Application of computational methods in fish identification based on mitochondrial DNA sequences

R. K. Negi1,3,*  Bheem Dutt Joshi2,3,  J. A. Johnson3 and S. P. Goyal2

1Department of Zoology, University of Delhi, Delhi 110 007, India
2Wildlife Institute of India, Dehradun 248 001, India
3Department of Zoology and Environmental Sciences, Gurukula Kangri University, Haridwar 249 404, India

The great discrepancy in sequence divergence of congeneric (0.4–0.6%) and conspecific (3%) individuals makes it difficult to identify species using DNA. A 650 base pair fragment of the cytochrome c oxidase subunit 1 (COI) gene from the fish Pethia conchonius was analysed using 30 samples. All the samples were identified as P. conchonius and two other congeneric species showing <2% sequence divergence with 29 samples out of 30 used, of P. conchonius. Two of the P. conchonius samples clustered with Puntius terio. Using different computational methods, we identified the sequence that was tagged as Puntius chola in the NCBI database as the P. conchonius sequence.

Keywords: Character attribute, Puntius, sequence divergence.

Species identification is an important aspect for exploring biodiversity, wildlife forensics, ornamental trade and other fields of biology. Fishes are a highly diverse species constituting more than 50% of all vertebrates globally.1,2 The segment near 5′-terminus of the mitochondrial cytochrome c oxidase subunit 1 (COI) is frequently used in taxonomy, and has been selected as the barcode region of the entire animal kingdom.3,3 The effectiveness of this gene has been validated in different animal groups, and

References:


Acknowledgements. Bhukya Saide is grateful to the Department of Biotechnology, New Delhi, India for financial support. Md. Imran acknowledges the SRF fellowship from DBT, New Delhi.

Received 7 December 2015; revised accepted 1 February 2016

doi: 10.18520/cs/v110/i11/2170-2172
most species that have been studied show distinct barcode arrays. Identifying closely related species (species complexes or hidden species) and those species recently diverged in the evolutionary time-scale, is difficult. The COI gene may be used to identify species, as this gene exhibits low levels of intra-specific sequence divergence and high levels of inter-specific sequence divergence in closely related taxa.\textsuperscript{3,4} However, DNA barcoding has been carried out for a large number of species\textsuperscript{5,6} and different threshold levels have been reported for conspecific and congeneric genetic distances (0.02–0.06 and 0.09–0.110 respectively) of different fish species.\textsuperscript{7,8} Some studies have used a genetic distance of ≥2% to define species boundary.\textsuperscript{9,10} At the same time, some species exhibit a lower sequence divergence than the aforementioned conspecific and congeneric genetic distances. This makes it difficult to use a single threshold level. Therefore, genetic distances vary from species to species.\textsuperscript{11,12}

Among fishes, genus Puntius/Pethia has more than 130 valid species with many hidden or species complexes.\textsuperscript{13,14} It is difficult to identify species from the DNA database using just one method. Sometimes false results are obtained using online reference repositories.\textsuperscript{15} Similarity-based and sequence divergence data may be interpreted incorrectly. Therefore, the use of multiple DNA sequence computational methods is suggested in place of single similarity-based methods that yield false results.\textsuperscript{15}

In this context, different computational and statistical methods are useful in identifying species from genetic data, such as distance, phylogenetic tree (topology),\textsuperscript{2,16} level of sequence variations\textsuperscript{17} and character attribute (CA)\textsuperscript{18,19} methods. These methods correctly identify recently diverged species. The aim of the present study is to identify $P$. conchonius using COI gene and identify the sequences submitted to NCBI database that have been incorrectly identified (false negative) as well as identify the character attributes of closely related species.

A total 30 samples of $P$. conchonius were used. Ten of these samples were collected from different rivers/streams in Uttarakhand. The sites include Song river, Diowala (30.18°N, 78.13°E), a tributary of the Ganga, Asan river, Vikasnagar (30.43°N, 77.72°E), a tributary of the Yamuna, and Kho river, Kotdwar (29.73°N, 78.52°E), a tributary of the Ramganga (Figure 1). Genomic DNA was obtained from samples (fin clip tissue) using a QIAamp Blood and Tissue Kit according to the manufacturer’s protocol (QIAGEN, Germany). A fragment (650 bp) from the extracted DNA was amplified using mitochondrial COI (LCO1490 F-5’-GGTCAACAAATCATTGAAGATATTGG3’; HCO2198R-5’-AAACTTCAGGTGACCAAATACTGA3’) gene.\textsuperscript{20} Polymerase chain reaction (PCR) amplification was carried out using 15 μl of a PCR master mix with 1.5× PCR buffer, 2.5 mM MgCl$_2$, 200 μM dNTP, 0.4 μM of each primer, 0.5 U Taq
Table 1. Character attributes observed in different species of genus Puntius and Pethia. Nucleotide positions were observed from the complete genome of Pethia conchonius

<table>
<thead>
<tr>
<th>Species</th>
<th>Haplotype</th>
<th>Sequence ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pethia conchonius</td>
<td>–</td>
<td>NC 022856.1</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>Hap1</td>
<td>PCD4* (n = 26)</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>Hap2</td>
<td>PCD5* , 6 (n = 2)</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>Hap3</td>
<td>PCA3* (n = 1)</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>Hap4</td>
<td>PCA9* (n = 1)</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>–</td>
<td>AB863607.1*</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>–</td>
<td>JQ667569.1*</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>–</td>
<td>JX260948.1*</td>
</tr>
<tr>
<td>Puntius terio</td>
<td>–</td>
<td>JX260958.1*</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>–</td>
<td>KP712125.1*</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>–</td>
<td>JX260946.1*</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>–</td>
<td>KF511543.1*</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>–</td>
<td>KF511542.1*</td>
</tr>
<tr>
<td>P. conchonius</td>
<td>–</td>
<td>JX260947.1*</td>
</tr>
</tbody>
</table>

A total of 12 parsimonious sequences were obtained from 30 samples taken in different sites (Kho river, Kotdwar; A, Sampling site. Doiwala, Song river. *GenBank.

The quality of the sequences was checked on Sequencer 4.7 (Gene Codes, USA). Sequences obtained were edited and cleaned, and finally clean sequences (490 bp) were obtained. ClustalW multiple alignments (CMA) in Bioedit (version 7.0.9.0) were used for further analysis. The generated sequences were compared with published sequences of COI gene in NCBI database and verified. Phylogenetic identification was carried out using a neighbour-joining (NJ) tree, with a bootstrap value of 1000 in the MEGA 6 software package. Character attributes were identified manually to find the specific nucleotide in the data sets in Bioedit and MEGA.

COI gene sequences of P. conchonius were generated successfully from 30 individuals. All the generated sequences were of good quality and were identified as P. conchonius with 100% similarity, using BLAST (NCBI). Sequences of each haplotype were submitted to GenBank (accession numbers KT957189–KT957196). Only sharp peaks were considered for mutations (Figure 2). Some other sequences (same genus) were found in the NCBI BLAST search, and these sequences showed 98.8% similarity, but they had been tagged as other species. One false negative sequence was also found. This had been tagged as P. chola, it showed 100% similarity with P. conchonius. A total of 12 parsimonious sites were observed, with six conserved sites that were different from the other sequences of species, viz. P. terio and P. khugae for Pethia conchonius. In contrast, the haplotype PCD5&6 (n = 2) has four base pair differences and a genetic distance of 0.008 from the other P. conchonius generated in the present study, and the sequences available in NCBI could be different species, and hence, need to be analysed, taking into consideration more sequences and morphological differences. The sequence of P. terio

Figure 2. Clean peaks of each haplotype generated in the present study highlighted in black.
Table 2. Pairwise sequence divergence in different species of Puntius

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PC</td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>PK</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>PC</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>PK</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>PC</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>PK</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>PC</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>PK</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>PC</td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>PK</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>PK</td>
<td>0.019</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>PK</td>
<td>0.019</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>PC</td>
<td>0.017</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PC = Pethia conchonius; PCh = Puntius chola (false negative); PT = Puntius terio; PK = Pethia kugae.

In contrast, in some cases, conspecific sequence divergence levels close to 3% and congeneric sequence divergence levels >8% have been found. However, these data suggest that the species involved have diverged a long time back. Large differences in the intra species sequence divergence may be a consequence of geographic separation and a disjunct species distribution pattern. Considering this, different computational methods are needed for identification of morphologically hidden species. Different methods like NCBI, BLAST (similarity), distance, character attributes and phylogenetic-based methods, that can discriminate species exhibiting low sequence divergence, have been used in our study. Character-based methods in particular are robust at identifying species.

All P. conchonius and P. terio individuals were grouped into a single clade with a low (70%) bootstrap value, which may be indicative of separate species that diverged recently from a common ancestor. The clade of P. conchonius was split into two major clades and further separated in the internal branch, showing intra-species splits. Only one of the 30 P. conchonius samples used clustered in the tree with P. terio and one sequence from NCBI (JX260947). The topology of the tree is similar to that of some other fishes. All the three sequences share the compound character attributes (characters present in other species also), as a result of which they cluster in a single clade.

One sequence identified as being from P. chola (JX260946) was identified in NCBI as being from P. conchonius by all the three methods, and one sequence identified as being from P. conchonius (JX260947.1) was identified as being from P. terio (Figure 3).

In conclusion, all the P. conchonius sequences were identified successfully. Two samples (Hap-2), showing a four base pair difference, could be different species, but this needs to be clarified using more sequences. One false negative was also identified, named as Puntius chola.
where the sequences of *Puntius terio* show very close proximity with *P. conchonius*. The present study suggests that different computational methods need to be used to identify species from DNA data, as there are several drawbacks in the use of online repositories\(^\text{15,28}\).


ACKNOWLEDGEMENTS. We extend our heartfelt thanks to the Dean, Faculty of Life Sciences and Head, Department of Zoology and Environmental Sciences, Gurukula Kangri University, Haridwar for their consistent support during the work. We also thank the Director and Dean, Wildlife Institute of India, for encouraging this work. We would also like to express our gratitude to our lab mates, who shared their personal experiences with us. We also thank the technical staff of the laboratories of the university and the institute.

Received 23 September 2015; revised accepted 28 January 2016

11.58250/ces/v110/i11/2172-2176