.

Appropriate identification tool for early life-history stages is essential to confirm purity of seed for introduction

*For correspondence. (e-mail: sumithomec@yahoo.com)
and culture in some protected habitat since: (i) the selected fish species have a similar breeding ground, with monsoon being the breeding season; (ii) artificial propagation research efforts and limited culture activity are dependent upon fish stocked from natural collections; (iii) there is extermination of juveniles of numerous non-target fishes due to lack of proper identification and (iv) there exists differential market potential of various fishes.

Even when two species are almost indistinguishable morphologically, they are likely to be easily distinguished genetically with the help of molecular markers. The cytochrome c oxidase I (COI) region in the mitochondrial genome was demonstrated as an appropriate type-I molecular marker for discriminating closely related species across diverse animal phyla, and this was used for marine and freshwater fishes. However, it is inappropriate to assume that molecular markers can provide the final answer to species identification; they are instead an additional marker system. In this context, the present study was directed to formulate precise identification manual for threatened species through introspecting a small segment of the genome, referred to as genetic ‘barcodes’ and provides information about the COI sequence data of the selected fish species.

The fishes were collected from MTPRS near the Indo-Bangladesh border. The depth of the river varies from a maximum of 3–3.5 m during monsoon, to a minimum of less than 1 m during summer with a silty-clayey
The basin consists of recent alluvium soil type flanked by arable land, especially Graminoid and scrub vegetation, with sporadic rural settlements. Throughout the catchment, the stream waters are heavily used for cultivation of crops (Figure 1).

Sampling was done at 20 sites along MTPRS (24°57’N, 88°20’E) on the Indian side with the help of local fishermen. The specimens were retrieved from the net and identified morphologically to the lowest taxonomic level following standard references. All species names adhere to Fishbase. Approximately 100 mg of white muscle tissue from 2 to 5 individuals of each species was preserved in 95% ethanol and brought to the laboratory under chilled condition. The samples were stored at −80°C until further use. Specimen details and GenBank accession numbers are shown in Table 1.

DNA was isolated and its quality was checked in agarose gel. The concentration was estimated using a UV spectrophotometer. The DNA was diluted to a final concentration of 100 ng/μl.

The COI gene was amplified in 50 μl volume with 5 μl of 10× Taq polymerase buffer (Sibenzyme), 2 μl of MgCl₂ (50 mM), 2 μl of dNTP (10 mM) mixture (Sibenzyme), 0.5 μl of each primer (0.01 mM), 0.5 U of Taq polymerase and 100 ng of genomic DNA. The primers used for the amplification of the COI gene were: F1 – 5’TCAACCAACCACAAAGACATTGGGCAC3’ and R1 – 5’TAGACTTCTGGTGCCAAAGAATCATAY’.

The thermal regime consisted of an initial denaturation for 2 min at 95°C followed by 35 cycles of 94°C for 40 sec, 54°C for 40 sec, 72°C for 1 min 10 sec, a final chain extension of 10 min at 72°C and held at a temperature of 4°C. The PCR products were visualized on 1.2% agarose gels, and the most intense products were selected for sequencing after quantification through spectrophotometry. The products were purified using Qiagen gel extraction kit. Sequencing reactions were performed bidirectionally using 96 capillary high throughput sequencer ABI 3730 XL to generate sequences with accurate DNA base calling following manufacturer’s instructions.

Sequences were aligned using CLUSTAL W and submitted to GenBank (Table 1). The extent of sequence difference between species was calculated by averaging pairwise comparisons across all individuals. The COI sequences of the conspecific individuals were aligned to yield a final sequence of 655 bp. The average nucleotide frequencies along with standard error were calculated. Pairwise evolutionary distance among haplotypes was determined by the Kimura two-parameter method using the software program MEGA (version 4) (Molecular Evolutionary Genetics Analysis).

The neighbour-joining (NJ) tree was constructed using MEGA 4 and to verify the robustness of the internal nodes of NJ tree, bootstrap analysis was carried out using 1000 pseudo replications.

The results were presented for eight species representing five families and four orders (Table 1). A total of 22 sequences were generated from the eight species using multiple specimens for all species. Sequencing of the COI gene produced 655 nucleotide base pairs per taxon. Sequence analysis revealed average nucleotide frequencies as T = 29.7%, C = 26.4%, A = 26.2%, G = 17.7% (Table 2). All the fish species studied showed relatively higher GC (45.1%) content; the highest GC content was found at the first codon (56.4%) position with progressive declining trend towards second (42.7%) and third (32.9%) codon positions (Figure 2). The average transitional (si = 62) pairs were more frequent than average transversional (sv = 47) pairs, with an average ratio (R = si/sv) of 1.3. The mean nucleotide diversity was calculated. The neighbour-joining (NJ) tree was constructed using MEGA 4 and to verify the robustness of the internal nodes of NJ tree, bootstrap analysis was carried out using 1000 pseudo replications.

Five individual fishes belonging to two genera and family Notopteridae under the order Osteoglossiformes were analysed. The sequence analysis revealed average nucleotide frequencies for T = 28.9%, C = 26.4%, A = 27.6%, G = 17.1% for C. chitala and T = 29.05%, C = 26.25%, A = 27.2%, G = 17.6% for N. notopterus (Table 2). The highest GC content was found at the first codon (56% and 56.4%) position with progressive declining
trend towards second (42.7% and 42.2%) and third (32% and 32.4%) codon positions (Figure 2) for *C. chitala* and *N. notopterus* respectively. The average transitional (*si* = 28) pairs were more frequent than average transversional (*sv* = 11) pairs, with an average ratio (*R* = *si*/*sv*) of 2.6. The average genetic distance within species was 0% whereas the average genetic distance between species was 11%. The NJ tree revealed distinct clusters shared by species of the same genera (Figure 3). All assemblages of conspecific individuals had 100% bootstrap values and the congeneric species formed the same clade.

Ten individual fishes belonging to three genera and family Schilbeidae under the order Siluriformes were analysed. The sequence analysis revealed average nucleotide frequencies as *T* = 31.4%, *C* = 24.45%, *A* = 27.38%, *G* = 16.75% for *A. coila*, *T* = 29.73%, *C* = 26.75%, *A* = 26.3%, *G* = 17.32% for *N. atherinoides* and *T* = 29.3%, *C* = 26.6%, *A* = 25.3%, *G* = 18.8% for *E. vacha* (Table 2). The highest GC content was found at the first codon (55.28%, 56.9% and 55.8%) position with progressive declining trend towards the second (42.6%, 42.6% and 42.73%) and third (25.83%, 36.5% and 33.55%) codon positions (Figure 2) for *A. coila*, *E. vacha* and *N. atherinoides* respectively. The average transitional (*si* = 53) pairs were more frequent than average transversional (*sv* = 25) pairs, with an average ratio (*R* = *si*/*sv*) of 2.1. The average genetic distance within species was 1.33%, whereas the average genetic distance between species was 19.3%. The NJ tree revealed distinct clusters shared by the species of the same family (Figure 3). All assemblages of conspecific individuals had high bootstrap values and the confamilial species formed the same clade.

Four individual fishes of *O. bimaculatus* belonging to the family Siluridae under the order Siluriformes were analysed. The sequence analysis revealed average nucleotide frequencies as *T* = 28.65%, *C* = 27.05%, *A* = 26.13%, *G* = 18.8% (Table 2). The highest GC content was found at the first codon (59.2%) position with progressive declining trend towards the second (42.6% and 42.73%) and third (32.4%) codon positions (Figure 2). The average genetic

<table>
<thead>
<tr>
<th>Fish</th>
<th>Average nucleotide composition (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
</tr>
<tr>
<td><em>Chitala chitala</em></td>
<td>28.9 ± 0</td>
</tr>
<tr>
<td><em>Notopterus notopterus</em></td>
<td>29.05 ± 0.21</td>
</tr>
<tr>
<td><em>Ompok bimaculatus</em></td>
<td>28.65 ± 0.44</td>
</tr>
<tr>
<td><em>Eutropiichthys vacha</em></td>
<td>29.3 ± 0</td>
</tr>
<tr>
<td><em>Ailia coila</em></td>
<td>31.4 ± 0.76</td>
</tr>
<tr>
<td><em>Neotropius atherinoides</em></td>
<td>29.73 ± 0.15</td>
</tr>
<tr>
<td><em>Pangasius pangasius</em></td>
<td>29.9 ± 0</td>
</tr>
<tr>
<td><em>Gudusia chapra</em></td>
<td>30.4 ± 0</td>
</tr>
</tbody>
</table>

**Table 2.** Mean percentage base composition (with SE) comparing cytochrome *c* oxidase 1 (COI) sequence of prioritized fishes
distance within species was 0%. The average genetic distance of individuals belonging to O. bimaculatus with other fishes under study (i.e. between species) was 20.3, with value ranging from 17% to 23%. The NJ tree revealed distinct clusters shared by the conspecific individuals having high bootstrap values (Figure 3).

A single individual fish of P. pangasius belonging to the family Pangasidae under the order Siluriformes was analysed. The sequence analysis revealed average nucleotide frequencies as T = 29.9%, C = 26.4%, A = 24.6%, G = 19.1% (Table 2). The highest GC content was found at the first codon (57.4%) position with progressive declining trend towards the second (42.7%) and third (39.8%) codon positions (Figure 2). The average genetic distance within species was 0%. The average genetic distance of individuals belonging to G. chapra with other fishes under study (i.e. between species) was 26.14, with value ranging from 25% to 28%. The NJ tree revealed distinct clusters shared by the conspecific individuals having 100% bootstrap value and formed an entirely different clade in the NJ tree (Figure 3) from the members of order Siluriformes and Osteoglossiformes.

In this study, the COI gene of eight species belonging to three orders and five families, collected from MTPRS was sequenced (Table 1). Efficient management of the threatened fish stock should start from unambiguous identification of any life-history stage of these fishes. DNA-based barcode identification system is believed to provide a simple, universal tool for identification of these fishes16.

The task of barcode generation is of utmost importance, since the task of routine species identification based on morphological and meristic features has four significant limitations: (i) incorrect identification of species due to phenotypic plasticity and genetic variability in the characters involved22; (ii) inability to recognize morphologically cryptic taxa23,24; (iii) keys designed for a particular life-history stage or gender may not be effective for the others and (iv) misdiagnoses due to lack of expertise and the dwindling pool of taxonomists signalling the need for a new approach to taxon recognition22.

The universal primers amplified the target region in all the fishes generating barcodes of 655 bp. The sequences were found to be conserved enough to allow amplification by a single set of forward and reverse primers in different fish families, yet diverse enough to permit unambiguous identification of fish species. No insertion, deletion or stop codons were observed in any of the sequences, which confirmed all amplified sequences being functional mitochondrial COI sequence. The 655 bp sequence length suggested that nuclear DNA sequences originating from mitochondrial DNA sequences (NUMTs) were not sequenced as vertebrate NUMTs are typically smaller than 600 bp (ref. 25).

Mitochondrial genomes show profound shifts in nucleotide usage among major taxonomic groups26 and this

---

Table 3. Summary of genetic divergence (K2P percentage) within various taxonomic levels calculated on the basis of COI sequence of prioritized fishes

<table>
<thead>
<tr>
<th>Comparison within</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Average</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>0.0</td>
<td>2.0</td>
<td>0.43</td>
<td>0.001</td>
</tr>
<tr>
<td>Families</td>
<td>11</td>
<td>19</td>
<td>16.8</td>
<td>0.039</td>
</tr>
<tr>
<td>Orders</td>
<td>11</td>
<td>21</td>
<td>18.0</td>
<td>0.030</td>
</tr>
</tbody>
</table>

---

Figure 3. Neighbour-joining tree constructed with the sampled fishes on the basis of COI gene sequence with bootstrap values.
can have serious impacts on phylogenetic analyses\textsuperscript{27–29}. The average GC content was found to be 45.1\% (Table 2) in this study and the highest GC content was found at the first codon position with progressive declining trend towards the second and third codon positions (Figure 2). This result corresponds reasonably well with those on complete mitochondrial genomes of \textit{Osteichthyes} species, deriving GC contents of 43.2\% (ref. 25) and 47.1\% (refs 8 and 16) for the COI gene sequence. The highest GC content (46.7\%) was found in \textit{G. chapra}, whereas the lowest (41.2\%) was observed in \textit{A. coila}. In a study embracing 16 animal phyla, the GC content was found to vary within a range of 22–53\%, with bony fish species \textit{Chirocentrus dorab} (52.8\%) and \textit{Salanx ariakensis} (53.4\%) showing higher GC content in general among all other species studied\textsuperscript{30}. Hence, it can be concluded that fishes in general display higher GC content.

The average transition and transversion ratio was 1.3, showing bias for transitional events. Members of the family Notopteridae, Schilbeidae and order Siluriformes demonstrated highest nucleotide changes occurring at the third codon position followed by the first and the second as evident from Table 4. This bias has been supported by observations in other fishes, like brook char (\textit{Salvelinus fontinalis}) control region\textsuperscript{31}, brown trout (\textit{Salmo trutta})\textsuperscript{32}, D-loop region of sturgeon\textsuperscript{33} and \textit{Poecilia reticulata}\textsuperscript{34}, as well as in coding regions of fishes\textsuperscript{35,36}. No transversional pairs were found at the first and second codon positions, ensuring minimal changes in the amino acid to be coded taking advantage of wobble effect. This reflects the fact that most synonymous mutations occur at the third position with a few at the first position and rarely at the second position. This observation was also in agreement with the earlier results\textsuperscript{37,38}.

The mean nucleotide diversity (\textit{Pi}) among all the species was estimated as 0.167. The COI locus harbours a high mutational rate for mtDNA\textsuperscript{37}. The average K2P distance of individuals within species was estimated as 0.43\%, whereas it was 16.8\% at the family level and 18\% for orders (Table 3). Hence, there was more than 39-fold sequence difference among confamilial species than conspecific individuals. The variation was even more (41.86 fold) among the individuals belonging to different orders. Hence, mean genetic distance within species was much smaller than the average distance between individuals of different species indicating marked genetic divergence beyond species level that has been referred to as the barcoding gap\textsuperscript{39}. All these observations were in agreement with the studies on Australian fish species\textsuperscript{4}, Alaskan skates\textsuperscript{39} and Indian marine fishes\textsuperscript{17}.

Although barcode analysis was performed to delineate species boundaries only, there was some phylogenetic signal in the COI sequence data\textsuperscript{4}. The NJ tree (Figure 3) revealed three distinct clusters formed by members of Siluriformes, Osteoglossiformes and Clupeiformes separately. The conspecific individuals were always found to cluster under the same node supported by high bootstrap value, whereas dissimilar species were clustered under separate nodes, enabling unambiguous identification of species. The nodes were supported by high bootstrap values. However, the true phylogeny of fishes from a 655 bp fragment of mitochondrial DNA through K2P distance and NJ method may not be conclusive enough\textsuperscript{4}, and hence more gene regions should be used (including nuclear genes) and additional analytic methods deployed, including maximum parsimony and maximum likelihood.

Hence, species-specific signals in the COI sequence should help in accurate and unambiguous identification of various developmental stages of threatened fish species. Our results suggest that COI barcoding can be taken up as a pragmatic approach for resolving unambiguity in the identification of the fish fauna with applications in its management and conservation. One management option could be a collection of early life-history stages of selected fish species from MTPRS, identified through DNA barcodes in monsoon or autumn, when fish abundance is maximum followed by their introduction in some freshwater protected area (FPA). The protected stock within FPA can be utilized for rehabilitation work in degraded aquatic systems and also to demonstrate the benefits of healthier ecosystems to the community and stakeholders. Protected areas could also serve as ‘banks’ of organisms for the replenishment of unprotected or degraded areas\textsuperscript{40}. This could be an invaluable tool for fisheries managers, fisheries ecologists and fish retailers, and for those wanting to develop fish identification microarrays.

---

RESEARCH COMMUNICATIONS


ACKNOWLEDGEMENTS. We thank the Head, Department of Zoology, University of Calcutta, Kolkata for providing infrastructural facilities and the Principal, Bethune College, Kolkata and Dipak Roy (Roy’s Fish Farm, Malda) for support.