DNA BARCODING AND PHYLOGENETIC ANALYSIS OF FISHES OF POKHARA VALLEY

Submitted to

CENTRAL DEPARTMENT OF BIOTECHNOLOGY

Tribhuvan University

Institute of Science and Technology

Kirtipur, Kathmandu, Nepal

2014

By

Kalpana Subedi

Supervisor

Dr. Tilak R. Shrestha

Registration no.: 5-2-37-1163-2007

DEDICATION

I dedicate this thesis to my parents, Gopal P. Subedi and Goma Subedi who have always been my inspiration and so close to me that I found them with me whenever I needed. It is their unconditional supervision and affection that motivates me to resolute eminent targets. I also dedicate this thesis to my brother Kiran Subedi and sister Karuna Subedi, who are my nearest ones and have provided me with a strong love and care and never let any sorrow get into my life. I devote this work to my Lord Pashupatinath for listening to my prayers and bestowing me with the strength to defy every challenge on my way. Thank you so much all.....

ACKNOWLEDGEMENT

First of all, I am grateful to The Almighty God for establishing me to complete the dissertation. Besides, I owe a debt of gratitude to many individuals.

I would like to express the deepest appreciation to my supervisor, Dr. Tilak R. Shrestha, Associate Professor, Central Department of Biotechnology, Tribhuvan University, Nepal, who has the attitude and substance of genius. He imparted a spirit of adventure in regard to research continually in a convincing manner. Without his guidance and persistent help this dissertation would have not been possible.

I sincerely thank Dr. Rajini Malla, Head of Department, Central Department of Biotechnology, TU, for her recommendation to complete my thesis from the reputed laboratory.

Much of my experimental work would not have been completed without the assistance and guidance of Anita Tiknaik mam, Rahul Jamdade sir, Dinesh Nalage sir and Chandrakant Jhadav sir in the PHCDBS, BAMU, Aurangabad, India.

Of course, I owe a special gratitude to number of personalities who contributed in the research indirectly; NV Ivanova and the group for their contribution with the primers; J Messing for M13 tailed primers and Hajibabei and the team for their contribution with cycle sequencing primers.

I wish to thank all the members of the Central Department of Biotechnology. The faculty, staffs and students here really have made my journey very encouraging and memorable.

The assistance, cooperation and experience of my dear fellow and friends were essential for the completion of my work. I'd like to thank Diptee Chaulagain for being with me throughout. I am also thankful to Preeti Regmi and her family for all the time and help during sample collection. I am grateful to Sajid Khan and Dinesh sir for their care and concern during my stay in Aurangabad. I want to remember all of my friends from 3rd batch for their cooperation and fellow feeling.

I'd like to express my sincere gratitude to my parents, Gopal P. Subedi and Goma Subedi who took care and encouraged me since my childhood. I warmly thank my siblings, Kiran and Karuna for their love and support.

I want to pay special regards to one and all who, directly or indirectly, have lent their helping hand in this venture. Everybody contributed to accomplish my aspiration. CHEERS.

ABSTRACT

DNA Barcoding and Phylogenetic Analysis of Fishes of Pokhara Valley

Despite extensive taxonomic studies, identification of fishes can be problematic often even to the experts due to various reasons. In this context, DNA barcoding can be a promising tool for species identification and biodiversity surveys through the use of short, standardized gene targets, ~652 bp of mitochondrial DNA. This tool can be more broadly applied if a comprehensive reference sequence library for all fish species can be constructed. Here, we make a small contribution to this grand challenge by barcoding some freshwater fishes from Pokhara. The standard barcode fragment of COI was used to barcode 14 individuals, representing 14 taxonomically recognized species in 13 genera, 7 families and 5 orders. A 99% sequence similarity threshold was employed as a matching criterion for specimen identification to the species level. After editing all obtained sequences using Codon Code Aligner 4.0 program, specimens and sequence data were archived and investigated using analytical tools available on BOLD and MEGA. The GC content was 44.97% on average. Mean genetic distance between families was 18.7%. The synonymous changes were much greater than the non-synonymous changes, especially in the 3rd codon position where variation is dominated. There were 174 conserved, 43 variable, 14 parsimony-informative and 29 singleton amino acid sites; while 367 conserved, 285 variable and 218 parsimony-informative sites were present out of 652 bp nucleotides. The NJ, ML and MP analysis indicated different clades corresponding to the recognized groupings; members of same families clustered together. Molecular species identification was in concordance with current taxonomical classification in all cases achieving success rate of ~94%. In addition to DNA barcodes, our study also provides supporting data in the form of specimen images, morphological characters, taxonomic bibliography, preserved vouchers and COI sequences. This work highlights the functional utility of barcodes for the discrimination of diverse ichthyofauna. We infer that DNA barcoding can be a valuable tool to increase accuracy, objectivity and comparability of taxonomic assessment in biodiversity studies. Finally, our study constituted an important contribution to the iBOL, providing barcode sequences for use in identification of the species by experts and non-experts, and allowing them to be available for use in other applications. Further research is needed to verify the deeper divergence within species and genera with larger sample size.

Keywords: Mitochondria, Cytochrome oxidase, Cytochrome oxidase subunit I (COI) gene, Taxonomy, DNA sequencing, Species identification, GenBank, BOLD.

TABLE OF CONTENTS

Title Page……I

Recommendation……II

Certificate……III

Certificate of Evaluation……IV

Dedication……V

Acknowledgement……VI

List of Abbreviations and Acronyms……VII

Table of Contents……XI

List of Tables, Figures and Appendices……XIV

ABSTRACT……**XVII**

CHAPTER 1: INTRODUCTION

1.1 Background……1

1.1.1 Pokhara Valley……2

1.1.2 Lake Begnas……2

1.1.3 Ichthyology……2

1.1.4 History of Ichthyology……4

1.1.5 Taxonomy……5

1.1.6 1 History of Taxonomy……6

1.1.7 Morphological Taxonomy……7

1.1.8 Fish Taxonomy……8

1.2 Current Studies……10

1.3 Hypothesis……10

1.4 Objectives……11

1.4.1 Broad Objectives……11

1.3.2 Specific Objectives……11

1.5 Rationale/Justification of the study……11

1.6 Scopes of the study……12

CHAPTER 2: LITERATURE REVIEW

2.1 Molecular Taxonomy or DNA Taxonomy……13

2.1.1 Mitochondrial DNA……13

2.1.2 COI as Barcode Region……15

2.1.2.1 Cytochrome c Oxidase……15

2.1.2.2 COI Gene……18

2.1.2.3 Other Popular Barcoding markers……19

2.1.3 NUMTS (Nuclear mitochondrial pseudogenes)……20

2.1.4 Indels……21

2.2 Barcoding Databases: A brief Intro……22

2.2.1 Components of Barcoding Projects……22

2.2.2 Fish BOL……26

2.2.3 FISHBASE……27

2.2.4 BOLD……27

2.2.5 Barcode Index Number (BIN)……29

2.3 Phylogenetics……30

2.4 DNA Barcoding and Population Genetics……32

2.5 DNA Barcoding: Merits, Scopes and Challenges……34

2.5.1 Merits of DNA Barcoding……34

2.5.2 Scopes of DNA Barcoding……34

2.5.3 Challenges of DNA Barcoding……35

2.6 Status of Molecular Taxonomy in Nepal……36

CHAPTER 3: METHODOLOGY

3.1 Study Area/Sampling Stations……37

3.2 Collection of Fishes……37

3.3 Photography……37

3.4 Tissue Sampling……38

3.5 Fish Identification Methods……38

3.6 DNA Extraction Using CTAB Method……38

3.7 Quantification of DNA……39

3.8 Qualitative Analysis of DNA……39

3.9 PCR Amplification of Gene……40

3.10 PCR Amplicon Check up……41

3.11 PCR Clean up……41

3.12 Cycle Sequencing Reaction……42

3.13 Ethanol Wash of Cycle Sequenced Product……43

3.14 Sequencing……43

3.15 DNA Sequence Alignment……43

3.16 Deposition of Data……44

3.17 Data Analysis……44

CHAPTER 4: RESULTS

4.1 Morphological Classification……45

4.2 DNA Processing Results……45

4.3 COI Gene Profiles……46

4.3.1 Species Identification……48

4.3.2 Grading for Taxonomic Reliability……49

4.3.3 Selection of the Best Fit Model for Analysis……49

4.3.4 Nucleotide Composition……50

4.3.4.1 GC Content……50

4.3.5 Amino acid Composition……51

4.3.5.1 Amino acid Variability……51

4.3.6 Transition/Transversion Bias……52

4.4 Pairwise Distances……56

4.5 Barcode Gap Analysis……56

4.6 Percent Similarity……57

4.7 Genetic Diversity Using Phenograms……62

CHAPTER 5: DISCUSSION

5.1 Mitochondrial COI as Barcode……67

5.2 Species Identification Based on BLAST and BOLD……67

5.3 Ranking System……68

5.4 Compositional Analysis of COI Sequence……69

5.4.1 Nucleotide variation vs. Amino acid variation……70

5.5 Sequence Divergence……71

5.6 Phylogenetic Analysis……72

5.7 Molecular Taxonomy Complements Morphological Taxonomy……74

5.8 Barcoding in Our Perspective……74

CHAPTER 6: SUMMARY AND CONCLUSION

6.1 SUMMARY……75

6.2 CONCLUSION……76

RECOMMENDATIONS……77

REFERENCES……78

APPENDICES……88

LIST OF TABLES, FIGURES AND APPENDICES

A] TABLES **Page no.**

Table 2.1 Common species level molecular markers 29

Table 3.1 PCR reagent composition and reaction volume 40

Table 3.2 PCR conditions for COI gene of fishes 40

Table 3.3 Cycle sequencing reagent concentration 42

Table 3.4 Cycle sequencing PCR reaction 42

Table 3.5 Master Mix composition for Cycle sequencing product washing 43

Table 4.1 Family wise number of individuals studied 45

Table 4.2 Details of fishes studied for their molecular taxonomy using DNA Barcoding 47

Table 4.3 BOLD ID and NCBI accession of the submitted sequences of the species studied 48

Table 4.4 Percentage similarity of COI gene of the specimen obtained from BLAST 48

Table 4.5 Attribution of grades (A to E) to DNA barcodes of 13 fish species from Begnas Lake 49

Table 4.6 Nucleotide composition (%) of the COI sequences under study (13 sequences) 51

Table 4.7 MCL transition / transversion bias 54

Table 4.8 Nucleotide frequency at various positions 54

Table 4.9 Pairwise distance of COI sequences of fishes 56

Table 4.10 Species with their nearest neighbor and distance to them 57

Table 4.11 Species sample code with their sequences and barcodes 59

Table 5.1 Summary of the DNA barcoding surveys of the freshwater fishes highlighting the number of species, higher taxa, families and genera with multiple species analyzed 71

B] FIGURES

Figure 1.1 Sample site map. Map of Kaski district showing Pokhara valley with Lake Begnas. 3

Figure 1.2 Basic taxonomy of fishes 8

Figure 2.1 Mitochondrial DNA showing location of genes and other key regions 16

Figure 2.2 Cytochrome Oxidase 16

Figure 2.3 Mitochondrial DNA of *Cirrhinus mrigala* 17

Figure 2.4 Mt-COI gene: a detailed view 18

Figure 2.5 Schematic view of a linearized mitochondrial DNA showing the relative positions of most coding and noncoding regions 18

Figure 2.6 DNA Barcoding workflow 25

Figure 3.1 *Mastacembelus armatus* 37

Figure 3.2 *Oreochromis mossambicus* 37

Figure 4.1 Quality check for genomic DNA (1%agarose gel) 45

Figure 4.2 PCR amplified product 46

Figure 4.3 Synonymous and nonsynonymous substitutions per site 50

Figure 4.4 Amino acid composition 53

Figure 4.5 Nucleotide frequency of all the taxa studied 53

Figure 4.6 COX1 substitution plots. Number of transition and transversion at different codon position 58

Figure 4.7 Histogram of Barcode gap analysis 58

Figure 4.8 Phylogenetic tree inferred using Neighbor joining method based on K2P distance using MEGA5.2 software 62

Figure 4.9 Molecular phylogenetic analysis by Maximum Likelihood method based on Tamura-Nei model by using MEGA5 software, version2 63

Figure 4.10 Maximum Parsimony tree constructed using Tamura-Nei model and the

closest neighbor interchange method of the MEGA 5.2 software package 63

Figure 4.11 An NJ phylogram showing COI barcode divergences in specimens of Order Perciformes analyzed in the present work and of GenBank species 64

Figure 4.12 K2P distance NJ tree of COI sequences from the species of the Order Cypriniformes analyzed in the present work and of GenBank 65

Figure 4.13 NJ tree based on the mitochondrial DNA COI nucleotide sequences of Siluriformes analyzed in the present work and of GenBank species 66

C] APPENDICES

Appendix 1 Composition of various reagents used 88

Appendix 2 Classification of fishes studied in the research along with the IUCN conservation status 89

Appendix 3 Concentration and absorbance of the extracted DNA, along with the dilution to be made for amplification 91

Appendix 4 Percent GC content in fishes studied 91

Appendix 5 Similarities of the species with one another in percentage using Clustal W 92

Appendix 6 CLUSTAL O (1.2.0) multiple sequence alignment 92

Appendix 6: Species identification using reference library from BOLD system. 95

Appendix 7 List of all the specimens belonging to 3 families Cypriniformes, Perciformes and Siluriformes extracted from GenBank for analysis with the accession numbers 96

Appendix 8 Quick view of BOLD System:v3 97

Appendix 9: Specimen Data Sheet of the studied fishes in BOLD. 98

LIST OF ABBREVIATIONS AND ACRONYMS

µg : microgram

µl : microlitre

bp : base pair

dNTPs : Deoxyribonucleotides

ibol : Barcode of life initiative

matK : Maturase K

mg : milligram

min : minute

ml : milliliter

mtDNA : Mitochondrial DNA

mA : milli ampere

mM : milli molar

ng : nanogram

nm : nanometer

rbcL : Large subunit of ribulose 1,5-biphosphate carboxylase/oxygenase

rDNA : Ribosomal DNA

rRNA : Ribosomal Ribonucleic acid

rpm : revolution per minute

sec : second

tRNA : Transfer Ribonucleic acid

AFLP : Amplified Fragment Length Polymorphism

BIN : Barcode Index Number

BLAST : Basic Alignment Search Tool

BOL : Barcode of life

BOLD : Barcode of life data system

CBOL : Consortium of BOL

CMDN : Centre for Molecular Dynamics- Nepal

COI : Cytochrome c Oxidase Subunit I

Cox1 : Cytochrome oxidase 1

CR : Control Region

CTAB : Cetyl Trimethyl Ammonium Bromide

Cytb : Cytochrome b

D/W : Distilled water

DDBJ : DNA Data Bank of Japan

DMAS : Data Management and Analysis System

DMSO : Dimethyl Sulphoxide

DNA : De-oxy ribonucleic acid

ED : Evolutionary Distance

EDTA : Ethylene Diamine Tetra aceticacid

EMBL : European Molecular Biology Lab

ESUs : Evolutionary Significant Units

Exo-SAP : Exonuclease I-Shrimp Alkaline Phosphatase

FASTA : Fast Analysis

GBIF : Global Biodiversity Information Facility

Indels : Insertions/Deletions

ITIS : Integrated Taxonomic Information System

ITS : Internal Transcribed Spacer

K2P : Kimura-2-Parameter

LIMS : Laboratory Information Management System

MEGA : Molecular Evolutionary Genetic Analysis

ML : Maximum Likelihood

MOTUs : Molecular taxonomic units

MP : Maximum Parsimony

MSA : Multiple Sequence Alignment

NA : Not applicable

NAST : National Academy of Science and Technology

NCBI : National Center for Biotechnology Information

NJ : Neighbor Joining

OD : Optical Density

OTU : Operational Taxonomic Unit

PCE : Phenol/Chloroform Extraction

PCGs : Protein Coding Genes

PCR : Polymerase chain reaction

RESL : Refined Single Linkage

RFLP : Restriction fragment length polymorphism

Rhod : Rhodopsin gene

RT : Room Temperature

SDS : Sodium Dodecyl Sulphate

SSR : Simple Sequence Repeat

TBE : Tris Borate EDTA

CHAPTER 1

INTRODUCTION

1.1 Background

Fishes are immeasurable assessment to human. They have long been consumed as a nutritious diet by many people. At present they form an important element in the economy of many nations while giving inestimable recreational and psychological value to the naturalist, sports enthusiast, and home aquarist. Fishes have number of economic importance; they are used as food, oil, meal, medicine; for making skin and leather; for disease controlling and therapeutics; for sports and recreation as well as for scientific studies. Fishes comprise nearly half of all vertebrate species; the group includes approximately 15,700 marine and 13,700 freshwater species (FishBase: www.fishbase.org). Total 186 species of fish, including the indigenous and exotic, belonging to 11 orders have been recorded from Nepal (Shrestha J, 1995; Subba BR et *al.,* 1996). The large diversity of fish species in Nepal is explained by the diversity of climatic zones, from subtropical to high mountains, and the fact that Nepal lies at the transition point of the Indo-Malayan and Palaearctic biogeographical realms (Shrestha J, 2002).

Nepal is a landlocked country, even though it is richest in water resources after Brazil. There are plenty of rivers with perennial water supply from Himalayas, large number of lakes and few reservoirs where the fishes inhabit. The rivers of Nepal are broadly classified into three categories; major, medium and minor based on their origin and water discharge. The first category includes four major rivers - the Koshi in the eastern region, the Gandaki in the central region, and the Karnali and Mahakali in the far-western region. The first three major rivers originate from the northern slopes of the greater Himalayas and subsequently cross the Himalayas, while the fourth river, the Mahakali, originates from the high mountains of Nepal Himalayas. A number of medium and small lakes are scattered throughout the country ranging from sub-tropical warmer areas to freezing altitudes. Much studied lakes of Pokhara valley - Lake Phewa (523 ha), Begnas (328 ha) and Rupa (135 ha) have well established aquaculture and capture fishery (Rai AK *et a.l,* 1995). From the lakes of Pokhara, 25 indigenous fishes have been reported (Ferro W *et al.,* 1980). The Mahendra Tal (Lake Rara) is the biggest lake. The lake has three endemic fish species (Rajbanshi KJ, 2001). Begnas Lake is the second largest lake of Pokhara Valley. Considerable amount of fishes are found there. The lake is situated at an elevation of 650 m, covers 328 ha and has a maximum depth of 10 m.

1.1.1 Pokhara Valley

Nepal is a landlocked country and its natural waters are classified into five categories:(i) rivers and streams, (ii) lakes, (iii) reservoirs, (iv) swamps, and (v) lowland paddy fields. Water bodies including rivers, streams, lakes and reservoirs, are used for multiple purposes such as drinking and other household water uses, industrial use, irrigation, aquatic crops production, hydropower generation, recreation and tourism, fisheries, including conservation of aquatic genetic pools, etc. They also provide a habitat for aquaculture production. There are many medium and small lakes in the country, with about 5,000 ha of water surface area. These lakes have different origins and can be classified as (a) glacial, (b) tectonic, and (c) oxbow lakes (Shrestha MK *et al.,* 2001). The mid-hill lakes are tectonic. Pokhara valley is one of the large midlands of Nepal, situated in the western part of the country. The climate of this valley is subtropical with a well defined rainy season. This valley has 3 sizeable lakes namely Phewa, Begnas, Rupa and several small lakes. These sizable lakes constitute one of the main sources of fish protein and also contribute to the natural beauty of this valley (Swar DB *et al.,* 1980). The lakes of Pokhara support a diverse fish community. Of the total of 186 indigenous fish species which have been recorded for Nepal by Shrestha J. (1995), more than 15% of the fish species are found here.

1.1.2 Lake Begnas

Begnas Lake, the second largest lake of the 8 lakes in the Pokhara valley of Nepal, is located in the Siswa village on the eastern part of Pokhara and is 13 km away from the city of Pokhara. It is a multipurpose lake used for irrigation, commercial fish production, fisheries research, and recreation. Begnas Lake is shallow with very dense vegetation around the shore. It has a single spring-fed inlet and a single outlet which finally joins the Seti River (Swar DB *et al.,* 1980). The main source of lake is the Shyankhundi stream which flows into the lake from west to south. Since this flow is insufficient, rainwater is collected to fill up the lake (Rai AK, 2000). Lake Begnas has a watershed area of approx. 20 km^2, surface area of 225 ha, maximum depth of 10 m and mean depth of 6.6 m (Ferro W *et al.,* 1978; Ferro 1981/82; Rai AK 2000). It is situated at 28^{o}17' N and 84^{o}07' E and 650 m above mean sea level. It is subtropical and moderately eutrophic lake with surface water temperature at noon from 15^{o}C in February to 30^{o}C in July (Swar DB *et al.,* 1988). The local fish fauna is comprised by members of families – Cyprinidae, Siluridae, Anguillidae, Belonidae, Channidae and Mastacembelidae which vary in food habit from exclusively carnivorous to omnivorous (Ferro W *et al.,* 1980). The Cyprinidae are represented by 7 species- *Barilius barna, Barilius bendelensis, Cirrhinus rewa, Labeo*

gonius, Puntius sarana, P. sophora and *Tor tor*, while the remaining 5 families are

Figure 1.1: Sample site map. Map of Kaski district showing Pokhara valley with Lake Begnas.

(Source: http://www.thekingdomofnepal.com/)

represented by single species, namely *Mystus cavasius, Anguilla bengalensis*, *Xenentodon cancila*, *Channa gachua* and *Mastacembelus armatus*, respectively. *T. tor, L. gonius, C. rewa* ad *P. sarana* are major economic species. Lake Begnas was stocked with 4 exotic carp species- *Hypophthalmichthys molitrix*, *Aristichthys nobilis, Ctenopharyngodon idellus, Cyprinus carpio* and *Labeo rohita* (Swar DB *et al.,* 1988). As this lake is prone to anthropogenic activities such as agricultural inputs and human settlements, the fresh river water might have spatial variation in sediment and dissolved load (Khadka UR *et al.*, 2012), which can ultimately result into altered physiochemical reaction thereby affecting the flora and fauna inhabiting there. So, it is one of the threatened habitats of Nepal (Khadka UR *et al*., 2012).

1.1.3 Ichthyology

The study of fishes is broadly termed as ichthyology. It is mostly concerned with studies of diversity, distribution, and interrelationships of fishes. It also deals with the physiology or functional morphology of fishes, seeking to determine how the various body parts of fishes interact to facilitate feeding, locomotion, respiration, or other vital functions (Etnier DA *et al.,* 2001). According to FishBase, 31,500 species of fish have already been discovered and described by early 2010, which is more than the combined total of all other vertebrates. The practice of ichthyology is associated with marine biology, limnology and fisheries science. There are 32,000 living species of fishes which are distributed among approximately 515 families and 4,494 genera. Of the approximate 970 living species of Chondrichthyes (sharks, skates, rays, and chimaeras), more than half (534 or about 55%) are rays. 96.6% of all living fish species are Actinopterygians or bony fishes. Out of them 96.4% belong to a group Teleosti. Cyprinidae, Gobiidae, Cichlidae, Characidae, Loricariidae, Balitoridae, Serranidae, Labridae, and Scorpaenidae are nine largest families of 515 fish families which contain about 30% of all species. Among bony fishes, 41.2% are freshwaters, while 58.2% belong to marine habitat. About 300 new species of fishes are described each year. Most of these are freshwater forms which come from high-diversity tropical habitats, but a significant number are marine fishes which come from the deep-sea.

1.1.4 History of Ichthyology

Although there were various people who investigated on fishes before, Peter Artedi, a Swede from University of Uppsala, is often considered as 'The father of Ichthyology'. His work "Ichthyologia" was published in 1738 by Linnaeus after his death. Artedi had recognized 47 genera and 230 species of fishes. He believed that genus represented a

group of species which agreed with each other in general but differed in minor characters. He grouped the genera into 'maniples', same as the Family. Maniples were arranged into orders and these into a class. Linnaeus was greatly influenced by his method of classification. Many other Ichthyologists such as **Otto Fabricus** (1744-1822), **Petrus Forskål** (1736-1763), **Petrus Pallas** (1741-1811), **Antione Risso** (1777-1845), **Thomas Pennant** (1726-1798), **Wlhelm G. Tilesius** (1769-1857), **Georg Wilhelm Steller** (1709-1746) followed his work. The first Ichthyologist to publish an actual description was an American named LeSueur (1821). Albert C.L.G. Gunther (1830-1914) wrote a series of volumes on the fishes of the world, entitled *Catalogue of the Fishes of the British Museum. It was published in* 8 volumes. The next attempt was made by Johann Baptis von Spix (1781-1826) and Louis Aggassiz (1807-1873). They worked on Brazilian fishes. Johann Müller (1801-1858) and Friedrich G. J. Henle (1807-1885) produced the first authoritative work on sharks (*Systematische Beschriebungen der Plagiostomen*) in 1841. Peter Bleeker (1819-1878) published a book named *Atlas Ichthyologique des Indes Orientales Néerlandaises* on the fishes of the tropical Indo-Pacific. (http://www.angelfire.com/biz/piranha038/taxon.html).

1.1.5 Taxonomy

Taxonomy is the theory and practice of classifying organisms. It is defined as the science related to discovery, recognition, definition, and naming of groups of organisms. It was one of the first sciences compelled to organize a large body of data. Linnaeus and his immediate disciples introduced the binomial system of nomenclature and the hierarchical system of higher taxa. It was a first attempt to do this, at a time when the diversity of life was underestimated by several orders of magnitude (Godfray HCJ *et al.,* 2004). Taxonomy provides opinions on species boundaries, and on the phylogenetic relationship between species. It provides a stable naming system that work as a gateway to huge store of information about a species (Godfray HCJ, 2007). In taxonomy, species is regarded as the basic unit but this concept varies among taxonomists. Species is defined as a group of interbreeding or potentially interbreeding populations reproductively isolated from all others (Etnier DA *et al.,* 2002). The basic taxonomy scheme divides living organism into Domains, Kingdoms, Phylums or Divisions, Classes, Orders, Families, Genera and Species. Sometimes Subphylums /Subdivisions, Subclasses, Suborders and Subfamilies are also used (www.aquaticcommunity.com). In the most basic classification system, species believed to be closely related are grouped within genera, connected genera within families, families into orders and related orders within families (Etnier DA *et al.,* 2002).

The fact that DNA sequence variegation can be calculated either directly or indirectly through protein analysis has been known since several decades. A starch gel electrophoresis of proteins was first used to identify species more than 40 years ago. Nearly 30 years ago, single gene sequence analysis of ribosomal DNA was being used to investigate evolutionary relationships at a high level and mitochondrial DNA approaches dominated molecular systematic in the late 1970s and 1980s (Ward RD *et al.,* 2005). A wide variety of protein- and DNA-based methods have been used for the genetic identification of fish species (Ward RD *et al*., 2005). Molecular barcoding is one of the major emerging ideas in species-level taxonomy, and will be more demanding in the years to come. At the moment molecular barcoding is a niche activity and probably only cost-effective when traditional morphology fails (Godfray HCJ *et al.,* 2004).

An improved system of listing biodiversity and disseminating taxonomic information is required to resolve the limited knowledge of species diversity in many areas of the globe, along with anthropogenic disturbance of ecosystems (Monaghan MT, 2006). With the goal of accelerating the rate at which new species are discovered and described, new ambitious methods for species delimitation have been developed and compared, including DNA barcoding, DNA taxonomy and Web-based taxonomy (Wiens JJ, 2007).

1.1.6 History of taxonomy

Taxonomy simply means classifying living organisms according to their natural relationships. The Greek philosopher, Aristotle is credited with starting taxonomy. He was the first to attempt to classify all the animals according to habitat and body form. He grouped all living things into 2 kingdoms: plants and animals. Within the animal kingdom, he further divided animals based on their anatomical and physiological similarities and differences. Animals with blood (vertebrates) were subdivided into live-bearing (mammals) and egg-bearing (birds, fishes). Animals without blood (invertebrates) were subdivided into insects, crustaceans and mollusks. However, by 16th century many new species had been discovered which was unable to be classified based on Aristotle's system. An English naturalist, John Ray then introduced the genus and species method of naming organisms. His methods of distinguishing were also superficial which led to wrong classification. Some other early taxonomists were Caesalpino, Bauhin, Tournefort who contributed in plant classification system. Among all, Carolus (Carl) Linnaeus, a Swedish botanist is the best known taxonomist. He is considered as "Father of Taxonomy" because he was the first to combine a hierarchical system of classification from kingdom to species with the method of binomial nomenclature. "The System of Nature (Systema Naturae)" published by him is remarkable for an overall framework of classification. In hierarchical system, 7 taxa or levels are there, starting

with the category of greatest diversity down to the smallest category: Kingdom, Division, Class, Order, Family, Genus and Species. By mid 80s, taxonomy was no longer the interest of scientists; as a result it was facing a worst crisis. Charles Darwin's publication "The Origin of Species" in 1859 brought a new idea of taxonomic categorization based on evolution. Darwin's theory presented a hypothesis that if two groups of organism shared similar characteristics and were placed in same taxon, they probably shared common ancestor. His theory of evolution has allowed the scientists to see diversity as the result of a dynamic process rather than a static figure. Today, eventually scientists have learned to classify the organism using genetic sequences. (http://www.biologyreference.com/Ta-Va/Taxonomy-History-of.html; http://www.angelfire.com/biz/piranha038/taxon.html, http://www.shmoop.com/taxonomy/taxonomy-history.html, http://davesgarden.com/guides/articles/view/2051/)

1.1.7 Morphological taxonomy

Morphological taxonomy is the primary tool used by taxonomists all over the world to reveal the inexplicable biodiversity. It is the most primitive methodology which relies on morphological characteristics. The whole concept of morphological taxonomy is contribution of 'The father of Taxonomy', Carl Linnaeus. He gave many things to biological science, namely: a) morphological criteria for species discrimination, b) hierarchical system for classification to catalogue them, c) binomial nomenclature for scientifically naming the classified species etc. Even in the era of DNA taxonomy, it is essential to classify the organism morphologically as well. It's already been more than 300 years since this method of classification is being practiced. Till now most of the organisms in the earth have been explained on the basis of morphological taxonomy. This method has merits in such places where applied researches, hi-tech instruments, funds and trained manpower for molecular taxonomic researches are limited or lacking. It also has some defects. The construction of morphological keys is difficult and laborious. This system is often criticized for its reliability. Even among the taxonomists there is conflict regarding the traits selected as keys. Therefore, with the advancement in molecular evolution, various non-morphological methods such as biochemical, physiological, cytogenetic approaches and molecular taxonomy have been tremendously developed for the purpose of classification which previously relied only on morphological features (Krishnankutty N *et al.,* 2008). It can't be denied that morphological taxonomy can direct to extremely high or low estimation of biodiversity because morphology is a complex non-neutral marker. These obstacles of morphological taxonomy can be subdued by using DNA barcoding to classify living organisms. DNA

barcoding, also called as DNA taxonomy is used for identification and allocation of specimens to taxonomic groups that has been described in the past. Thus it helps in classification of new taxa (Lefebure T *et al*., 2006).

1.1.8 Fish Taxonomy

Many genetic, physiological, behavioral, morphological and ecological data are available for taxonomic and evolutionary studies. Fish species have characteristic shapes, sizes, pigmentation patterns, disposition of fins, and other external features that aid in recognition, identification, and classification (Strauss *et al*.).

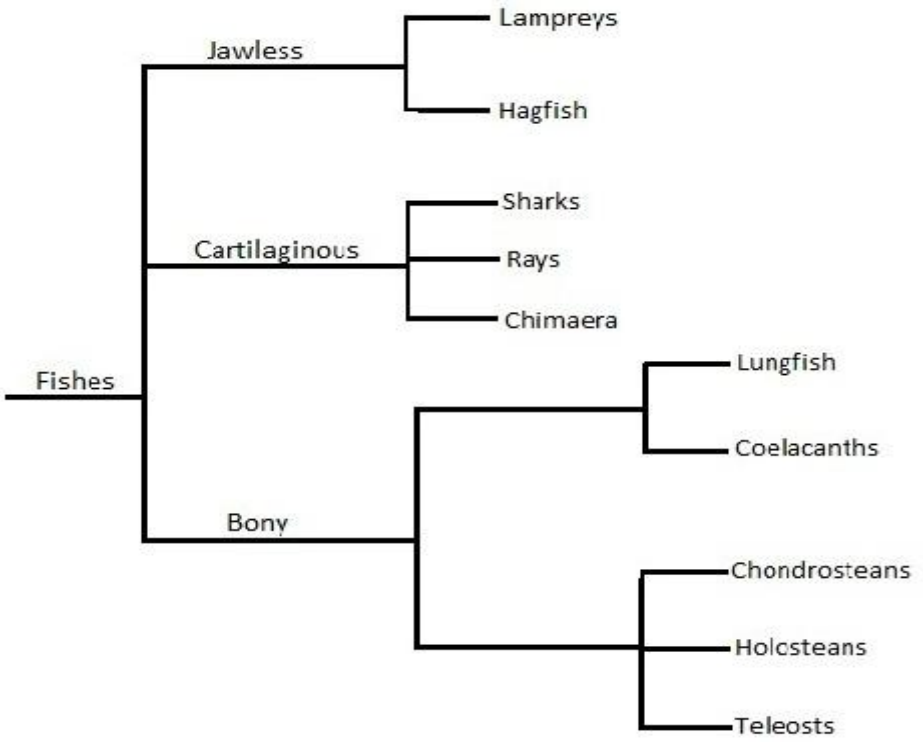


Figure 1.2: Basic taxonomy of fishes

(*en.wikipedia.org/wiki/Template:Basic_fish_taxonomy*(

This falls under morphology based approaches. Conventional taxonomic methods are undoubtedly of great value for this purpose but they have their limitations especially when the species under study are too small or when geographical and other factors, unidentified as yet complicate the picture (Vishwanathan R *et al*., 1956). Estimation of the population, size, and taxonomy of fish is important in order to manage fish populations, regulate fisheries, and evaluate the impact of manmade structures such as dams. Previous research methods for automated fish identification and taxonomy have depended on a human expert to design the features the identification algorithm use.

Adapting to other environments with different fauna is difficult, time consuming, and costly (Lillywhite K *et al.,* 2011).

Nepal, being rich in biodiversity, has a variety of natural privileges adequately. Fishes are the major section of the biodiversity which exhibit enormous diversity in their morphology, in the habitats they occupy, and in their biology (Nelson JS, 2006). These diversities are yet to be identified properly and named and it's a difficult task to be accomplished relying only on the morphological basis. Moreover, the task of routine species identification has many limitations. First, both phenotypic plasticity and genetic variability in the characters employed for species recognition can lead to incorrect identifications. Second, this approach overlooks morphologically cryptic taxa which are common in many groups. Third, morphological keys are effective only for a particular life stage or gender in most of the cases. These constraints demand new approach to taxon recognition. Thus, DNA based identification system can be opted for the identification purposes. Genomic approaches to taxon diagnosis exploit diversity among DNA sequences to identify organisms. In a very real sense, these sequences can be viewed as genetic 'barcodes' that are embedded in every cell (Hebert PDN *et al.*, 2003).

DNA barcoding simply implies the identification of any organisms using DNA. Barcodes are relatively short, specifically defined DNA sequences used to recognize organisms by comparing the barcode sequence from an unknown sample to a collection of sequences from known reference samples (Hebert PDN *et al.*, 2003). It was proposed by Hebert *et al*. as a method to identify species in 2003. Over 1.9 million specimens, belonging to roughly 17,2000 species, have been barcoded since then, including 9,502 fishes (Pereira LHG *et al.*, 2013). DNA barcoding uses standardized 500 to 800-bp sequences to identify species of all eukaryotic kingdoms using primers that are applicable for the broadest possible taxonomic group (Schoch CL *et al.*, 2012). In case of animals, mitochondrial DNA is designated as the barcoding material. Mitochondrial DNA is maternally inherited and it has higher mutation rate of base substitution compared to nuclear DNA. So, it is an efficient genetic marker in genetic differentiation studies (Qiongying T *et al.*, 2006). A portion of the cytochrome *c* oxidase 1 (COI) mitochondrial gene is used as the prime region for organism identification, taxonomic clarification and also for investigating phylogeographic groups within a single species.

The economic importance and identification challenges associated with fishes prompted the launch of an international Fish Barcoding of Life (FISH-BOL) initiative (http://www.fishbol.org/) with the aim of barcoding all fishes (Ivanova N *et al.*, 2007). It is a collaborative international research effort, which seeks to establish a reference library of DNA barcodes for all fish species derived from voucher specimens with

authoritative taxonomic identifications (Hanner R *et al*., 2005). Beside facilitating species identification, flagging potentially previously unrecognized species, and enabling identifications where traditional methods are not applicable, FISH-BOL will also provide a powerful tool for enhanced understanding of the natural history and ecological interactions of fish species. BOLD (Barcode of Life Data System) is a collaborative online corporation developed and maintained by University of Guelph in Ontario as the extensive DNA barcode libraries to help with resources for DNA barcoding community. BOLD is freely available to any researcher with interests in DNA barcoding. The BOLD database provides detailed information of COI-sequenced species which include the species name, voucher data, collection record, identifier of the specimen, COI sequence of at least 500 bp, PCR primer used to generate amplicon and trace files (Hubert N *et al*., 2008). Out of almost 30,000 fish species estimated in the world, barcodes for more than 10,000 fish species are currently recorded in the BOLD database (Wong LL *et al.,* 2011).

1.2 Current Studies

Nepal has at least 186 fish species, however unfortunately no species has been barcoded yet. Nevertheless, some medicinal plants have been barcoded and the data submitted to the BOLD. However, barcoding of many animals including birds, springtails invertebrates, skipper butterflies, blowflies, leaf beetles, nematodes, amphibians, ants, crustaceans, scuttle flies and many other have been conducted worldwide already and the projects are ongoing to date (Muchlisin ZA *et al*., 2013). In addition, DNA barcodes have been obtained for more than 10,000 species of fish all over the world and the COI sequences deposited in the BOLD online workbench and repository. DNA Barcoding have been accomplished for many fish species including Australian fishes, Australian sharks and rays, Canadian freshwater fishes, North American marine fishes, coral reef fish, Central American freshwater fishes, Indian marine fishes and Antarctic fishes (Kress WJ *et al.,* 2012). Currently, barcoding of some fishes of Rivers Koshi and SunKoshi of Nepal is also being carried out in Paul Hebert Centre of DNA barcoding and Biodiversity Studies (PHCDBS), Aurangabad.

1.3 Hypothesis

DNA barcoding helps in identification of known/unknown species regardless the geographical location, climatic condition and origin. In case of animals, COI gene is used as the comprehensive barcode region in order to discriminate species since it is well conserved within same species for molecular identification among many morphologically similar specimens. K2P distance which uses 2% distance threshold is

useful for barcode gap analysis. For the specimen to be belonged to particular species, less than 2% intraspecific divergence or above 2% interspecific divergence is needed, while when the variation increases i.e. among intergeneric or interfamilies species K2P value is greater than 2%. Besides, barcode sequences can also be used for phylogenetic analysis of the species by construction of Neighbor Joining, Maximum Likelihood or Maximum Parsimony trees, where close species are found to originate from single node and cluster together. Thus, we hypothesize that integration of DNA barcodes (partial cytochrome c oxidase subunit I sequences) into bioassessment protocols will provide greater discriminatory ability than just morphological identifications and this increased specificity could lead to more sensitive assessments of freshwater fishes from Nepalese water.

1.4 Objectives

1.4.1 Broad Objective

- To identify and catalogue the fishes from Pokhara Valley by performing their barcoding followed by phylogenetic analysis.

1.4.2 Specific Objectives

- To morphologically identify fishes and extract DNA from fin tissues.
- To amplify a specific region of the mitochondrial genome, Cytochrome oxidase subunit 1 (COI) by PCR and obtain the sequences of the amplicons.
- To upload the sequences in BOLD and submit to NCBI.
- To carry out COI gene based analyses, including nucleotide and protein profiling.
- To perform the phylogenetic analysis of fish species using multiple sequence alignment and tree-building tools.

1.5 Rationale/Justification of the study

Fishes are considered as good foods with respect to human health being rich source of proteins, vitamins and minerals; so are consumed largely in many forms. These fishes are also vital part of ecology and biodiversity. They need to be characterized for identification depending on geographical and ecological diversity. In order to preserve species diversity, which is in the verge of disappearance due to global climatic change and habitat destruction, it is essential to identify them correctly. Taxonomy has been used as the science of classifying living things according to the shared features since a

long time. But classical taxonomy falls short in the race to catalog biological diversity before it disappears. It necessitates a highly trained and judgmental specialist in order to distinguish subtle anatomical differences between closely connected species. This is tedious and time consuming. So, nowadays DNA barcoding is gaining popularity because it allows non-experts to objectively identify species – even from small, damaged, or industrially processed material. Barcoding allows comparing newly found species with the older ones in order to know about their relationship in genetic level.

1.6 Scopes of the study

This type of research carries tremendous scopes. Using the information from barcoding, classification of any animals or plants becomes lot more convenient. Barcoding would enable retail substitutions of target specimen and comparing this information to a species to be detected, assist in managing for long-term sustainability and improve ecosystem research and conservation. Since no any barcoding has been carried out for animals in our country, this research work can act as a pioneer for rest of the species to be classified authentically which are diverse in nature. The data uploaded in the BOLD can be very helpful to the researchers as well as students for expanding their knowledge, skills and implementing them for the unknown species identification.

CHAPTER 2

LITERATURE REVIEW

2.1 Molecular Taxonomy or DNA Taxonomy

2.1.1 Mitochondrial DNA

The genes encoding the ribosomal small subunit sequences, both of nuclear and mitochondrial origin are those with the broadest taxonomic coverage currently available. However, they are rather conservative genes, thus are not particularly useful for differentiating closely related species (Tautz D *et al.*, 2003). Protein coding genes and regulatory genes (control region) of mtDNA are used as markers for investigating intraspecies and interspecies genetic diversity. As mtDNA accumulates many base substitutions over a period of time, it provides comparative data for taxonomic, evolutionary and phylogenetic research (Kartavtsev YP *et al.*, 2009). One of the most quickly diverging, and thus very informative sequences, is the mitochondrial control region (Tautz D *et al.*, 2003). The vertebrate mitochondrial genome consists of a 16-19 kb of circular molecule, usually containing 37 genes encoding 13 protein-coding genes, 2 rRNAs, 22 tRNAs, and a variable control region (CR) or D-loop. Among mitochrondrial genes, the only 2 protein coding genes that occur in all eukaryotes are cytochrome c oxidase subunit I and cytochrome b. Mitochondrial DNA (mtDNA) is a useful tool in studies of phylogenetics, phylogeography, molecular evolution, and population and conservation genetics due to its relatively simple structure, predominant female inheritance, and high rate of evolution (Prosdocimi F *et al.*, 2011). Restriction fragment length polymorphisms (RFLP) was used at first for mtDNA variation studies. These studies set the stage for much work to pursue and were helpful in developing mtDNA as a molecular tool. When Kocher *et al.* (1989) published highly conserved primers that could amplify the DNA from a wide range of taxa by PCR, sequence analysis started to be focused on sequence analysis rather than RFLP (Ballard JWO *et al.,* 2004). The complete mitochondrial genome sequences have been reported for numerous vertebrates including many fishes (e.g., loach, carp, sea lamprey, cod, bichir, lungfish, coelacanth, dogfish etc.). The gene content and organization of fish mitochondrial genomes is quite conserved. This conserved characteristic facilitates their alignment and identification (Peng Z *et al.*, 2006). The mitochondrial genome of animals is a better target for analysis than the nuclear genome because of its lack of introns, its limited exposure to recombination and its haploid mode of inheritance (Saccone C *et al.* 1999). For molecular phylogenesis, Cytochrome oxidase 1 (Co-1), single mtDNA genes: Cytochrome b (Cyt-b) and 16s rRNA genes are popularly used as they are capable of analyzing between species up to family level. But single gene approach provides insufficient

phylogenetic data when applied above Order due to less information capacity and homoplasy effects (Kartavtsev YP *et al.*, 2009). Past phylogenetic work has often focused on mitochondrial genes encoding ribosomal (12S, 16S) DNA, but due to the prevalence of insertion and deletions (indels) their use in broad taxonomic analyses is restricted. The indels greatly complicate sequence alignments (Doyle JJ *et al.,* 2000). The 13 protein-coding genes in the animal mitochondrial genome are better targets because indels are rare since most lead to a shift in the reading frame (Hebert PDN *et al.*, 2003). Using various genes, for instance; Cytochrome b (Cytb) gene and Rhodopsin (rhod) gene, which show different evolutionary rates and genomic positions simultaneously, can increase barcode efficiency. Cytb has similar phylogenetic performance as COI gene. *Rhod* gene is an intronless teleost fish gene which provides quantitatively-equal inter-species identification labels of targeted nuclear PCR amplification products throughout its coding sequence. Both of these genes have been widely used for identifying fish species and resolving fish phylogenies (Sevilla R *et al.*, 2007). Focusing only on the mtDNA genome can raise some problems due to heteroplasmy, incomplete lineage sorting, possible events of hybridization and the fact that analyzing only maternal lineages can be misleading (Ballard JWO *et al.,* 2004). Furthermore, the peculiar nature of the mitochondrial organelle arise greater possibility of disagreement between the evolutionary histories and of the mtDNA genome and the species as a whole. Thus, it can limit their usage as a genetic marker in population and species research (Alexander LC *et al.*, 2009). The control region has not been proved capable of differentiating some species while the cytochrome b gene has shown a higher number of fixed, diagnostic characters and thus a major promising tool in the identification of these species. Sometimes, ribosomal mitochondrial genes are used as alternatives. They are easy to amplify and are abundant in databases. These genes are also reference of synapomorphies in loop regions (Ferri E *et al.*, 2009). However, studies on insects, birds and fish have proved that COI Barcodes can be a useful tool in correctly identifying species. They are useful in a group where taxonomic uncertainty still exists: delphinid cetaceans. Hebert *et al*. (2003) have argued that the existence of robust universal primers and the greater range of phylogenetic signal when compared to other mtDNA genes make this gene the ideal one (Amaral AR *et al.*, 2007).

The useful properties of 5’ COI as barcode gene for animals are summarized below:

i. It is present in all eukaryotes.
ii. It is relatively abundant in each cell being mitochondrial gene and can be recovered from suboptimal specimens.
iii. It contains enough sequence diversity to differentiate most animal species (exception: Cnidaria).
iv. It is short enough to be readily amplified and sequenced.

v. It can be amplified from diverse phyla with broad-range primers (Stoeckle M *et al.,* 2003).

Its application in molecular taxonomy has been criticized due to introgressive hybridization, mitochondrial pseudogenes in the nucleus, and the retention of ancestral polymorphisms. Nevertheless, species assignment failure rates do not typically exceed 5–10% according to Hebert and Gregory, 2005 (Carvalho DC *et al.*, 2011).

2.1.2 COI as Barcode Region

2.1.2.1 Cytochrome *c* Oxidase

Cytochrome oxidases are intinsic membrane metalloprotein complexes that contain haem iron, copper, zinc and magnesium and reduce oxygen to water as the terminal step in aerobic respiration (Saraste M, 1994; Brunori M *et al.,* 1987)**.** The two main classes of cytochrome oxidases are cytochrome *c* oxidases, and quinol oxidases. Cytochrome *c* oxidase is the terminal enzyme of the respiratory chain. It is essential for respiratory function because it irreversibly transfers electrons of the chain to molecular oxygen (Hocker JM, 1989). Cytochrome *c* oxidase activates dioxygen, the terminal electron acceptor of mitochondrial respiratory chain. So, it plays a crucial role in aerobic life. The enzyme catalyzes the one electron oxidation of ferrocytochrome c and the four-electron/ four-proton reduction of dioxygen to water (Brunori M *et al.,* 1987).

For cytochrome c oxidase, the overall reaction is:

$$4\ \text{ferrocyt c} + 4H^{+}_{N} + 4H^{+}_{N} + O_2 \ \Longrightarrow\ 4\ \text{ferricyt c}^{+} + 2H_2O + 4H^{+}_{P}$$

The cytochrome *c* oxidase complex is composed of only 13 individual protein subunits and is thus amenable to systematic evolution. Three subunits (I, II and III) of cytochrome *c* oxidase comprise the catalytic core of the enzyme and are all synthesized from mitochondrial DNA. The remaining subunits (IV, Va, Vb, VIa, VIb, VIc, VIIa, VIIb, VIIc and VIII) are synthesized from nuclear DNA found on a variety chromosomes (Herrmann PC *et al.,* 2003).

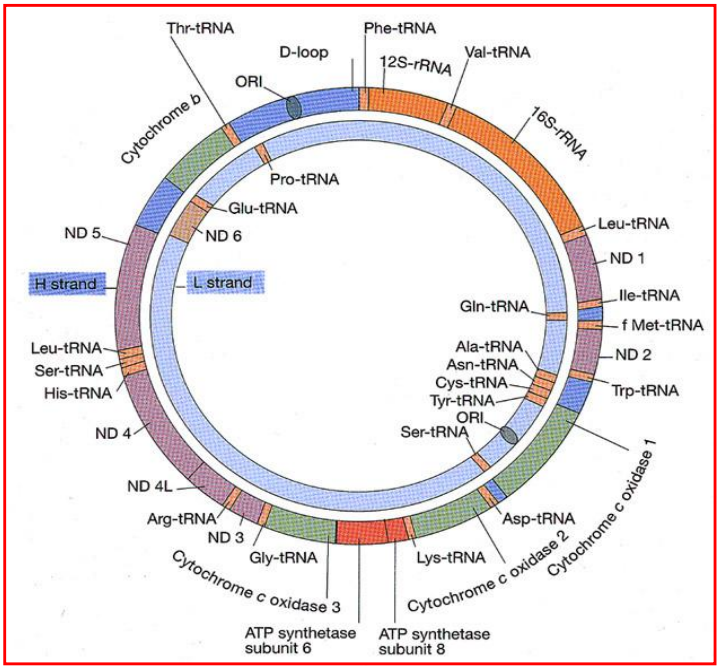


Figure 2.1: Mitochondrial DNA showing location of genes and other key regions.

(Source: www.google.com)

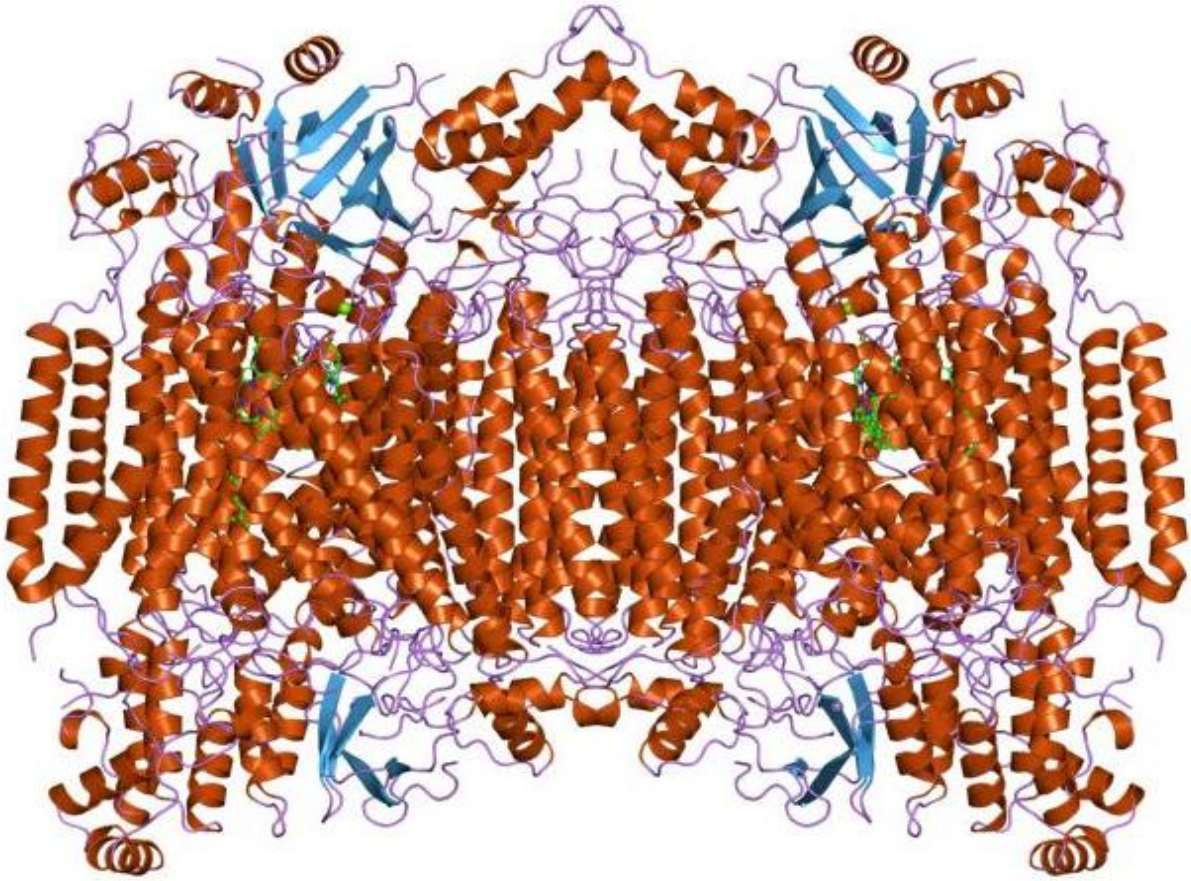
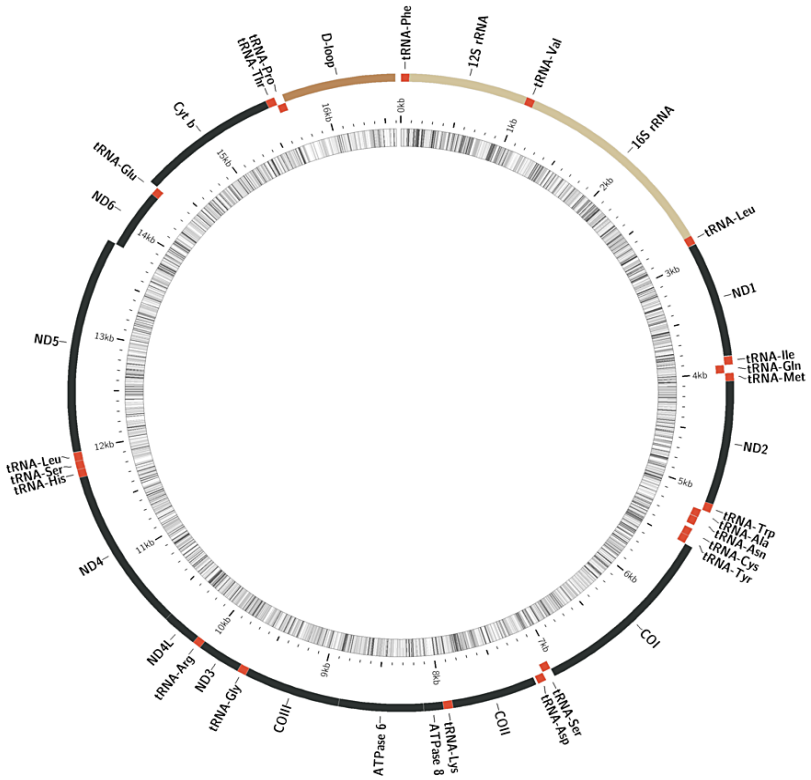


Figure 2.2: Cytochrome Oxidase

(Source: www.google.com)



GTGGCAATCACGCGCTGATTCTTCTCTACCAACCACAAAGACATTGGTACCCTTTATCTCGTATTTGGTG

VF2_t1-TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC

FishF2_t1-TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC

CCTGAGCCGGAATAGTAGGAACTGCCTTAAGCCTTCTTATTCGGGCCGAGCTAAGCCAACCCGGATCGCT
TCTAGGCGACGACCAAATTTACAATGTCATCGTCACTGCTCACGCCTTCGTGATAATTTTCTTTATAGTA
ATGCCCATCCTCATTGGAGGATTTGGAAATTGACTCGTCCCATTAATGATTGGGGCCCCAGACATAGCAT
TCCCCCGTATAAACAACATAAGCTTCTGACTTCTACCCCCATCATTCCTGCTACTACTAGCCTCTTCTGG
TGTTGAAGCTGGAGCTGGAACAGGATGAACGGTATACCCGCCTCTTGCAGGAAATTTAGCCCACGCAGGA
GCATCGGTAGACCTAACAATTTTCTCACTTCACTTAGCGGGTGTTTCATCAATTCTAGGGGCTATTAATT
TTATTACCACAACCATCAACATGAAGCCCCCAGCCATCTCACAATACCAAACACCTCTGTTCGTTTGATC
CGTGCTCGTAACCGCCGTACTGCTTCTTCTATCACTGCCAGTCCTAGCTGCTGGTATTACAATGCTCTTA
ACAGATCGAAACCTTAATACCACATTCTTCGACCCAGCAGGAGGGGGAGACCCAATTCTCTACCAACACT
TATTCTGATTCTTCGGCCACCCGGAAGTCTATATTCTTATTCTCCCAGGATTCGGTATCATTTCCCACGT

702AAGACTAAGAAGCCAGTGGGACTTCACAGTATCGACAAAGGAC-FishR2_t1

AARACYAARAAGCC TGTGGGACTCCACAGTATCGACAAAGGAC-FR1d_t1

CGTAGCTTATTACTCAGGTAAAAAAGAACCATTCGGCTACATAGGAATGGTTTGAGCAATGATGGCTATT
GGTCTCTTAGGATTTATCGTATGAGCCCATCACATGTTTACTGTTGGAATAGATGTAGACACTCGTGCAT
ACTTTACATCTGCAACAATAATTATTGCCATCCCAACTGGTGTAAAAGTATTTAGTTGATTAGCCACACT
TCACGGAGGCTCAATTAAATGAGAAACTCCTATGCTCTGAGCTCTTGGATTCATCTTCCTGTTTACAGTA
GGGGGACTCACAGGAATCGTATTATCTAATTCTTCACTTGACATTGTGCTTCACGATACATATTATGTAG
TTGCACACTTCCACTACGTACTATCAATAGGCGCCGTATTCGCTATTATAGCAGCCTTCGTGCACTGATT
CCCGCTACTAACTGGATATACCCTACACAGCGCTTGAACAAAAATCCACTTCGGAGTAATGTTCATCGGA
GTCAACCTCACATTCTTCCCACAACACTTCCTAGGTCTAGCAGGCATGCCACGACGATACTCCGACTACC
CAGACGCTTACGCCCTATGAAACACAGTATCATCTATTGGGTCATTAATTTCTTTAGTGGCAGTAATCAT
GTTCCTATTCATCTTATGAGAAGCCTTCGCCGCCAAACGAGAAGTATTATCTGTAGAACTAACTATAACA
AACGTAGAGTGACTTCACGGCTGCCCTCCTCCTTACCACACATATGAGGAACCAGCATTCGTCCAAATTC

AATCAAACTAA 1550

Figure 2.5: Mitochondrial DNA of *Cirrhinus mrigala.* The 1550 bp COI sequence (5486-7036) extracted from GenBank (Accession no. JQ838173.1) is shown above along with the indication of primer binding sites. Primers are highlighted yellow and the M13 tail in each primer is underlined. (Diagram source: Mitochondrial Genome Database of Fish)

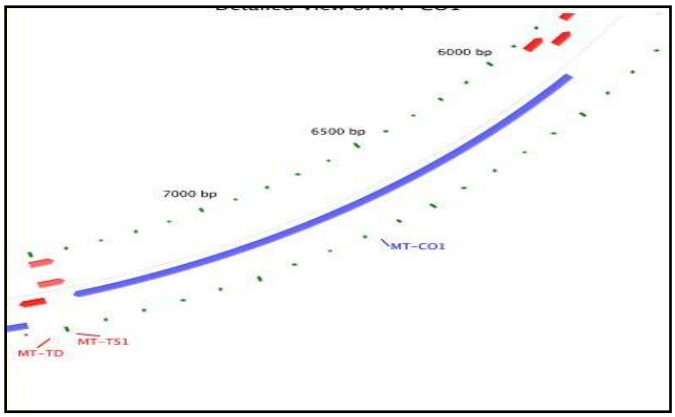


Figure 2.4: Mt-COI gene: a detailed view

(Source: http://ghr.nlm.nih.gov/gene/MT-CO1)

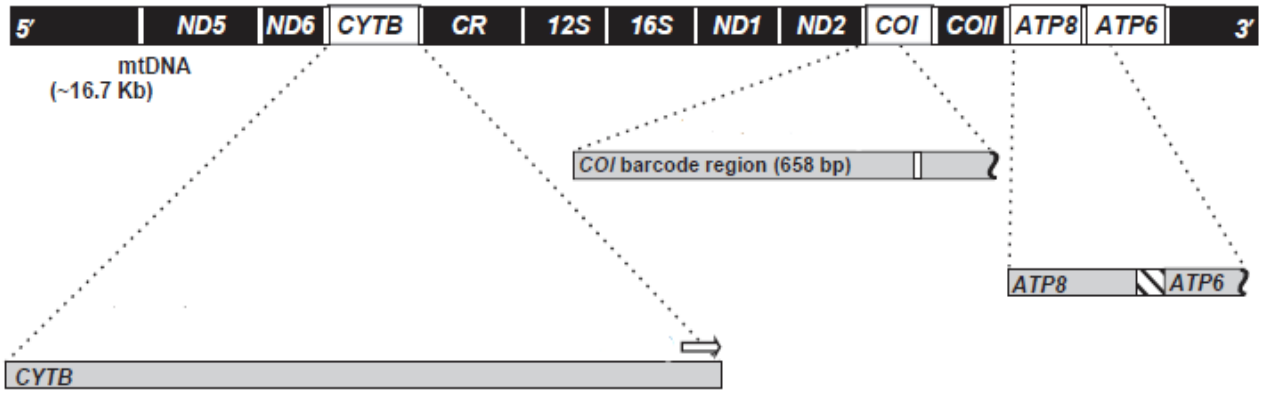


Figure 2.5: Schematic view of a linearized mitochondrial DNA showing the relative positions of most coding and noncoding regions (Chaves PB *et al.,* 2012).

2.1.2.2 COI gene

COI is a protein-coding gene, and as such, has an open reading frame. It is widely accepted marker for molecular identification to the species level across diverse taxa (Buhay JE, 2009). The mitochondrial DNA (mtDNA) cytochrome *c* oxidase subunit I gene (COI) has provided numerous examples as a reliable and universal tool for the identification of species such as the flatfish, tuna, anchovy, sharks, and also wildlife forensics investigations (Carvalho DC *et al.,* 2011). The COI gene is on average 2000–2200 nucleotides long (Lynn DH and Struder-Kypke MC, 2010). In vertebrates the total length of COI is about 1545 base pairs and a region about 650 bp long starting near the start of the cox1 reading frame is used as barcode globally (Ward *et al.,* 2007).

The cytochrome *c* oxidase I gene (COI) does have two important advantages. First, the universal primers for this gene are very robust, enabling recovery of its 5' end from representatives of most, if not all, animal phyla (Folmer *et al.* 1994). Second, COI appears to possess a greater range of phylogenetic signal than any other mitochondrial gene (Hebert PDN *et al.,* 2002). In COI gene, short polymorphic regions are flanked by highly conserved DNA priming sites making it easy to sequence in a wide range of taxa and thus fascinating as a standard locus for extensive DNA barcoding (Alexander LC *et al.,* 2009).

COI's third-position nucleotides show a high incidence of base substitutions, leading to greater rate of molecular evolution, about three times higher than that of 12S or 16S rDNA (Knowlton & Weigt 1998). As a matter of fact, the evolution of COI gene is rapid enough to allow the discrimination of not only closely related species, but also phylogeographic groups within a single species. Although COI may be matched by other mitochondrial genes in resolving such cases of recent divergence, this gene is more likely to provide deeper phylogenetic insights than alternatives such as cytochrome b because changes in its amino-acid sequence occur more slowly than those in this, or any other, mitochondrial gene (Hebert PDN *et al.,* 2003). From comparisons between the genome profile and the 13 individual gene regions, it is pointed out that the COI barcoding region is also representative of the efficacy of the mitochondrial genome as a whole of the twelve PCGs together (Elmeer K *et al.,* 2012). As in other mitochondrial protein coding genes, indels (insertions/deletions) are rare when COI is used, since most lead to a shift in the reading frame. They are, as a result, eliminated from the population (Pires AC and Marinoni L, 2010). Three criteria must be met at least to identify a gene region as appropriate for a DNA barcode: 1) significant species-level genetic variability and divergence; 2) short sequence length to facilitate DNA extraction and amplification, and 3) universal PCR primers. All these criteria have been found to be fulfilled by COI in the great majority of animal taxa (Elmeer K *et al.,* 2012).

2.1.2.3 Other Popular Barcoding Markers

a) Internal Transcribed Spacer (ITS): ITS is proposed as the standard barcode for fungi. The nuclear rRNA cistron, which consists of the 18S, 5.8S, and 28S rRNA genes, is popularly used for diagnostics and phylogenetics of Fungi. ITS region includes 5.8S gene and two spacers formed by splitting of rRNA cistron after post-transcription. It has the highest probability of successful identification for wide variety of Fungi because it gives clearly defined barcode gap between interspecific and intraspecific variation. Nearly 172,000 full-length fungal ITS sequences are deposited in GenBank at present. In some fungi, ITS region is also used to indicate delimitation by the measure of genetic

distances. It may also be as barcode for other organism beside fungi such as Chlorophyta, plants and Oomycota (Schoch CL *et al.,* 2012). ITS spacers from nuclear ribosomal DNA (nrITS) also represent the fundamental barcode in some parasitic plants with highly reduced genome (Hollingsworth PM *et al.,* 2011).

b) **18S Region:** The 18S nuclear ribosomal subunit rRNA gene (SSU) is used in phylogenetics but has less hypervariable domains (Schoch CL *et al.,* 2012).

c) **12S, 16S Region:** Most of the phylogenetics works in the past used to be focused on mitochondrial genes encoding ribosomal DNA, 12S and 16S. Because of the predominance of indels, their use is held back in broad taxonomic analyses these days (Hebert PDN et al., 2002).

c) **rbcL+matK:** The combination of rbcL (ribulose 1,5-biphosphate carboxylase/oxygenase) and matK forms a perfect plant barcode. Both rbcL and matK are portions of two plastid coding regions. The rbcL region consists of 599 bp region at 5' end of the gene, located at 1-599 bp, while matK barcode region consists of 841 bp at the center of the gene, located between 205-1046 bp in the complete Arabidopsis thaliana plastid genome sequence. Among the coding regions in plastid genome, matK is the most rapidly evolving. It is nearest to COI region of animal barcode in functioning. But, it can be difficult to amplify using existing primer sets especially in non-angiosperms. Unlikely, rbcL is easy to PCR amplify, sequence and align. But, it has comparatively low evolution rate. Thus, rbcL and matK, when used together can prove them as core barcode in plants due to straightforward recovery of rbcL region and discriminatory power of matK region (Hollingsworth PM *et al.,* 2011).

2.1.3 NUMTS (Nuclear mitochondrial pseudogenes)

Despite the haploid nature of mtDNA, non-identical mtDNA-like sequences may exist in one individual, and oftentimes they amplify with or instead of the target mtDNA (Schizas NV, 2012). Numts are copy of mtDNA that is integrated into the nuclear genome. They are also known as pseudogenes, homologs or paralogs. They vary widely among eukaryotes, with human and plant genomes harboring the largest repertoires (Antunes A and Ramos MJ, 2005). Numts come in many sizes, from all types of mtDNA sequence, and bear varying degrees of similarity to their mitochondrial counterparts. They typically occur in single copies at dispersed genomic locations. Numts arise both with and without

RNA intermediates. Their integration into the nuclear genome was originally associated with transposable elements or short dispersed repeats, but close examination of many different Numt loci reveals a lack of common features at integration sites (Bensasson *et al.,* 2001). These unusual mtDNA-like sequences have been found in protists, plants, fungi, and animals. Numts seem to be especially common in crustaceans, sea urchins, tunicates, and fishes and have been found more recently in sponges (Schizas NV, 2012).

Numts might be incorporated into the nuclear genome during the repair of chromosomal breaks by non-homologous recombination (Bensasson *et al.,* 2001). Numts are a major challenge in using mitochondria for DNA barcoding (Hazkani-Covo E et al., 2010). When non-specific primers are used, *Numt* sequences may be preferentially amplified because of the better matches between primer and pseudogene (Vallinoto M *et al.,* 2000). PCR ghost bands, extra bands in restriction profiles, sequence ambiguities, frameshift mutations, stop codons and unexpected phylogenetic placements are indications for mitochondrial pseudogenes. Sequence ambiguities result if the pseudogenes are at polymorphic sites or if they are encountered when sequencing from both strands. Increasing the proportion of amplified mtDNA can avoid Numts. This can be done by purifying mitochondria before DNA extraction, by long PCR amplification, or by using tissue that is rich in mtDNA relative to nuclear DNA like muscle (Bensasson *et al.,* 2001). However, such pseudogenes can be used as a powerful tool to estimate the relative evolutionary rates of mitochondrial genes. As these sequences evolve more slowly than their mitochondrial counterparts, and are thus generally more similar to the ancestral sequences, they can be used as outgroups in phylogenetic analyses (Vallinoto M *et al.,* 2000).

2.1.4 Indels

Insertions and deletions of nucleotides occur infrequently in coding region as they are strongly deleterious (Saitou N and Ueda S, 1994). Indels are difficult to model because little is known about their origin and the length of indels also needs to be dealt along with mutation rate. They are also difficult to handle because they are alignment dependent. Besides, they are often treated as alignment noise (Sjodin P *et al.,* 2010). Therefore, they are not well studied. Evolutionary distance(ED) calculated on the basis of indels results a very low increase of the distance over a long period of evolutionary time (Saitou N and Ueda S, 1994). ED is the per nucleotide site number of mutations occurred in the course of evolution of the sequences from their last common ancestors (Ogurtsov AY *et al.,* 2004). Insertion and deletions along with nucleotide substitution, gene duplication, unequal crossing-over and gene conversion are the types of mutations which are fundamental source for organismal evolution. It is vital to estimate the

spontaneous rate of each mutation type, including indels in order to overview the pace and mode of evolution at the nucleotide level. Besides, indels are constant in the course of evolution too (Saitou N *et al.,* 1994). Though indels are less common than single nucleotide mutations, they explain greater variation between species. Large scale indels are caused by the proliferation and illegitimate recombination of transposable elements, while short indels are generated by polymerase slippage. These both are very different from each other. There are varying views regarding the effects of indels. Some propose that deletions are more deleterious than insertions, while others argue that insertions are more deleterious as they increase the number of sites that can mutate into deleterious mutation (Sjodin P *et al.,* 2010).

2.2 Barcoding Databases: A brief Intro

2.2.1 Components of Barcoding projects

There are 4 constituents that comprise barcoding project which are explained below:

a. **The Specimens:** Specimens are very necessary elements in barcoding. They need to be collected sensibly for identification and well archived after barcoding. For more effective sample collection, cooperation with the taxonomists can be opted. By this it can be possible to put together a library of sequences that provides both broad species coverage and similar sampling intensity across species. After the collection, specimens should be preserved in cyanide or etanol or freezed. Formaldehyde, ethyl acetate should be avoided as they damage DNA. Long term storage can degrade DNA, so better freshly used. Cross-contamination should be prevented. Natural history museums, herbaria, zoos, aquaria, frozen tissue collections, seed banks, type culture collections and other repositories of biological materials are good sources for large amount of sample collection. However, hydrolysis and oxidation, exposure to ultraviolet light and preservation agents such as formaldehyde can degrade these specimens. In such cases, several short sequences can be amplified and then connected to generate a barcode.
b. **The Laboratory analysis:** There are established protocols from DNA isolation to sequencing of the barcodes which can easily be followed by the researchers. To produce a DNA barcode, a well-equipped molecular biology lab needs about $5 and a few hours only. As little DNA is required minute sample is sufficient for DNA isolation. For this, there are 2 types of protocols- DNA release and DNA extraction. DNA release protocol yields sufficient DNA for barcoding in case of fresh specimens. But for archival materials, DNA extraction methods such as PCE (Phenol/Chloroform extraction), CTAB can be more useful despite being time

consuming. PCR amplification of the isolated DNA depends on the primer used. Usually non-degenerate primers or inosine-based primers used along with optimized PCR can help in amplification of preferred barcode region only. Annealing temperature, concentration of dNTPs, magnesium, primers etc can be altered so as to yield sharp amplicon and eliminate the need of clean up. Various additives such as DMSO, trehalose can enhance amplification. Only 10 µl of reaction is sufficient for barcoding. Ethanol precipitation and magnetic bead protocols are widely used for reaction product clean up, also column-based methods. It reduces both reagent use and cost. Bidirectional sequencing is then carried out. It is better to edit sequences manually to avoid polymorphic sites and to maintain sequence quality. SEQUENCHER, SEQSCAPE, Codon code Aligner are some popular commercial software options which include features such as internal basecallers, automatic alignment, contig assembly and trimming of sequences. After obtaining the barcodes, the data are put in a database for further analysis.

c. The Database: The ultimate goal of the DNA barcode movement is the development of comprehensive barcodes for all lineages of eukaryotes. Thus, the construction of a public reference library of species identifiers for assigning unknown specimens to known species is the mostly prioritized thing in barcoding. The huge amount of barcode records generated from different barcoding projects need to be organized and analyzed, as well as easily searchable by sequence, species name or higher taxonomic groups.

There are currently two main barcode databases that fill this role:

- Barcode of Life Database (BOLD) - BOLD was created by University of Guelph in Ontario. It is a workbench of researchers where DNA barcode data can be collected, managed and analyzed. It includes 3 components- a laboratory information management system (LIMS), a data management and analysis system (DMAS) and a sequence identification engine. LIMS collect and store numerous barcode records required to maintain accuracy while tracking specimens passing through the multistep analytical chain. DMAS supports both storing and analyzing of barcode records. It allows work to proceed simultaneously in different labs, as such centrally managed improving communication and data loss or duplication. As a whole, it offers unambiguous traceability of the data stream back to the source, as DMAS includes information such as where the specimen was collected, where it is currently deposited, copies of sequence traces and photographs of specimen. BOLD-ID is the sequence identification engine which uses a combination of Local Alignment Search Tool (BLAST) and hidden Markov model based on a global protein

alignment for the COI gene. It comprises a simple user interface that allows COI sequences to be entered into a search field and automatically compared to the existing ones. The identification is confirmed by providing the photographs as well.

- The International Sequence Database Collaborative – It is the collaborative organization between GenBank in the Unites States, the Nucleotide Sequence Database of the EMBL (European Molecular Biology Lab) in Germany and the DDBJ (DNA Data Bank of Japan). All of these databases have agreed to CBOL's data standards for the barcode records.

d. The Data analysis: The Data Analysis Working Group of CBOL improves the ways that DNA barcode data can be analyzed, exhibited and utilized. This group has introduced the BOLD Portal which offers researchers newer and more pliable methods to store, manage, analyze and display their barcode data. In order to identify the specimens, the closest matching reference record in the database is found. Sequence records are automatically aligned. Distance-based Neighbor-joining tree can be exported by assembling species records as per requirement. (Hajibabei M *et. al.,* 2005; www.barcodeoflife.org).

The figure 2.8 in the next page illustrates each and every steps of how barcoding system functions. The flow chart comprises steps from specimen collection and studies, through DNA extraction and amplification to sequencing and data submission. The foremost stage in the process is the collection of specimens. The specimens are sub-sampled according to the necessity and then tissues are stored. It is not always necessary that the submission is always in the form of specimen, it might be DNA or PCR product or sequencing product. They are followed according to the requirements. If it is a tissue, it needs to be lysed for DNA extraction. If it is a PCR or sequenced product this step is skipped. After obtaining sufficient DNA, it is amplified using specific primers. Excessive DNA can be archived but needs to be frequently checked. The PCR is checked on agarose gels using UV. If the bands are good enough, they are taken for cycle sequencing and thermo cycling; otherwise PCR is repeated with some alterations whether in annealing temperature or the PCR mixture. Then, sequencing clean up is carried out followed by sequencing on DNA analyzer. After sequencing, editing of the sequencing on the basis of trace files is carried out both electronically and manually. If the data is validated after editing, sequence and trace files are uploaded to BOLD. Else it is reported as error and sent for investigation. The negative report may redirect for re-sequencing or even for re-amplification. The uploaded data needs to be passed from Project Manager for validation in BOLD. If accomplished, finally it can be published as well as submitted to GenBank. If not again it is reported as error and needs to be redone. Hence, it is a must

to monitor each and every step very cautiously as minor flaws can lead to wrong identification/interpretation.

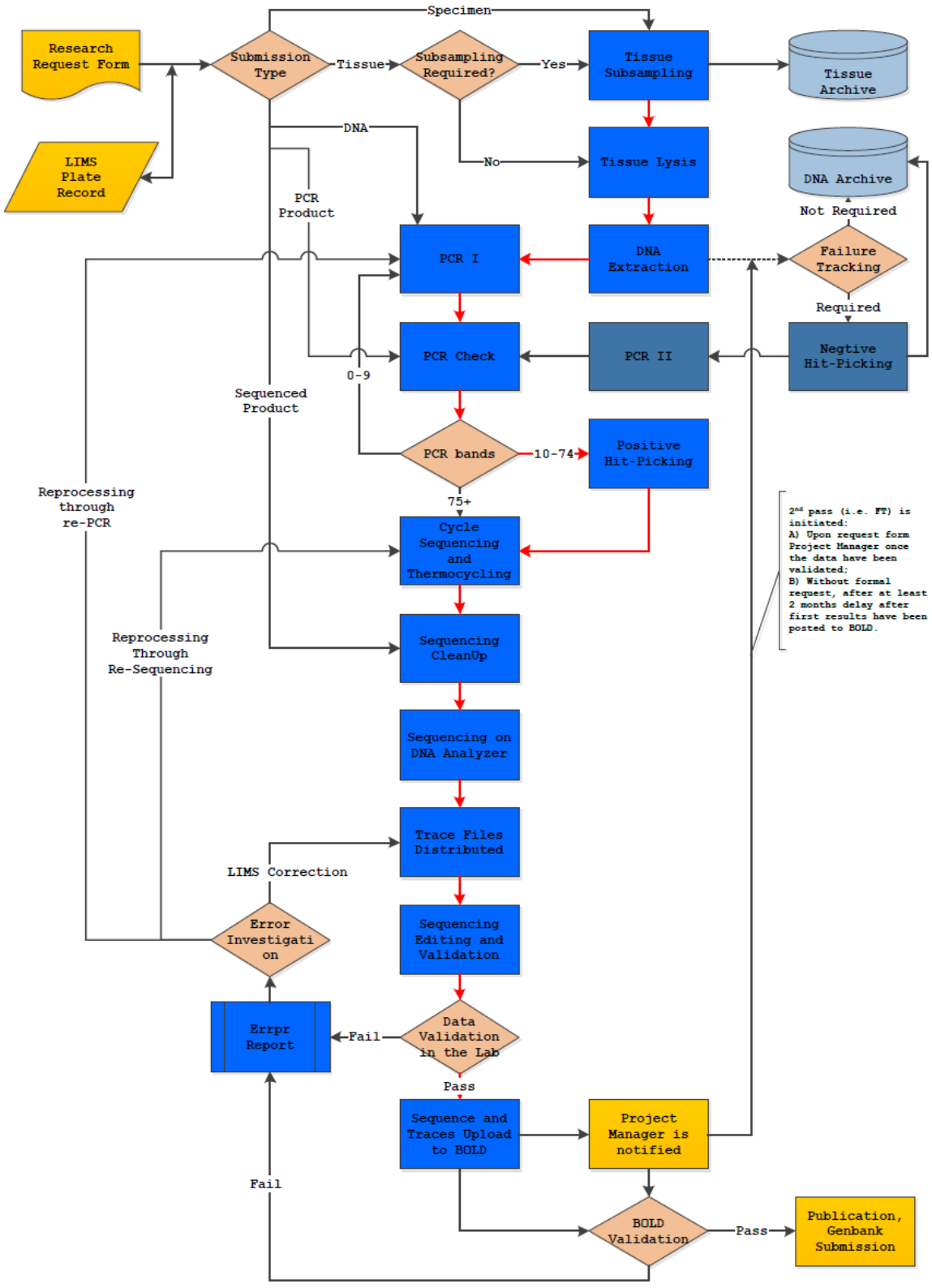


Figure 2.6: DNA barcoding workflow.

2.2.2 Fish BOL

The genetic barcodes can be stored in an open-access digital library that can be used to compare the DNA barcode sequences of unidentified samples from the field, garden, or market by matching them to known sequences with associated species names in the database. The Consortium for the Barcode of Life, CBOL (http://www.barcoding.si.edu/) is charged with systemizing barcoding activities around the world and promoting a database of documented and vouchered reference sequences to serve as a universal DNA barcode library for all life (Elmeer K *et al.,* 2012). CBOL was started in May 2004 and at present includes more than 120 organizations from 45 countries. It has been promoting to develop a barcode library for all eukaryotes by maintaining relationships with international researches. Since this project will generate enormous records, the library of the barcodes will also be very large. This resulted into the discovery of BOLD as a enterprise-scale software which upholds the novel barcoding aspects. CBOL is associated with the major genomics repositories (e.g. NCBI), biodiversity organizations (e.g. Global Biodiversity Information Facility (GBIF)), major barcoding centres and the multiple taxonomic communities to establish and strengthen data standards (Ratnasingham S *et al.,* 2007).

The Fish barcode of life initiative (FISH-BOL), launched in June 2005 (*Ward et al.,* 2009) is a global effort to aid assembly of a standardized reference sequence library for all fish species that is derived from voucher specimens with reliable taxonomic identifications. The fundamental task of FISH-BOL campaign is conducted by ten regional working groups representing Africa, Australia, Oceania/Antarctica, the Americas (North, Central and South America), Europe and Asia (India, North East Asia, and South East Asia) whose work is to supervise collections, identifications, and do the barcoding of fish fauna in their region (Swaetz ER *et al.,* 2008). The FISH-BOL campaign has adopted FishBase (www.fishbase.org) as the current global taxonomic authority file (Steinke D *et al.,* 2010). FISH-BOL has collaboration with catalog of fishes, Integrated Taxonomic Information System (ITIS), and FishBase to resolve an integrated checklist incorporating information from each of these sources. FISH-BOL uses the BOL Database (BOLD) as a workbench for assembling individual projects (Ratnasingham S *et al.,* 2007). BOLD offers a publicly available taxonomy browser (http://www.co1bank.uoguelph.ca/views/taxbrowser_ root.php) to aid taxonomical activities and to facilitate collaboration.

FISH-BOL has the primary goal of gathering DNA barcode records for all the world's fishes, about 31,000 species (Becker S *et al.,* 2011). By 2012, this campaign had barcoded about 8,000 fish species recognized for the cytochrome *c* oxidase subunit I (COI) gene (Kress WJ *et al.,* 2012). Currently, total specimens barcoded is 94836, out of

which there are 10185 individual species (www.fishbol.org/). With so many fishes that remain to be investigated, FISH-BOL will not only increase the profile of museum collections, but will also expand existing scientific networks and collaborations. FISH-BOL will attempt to barcode all fish on Earth. This is indeed an ambitious task but it will be realistic to barcode all the available samples currently in collections within a relatively short period of time (Swaetz ER *et al.,* 2008). The FISH-BOL campaign will build on the success of sister projects that focus on other taxa, namely the All Birds Barcoding Initiative (ABBI) and the All Leps (Lepidoptera) Barcode of Life campaign. The FISH-BOL project will be more challenging than the All Birds project, not only because fish are far more diverse but also because there is much less taxonomic information and expertise available. (Swaetz ER *et al.,* 2008). There are numerous benefits of barcoding fishes. It facilitates species identification for all users/taxonomists, resolves a previously unidentified specimen and enables the discrimination of species where traditional methods fail. FISH-BOL has a public resource in the form of electronic database which contains DNA barcodes, images, and geospatial coordinates of examined specimens, linkages to voucher specimens, information on species distributions, nomenclature, authoritative taxonomic information, collateral natural history information and literature citations (www.fishbol.org/).

2.2.3 FISHBASE

FISHBASE, the California Academy of Sciences' Catalog of Fishes and the Integrated Taxonomic Information System (ITIS) are major depositories for updated taxonomic and biological information on fish species worldwide. FISH BOL is currently using FISHBASE as its global taxonomic authority, but is also collaborating with Catalog of Fishes, ITIS and FISHBASE to incorporate their information into a resolved checklist for all fishes (Swaetz ER *et al.,* 2008). FishBase is the global most important Biodiversity information system on all fishes of the world, covering over 32,000 species. It records a wide range of information on all fish species currently known in the world about their biology, ecology, taxonomy, life history, trophic features, population dynamics and uses. FishBase provides also a range of country, regional, and ecosystem-specific information (http://www.worldfishcenter.org/fishbase).

2.2.4 BOLD

The BOLD (Barcode of Life Data Systems) data system is an informatics workbench that assists every stage of analytical pathways from collection of specimen to barcode library and is central to the DNA barcoding approach. The Barcode of Life Data Systems (BOLD,

www.boldsystems. org; Ratnasingham and Hebert 2007) is adopted by FISH-BOL. It provides an intricate platform for DNA barcode data storage, management, and includes species identification tools (Hanner R *et al.,* 2011). BOLD may also help to link the user to taxonomic expertise and give the country of origin easier access to biodiversity information. It provides a joined protocol for data acquisition, storage and analysis. The link between the barcode sequence, voucher specimen, image of the specimen and all the associated collecting and geographical information for future verification of species identity will add value to museum collections (Swartz ER *et al.,* 2008).

BOLD operates mainly in 3 ways: firstly, it is a repository for the specimen and sequence records that contribute to barcode library. Secondly, it is a workbench that helps to administer, assure quality and analyze the barcode data. Thirdly, it acts as a means of collaboration among various research communities which are geographically diverse by conjoining pliable security and data submission attributes with web-based delivery. BOLD has been designed specifically to link barcoding information to museum-registered specimens, so that taxonomic experts can go back to specimens to re-check identifications if necessary (Ratnasingham S *et al.,* 2007). This online resource offers tools that allow researchers to perform neighbor-joining clustering, to store information on the different groups studied, and to identify taxa using an updated sequence library, among other things (Pires CA et al., 2010). This is crucial for new records or even conflicting results for known species. It will also be possible to submit sequences to GenBank (NCBI) directly from this database, where they will hold the status of 'Barcode' to signify their quality and direct linkage of DNA sequences to museum collection data. Barcoding therefore holds the promise to increase the usage of collections, and raise the profile of collections at a time when many museums suffer from budgetary constraints due to inappropriate governance and reduced government funding (Hebert PDG, 2001).

After the barcode data records are ready for the public release, a copy of all sequences and key specimen data move to NCBI or other sister repositories (DDBJ, EMBL). BOLD in itself is an array of secondary sites which renders the biological science community with specialized services that can't be provided by the global sequence databases. The following 7 data elements of the specimen record are required to be entered to gain formal barcode status:

a. Species name
b. Voucher data
c. Collection record
d. Identifier of the specimen
e. COI sequence of at least 500 bp
f. PCR primers used to generate the amplicon

g. Trace files (Ratnasingham S and Hebert PDN, 2007).

Appendix 8 shows the glimpse of BOLD system which displays the pellucid view of how species or related information about it including images, sequence, trace file etc can be accessed easily and appendix 9 shows the specimen data sheet of the studied fishes which are uploaded in the BOLD.

Table 2.1: Common species level molecular markers. COI-barcode statistics are retrieved from BOLD. Statistics for other loci are retrieved from GenBank (Hajibabei *et al.,* 2007).

Gene[a]	Genomic location	Number of sequences			
		Animals	Plants	Protists	Fungi
COI-barcode[b]	Mitochondria	195 777	520	1931	410
16S-rDNA	Mitochondria	41 381	221	2059	285
cytb	Mitochondria	88 324	165	1920	1084
ITS1-rDNA	Nucleus	12 175	57 693	68 839	56 675
ITS2-rDNA	Nucleus	13 923	58 065	67 332	56 349
18S-rDNA	Nucleus	21 063	17 121	32 290	33 327
rbcL	Plastid	NA[c]	30 663	37 328	NA

2.2.5 Barcode Index Number (BIN)

BIN system is defined as the overall informatics system supporting the indexing, storage and retrieval of the OTU produced through the application of RESL. Operational taxonomic unit (OTU) is generated for the barcode sequence on BOLD using RESL algorithm. Refined Single Linkage (RESL) is an algorithm which computes barcode sequence records and enables ongoing adjustments in OTU boundaries. Each of the OTUs resulting from the analysis is assigned with a unique alphanumeric code with a standard structure (BOLD: 3 letters, 4 numbers). When a sequence data for barcode region is uploaded to BOLD, BIN pipeline analyzes it and the sequence that establish a new BIN add an entry to BIN index, whereas sequence assigned to existing BIN contribute their metadata to it. Appendix shows how BIN is presented as a single page that exposes the aggregate data for its members. A BIN data can b retrieved and downloaded for any taxonomic group. The key data elements off BIN include- taxonomy, distribution, images, sequence, and micro-attribution. BIN provides species level information needed to empower biodiversity science. Initially, BIN contains only the single record, which is joined through time by other sequences which match it or which show little divergence from it. A BIN boundary in a sequence space is more clarified by the addition of each new record to it. Thus, BIN system renders a vital identification service for the animal kingdom, where specialists are lacking for routine identification (Ratnasingham S *et. al.,* 2013).

2.3 Phylogenetics

Once the organismal barcode is generated, it needs to be read. Most recently published approaches to DNA barcoding have used distance measures to infer species affiliation. These include two frequently used methods—a simple BLAST approach and a tree-based genetic distance approach. These approaches generally use a raw similarity score to produce a nearest neighbor that is not necessarily the closest relative (Zhang AB *et al.,* 2008). Various alternative methods have been proposed to analyse DNA barcode data amongst which we can distinguish four main categories of approaches: (i) similarity approaches, based solely on the similarity between the total DNA barcode sequences or small parts of them (e.g. oligonucleotide motifs); (ii) classical phylogenetic approaches, using either genetic distances or maximum likelihood / Bayesian algorithms and assuming different mutational models (e.g. Neighbor-Joining, phyML, MrBayes); (iii) multiple-character based analysis ; (vi) pure statistical approaches based on classification algorithms without any biological models or assumptions (CAOS); and (v) genealogical methods based on the coalescent theory using demo-genetic models and maximum likelihood / Bayesian algorithms (Frézal L *et al.,* 2008).

A major shortcoming of using distances in DNA barcoding is that all classical studies and taxonomic schemes that accomplish the same thing that barcodes are meant to accomplish are character based, making the union of classical and DNA barcoding a difficult process if the use of distances is continued in barcoding studies. A second shortcoming is that similarity scores often do not give the nearest neighbor as the closest relative. A third shortcoming involves the lack of an objective set of criteria to delineate taxa when using distances (DeSalle R *et al.,* 2005).

An alternative approach including character based phylogenetic analysis is more appropriate for establishing or 'printing' barcodes. The character based approach is compatible with classical approaches allowing the combination of classical morphological and behavioral information. Character based approaches avoid the nearest neighbor problems of distances because they can reconstruct hierarchical relationships where common ancestry is inferred when two entities share derived characters (DeSalle R *et al.,* 2005).

Phylogenetic techniques are implemented in a Web based program that aligns a user-submitted gene sequence of unknown origin against a set of validated reference sequences, computes the evolutionary distances between the unknown and each of the reference sequences, and then builds a phylogenetic tree to display the affinity of the unknown sequence with the reference sequences (Ross HA *et al.,* 2003). Austerlitz et al. compared phylogenetic tree reconstruction with various supervised classification methods on both simulated and real data sets and found that maximum likelihood

phylogenetic always seem to be more accurate than distance based (Neighbor-Joining) phylogenetic inferences. But computation times are much higher for maximum likelihood phylogenetic reconstruction than for statistical classification. However, the accuracy of all the methods strongly depends on sample size and global variability of the taxa (Frézal L *et al.,* 2008). Neither BLAST (Altschul SF *et al.*, 1990) nor neighbor joining (Saitou N *et al*., 1987) tree building approaches allow for character-by-character diagnoses on branches of trees. Any such diagnosis would need to be Parsimony or Maximum Likelihood based (DeSalle R *et al.,* 2005). A typical molecular phylogenetics project involves a primary decision in relation to the target group for analysis (e.g. family), the assembly of representative taxa, the acquisition of sequence information, and the construction of phylogenetic trees by using optimality criteria such as Maximum Likelihood, Maximum Parsimony, or Bayesian analysis. It is important to emphasize that care must be exercised in the selection of both loci and representative taxa to optimize the recovery of a strongly supported phylogenetic tree (Hajibabaei M *et al.,* 2007). Consequently, most recent phylogenetic analyses use sequence information from multiple loci (covering several kilobases), often from different genomic compartments (i.e. nucleus, mitochondrion and chloroplast) to enhance resolution at different taxonomic levels and to avoid gene-specific biases (Hajibabaei M *et al.,* 2007). It is generally recognised that increasing the number of taxa aids recovery of the correct phylogeny by reducing branch lengths and homoplasy, both factors that can produce misleading phylogenies (Huelsenbeck JP, 1995). Researchers relied on heuristics and simplified analytical methods when dealing with phylogenies with large number of taxa (i.e. hundreds of species) (Hajibabaei M *et al.,* 2007). While barcode libraries have similarities to molecular phylogenetic data (both are sequence information from assemblages of species), DNA barcodes do not usually have sufficient phylogenetic signal to resolve evolutionary relationships, especially at deeper levels. Barcode sequence data can also provide a shared genomic cornerstone for the variable repertoire of genes that can be used to build the phylogenetic tree. It can be used as a link between the deeper branches of the tree to its shallow, species-level branches (Hajibabaei M *et al.,* 2006).

Phylogenetic tree of relationships is used for gene sequences comparison by researchers in diverse fields, including ecology, molecular biology, and physiology. Phylogenetic analysis of many gene families have been performed earlier, e.g. genes encoding: heat shock proteins, phytochrome, actin, transcription factors encoding gene MADS box genes (in plants) (Soltis DE *et al.,* 2003). The pattern of evolution of many morphological and chemical characters, including complex pathways such as nitrogen-fixing symbioses, mustard oil production, and chemical defense mechanisms have also been displayed by it. Evolutionary history of genes shows whether genes under investigation are the members of a single well-defined clade, all members of which appear to descend from a

recent common ancestor as a direct result of speciation (orthologous genes), or do the sequences represent one or more ancient duplications (paralogous genes) (Soltis DE *et al.,* 2003).

To begin phylogenesis, particularly in studies of genes from divergent taxa, it is necessary to align nucleotides and amino acid sequences that are at least homologous. Despite its fundamental importance, alignment remains the most difficult and poorly understood aspect of molecular data analysis (Soltis DE *et al.,* 2003). After alignment, it can be determined which position along the DNA or protein sequences are derived from a common ancestral position (Doyle JJ *et al.,* 2000).

Several methods of phylogeny reconstruction of molecular sequences such as maximum parsimony (MP), maximum likelihood (ML), distance-based methods such as NJ, and Bayesian inference (BI) have respective strength and weaknesses. Nonetheless, some measure of internal support (e.g. bootstrap, jackknife, and posterior probabilities) is also essential (Soltis DE et al., 2003).

2.4 DNA Barcoding and Population genetics

DNA barcoding is an initiative for species identification that overlooks sequence diversity in a 648 bp region of the mitochondrial gene coding for cytochrome *c* oxidase, subunit I (COI), a gene that plays an essential role in energy production (Lou M, 2012). DNA barcoding is not only used in conservation genetics and molecular ecology but also used in a number of other areas including forensic applications, population genetics and ancient DNA studies (Munch K *et al.,* 2008). Molecular phylogenetics and population genetics are the two branches of biology that have developed apparatus and applications employed to assess biological relationships with DNA sequences. Studies in molecular phylogenetics typically deal with evolutionary relationships among deeper clades, whereas those in population genetics target variation within and among populations of a single species (Hajibabaei M *et al.,* 2007). Numerous DNA based molecular techniques such as SSR, RAPD, AFLP, mt DNA have been used to find the population genetics relationships (Mu X *et al.,* 2012). Mitochondrial DNA markers are haploid and uniparentally inherited, so they are frequent targets for analysis and have made a particularly strong contribution to population-level studies (Avise JC, 2004). Population genetics studies examine variation within populations of a single species, and this sort of information has been successfully applied to geographical studies of populations, to investigate issues such as migration and genetic drift (Hajibabaei M *et al.,* 2007). The presence of different haplotype lineages may be explained by possible restricted gene flow due to the fragmented nature of freshwater ecosystems, which can include many physical and chemical barriers (Pereira LHG *et al.,* 2013). Various models

of population genetics have been proposed for the assignment of individuals to species in DNA barcode analysis; one of them being coalescent –based model (Abdo Z and Golding GB, 2007). Barcoding assignment methods can be divided into similarity methods based on the match between the query sequence and the reference sequences such as BLAST search, phylogenetic approaches, classification algorithms with no underlying biological models such as the nearest-neighbour method and methods based on population genetics (David O *et al.,* 2012).

There are many species identification approaches already with new ones being developed and performances among them have been explored (Ross, Murugan and Li, 2008; Austerlitz *et al.*, 2009; Parks, MacDonald and Beiko, 2011). Molecular taxonomic units (MOTUs) and evolutionary significant units (ESUs) are two of them for this purpose. They estimate diversity but fail to connect delineated units with known species (Blaxter *et al.*, 2005; Kizirian and Donnelly, 2004). In ecological niche modeling, environmental variables are identified and associated with the known distribution of a species, while in character-based methods, a unique combination of diagnostic characters are used to define a species. But the constant change occurring within species, reliance on a reference tree, and lack or subtlety of informative molecular characters may limit their use. However, distance-based, tree-based, or coalescent-based are the three classes of methods most accepted by the barcoding community (Lou M, 2012). A gap between intraspecific and interspecific variation is called barcode gap. In case of North American breeding birds, variation of cox1 sequences within species was found to be 20 times smaller than between species. Thus there was a clear gap. Utilizing this barcoding gap, a standard sequence threshold was proposed to define species boundaries of around 10 times the mean intraspecific variation for the group under study (Aliabadian M *et al.,* 2009). But the genetic distances and barcoding gaps variations are inadequate as they fail to consider species specific evolutionary rates. In phylogenetic or tree-based methods, the query belongs to the clade that it groups with. Coalescent method calculates the likelihood of coalescents for sequences known to originate from a particular species and then calculates the change in the likelihood when the query sequence is considered a member of this species (Abdo and Golding, 2007). It can be time consuming for data sets with a large number of sequences since it must generate enough coalescent trees to adequately sample all possible coalescent events (Lou M, 2012).

2.5 DNA Barcoding: Merits, Scopes and Challenges

2.5.1 Merits of DNA Barcoding

DNA barcoding has numerous benefits. It facilitates biodiversity surveys when large number of specimens from diverse taxa needs to be identified. Where traditional methods are unrevealing, barcoding enables the identification. Besides, even non specialists are able to use identifying tools fast, cheaply and reliably with more practical and fundamental applications (Radulovici AE *et al.,* 2010). It aids taxonomists to relieve the enormous burden of identifications, so they can focus on more pertinent duties such as delimiting taxa, resolving their relationships and discovering and describing new species. A large scale DNA barcoding effort will help to develop new techniques for DNA analysis, involving robust methods for DNA isolation from various specimens, and rapid and inexpensive sequencing techniques. It may attempt to resolve the phylogenetic relationships among all organisms by bringing each individual leaf into better focus (Stoeckle M *et al.,* 2004). Any person who has access to DNA sequencing, even if they lack taxonomic expertise can accurately identify species (Dasmahapatra KK *et al.,* 2006). It can identify species from even a small fragment. It works for all life forms from eggs and seeds through larvae and seedling to adults and flowers. It can differentiate among species that look alike, revealing dangerous organisms imposing as harmless ones and enabling a more accurate view of biodiversity. Barcodes provide an unambiguous digital identification feature, supplementing more parallel quantification of words, shapes and colors. DNA barcoding also provides bio-literacy tools for general public. Finally, once a comprehensive library is set up, it can enhance the public access to biological knowledge by creation of on-line encyclopedia of life on Earth, through which every species of plants and animals can be easily accessed along with vouchered specimens and their binomial names. Also any set of specimens could rapidly be discriminated and analyzed (Ramadan HAI *et al.,* 2012; Stoeckle M *et al.,* 2004).

2.5.2 Scopes of DNA Barcoding

DNA barcoding has quite expanded utilities in various fields. Barcoding is flourishing as a useful tool in diagnosing cryptic species which had previously been misidentified as single morphologically based species (Dasmahapatra KK *et al.,* 2006). Barcoding can be used to explore life cycles of any organism. Additionally, DNA barcoding will also facilitate basic biodiversity inventories. It also helps to reconstruct food webs by identifying fragments in stomach and many studies regarding it has already been carried out in several fish species. Plant physiology and soil science research can be done with this method by identifying roots sampled from soil layers. Biomedicines can make use of barcoding technique to verify the disease causing parasites and transmitters vectors. In

agricultural field too barcoding can help to determine the type of pests that's troubling the crops. It can be used to spot products prepared from certain species and pest species in imported goods. The trading of endangered species can be monitored and controlled by distinguishing them by molecular based technique like DNA barcoding (Ramadann HAI *et al.,* 2012). Thus, barcoding facilitates numerous applications including detection of putative cryptic species, identification of ambiguous life history stages, estimates shifts inspecies ranges, issues relating to tracking valuable/endangered species, analysis of food webs and trophic dynamics. In fisheries also DNA barcoding is proving to be beneficial regarding illegal fishing and fish fraud. DNA barcoding libraries of fishes constitutes a valuable resource for ichthyologists, fisheries biologists and other professionals as they require strictly reliable species identification on a routine basis and often for multiple species catches comprising various life history stages (Costa FO *et al.,* 2012).

2.5.3 Challenges of DNA Barcoding

Despite so many advantageous features, DNA barcoding is not untouched by some limitations. DNA barcoding identification system is based on a single character (~650 bp from 1st half of mt. COI gene) as a result the outcomes are sometimes unreliable and prone to errors. COI gene is not inherited as the nucleus located gene because it is located in the mitochondria which is maternally inherited. In case there occurred interspecific hybridization or infections (eg. Endosymbionts such as Wolbachia) which can transmit maternally, the mitochondrial genes can flow between biological species leading to different species identification rather than true one. This problem can be solved by supplementing nuclear barcodes along with mitochondrial barcodes. But nuclear loci evolve too slowly to be distinguished by barcoding and also they have intron regions with lots of insertions and deletions. They require cloning to obtain high quality sequence information from heterozygotes. Thus, it is challenging to find 600-1000 bp long nuclear protein coding region uninterrupted by introns, with high evolutionary rate to distinguish closely related species. There is high chance of misidentification, mislabeling, cross contamination between samples due to leaked DNA in ethanol jar with mixed samples or during amplification (Dasmahapatra KK *et al.,* 2006). Pseudogenes, contaminants amplified with universal primers or mitochondrial introgression can also be disturbing factors in barcoding success. Low resolutions in case of hybrids, recently diverged species, species complexes or slow evolving groups are troublesome at times. A new 'barcode-species' concept which will lead to an extreme amount of divergent clusters being recklessly raised to the species level, so called taxon over-splitting is of great concern. In addition, wide knowledge on reproductive isolation

biology of species in some cases, for instance marine animals, is necessary which is quite difficult to investigate (Radulovici AE *et al.,* 2010).

2.6 Status of Molecular Taxonomy in Nepal

Nepal is just stepping towards development of molecular techniques in several fields. And similar is the case with DNA barcoding. There is only one governmental organization, National Academy of Science and Technology (NAST) located at Khumaltar, Lalitpur which has been working in the related field. Various plant species have been identified using barcoding method in NAST, despite lacking the facility of sequencer. All the steps for barcoding are carried out at NAST and the cleaned up PCR products are sent to other labs outside the country for sequencing. One more organization which is non-governmental, Centre for Molecular Dynamics- Nepal (CMDN) situated at Kathmandu has also begun working in the field of DNA barcoding. They have recently barcoded tigers of Nepal, according to them but not public yet supposedly. Nepal is naturally very rich. There are uncountable valuable diversities of plants and animals which need to be correctly distinguished. But due to several technical, political and economical problems, establishment of well-facilitated DNA barcoding centre Nepal is still a long way to go.

CHAPTER 3

METHODOLOGY

3.1 Study Area/Sampling station

In this research, it has mainly been focused on molecular taxonomy of fishes from the rivers and lakes of Pokhara valley. So, sampling of fishes has been done from various water bodies in Pokhara including Begnas Lake and small rivulets. Samples were collected during the month of June. Selection of the sampling stations was done on the basis of fishing accessibility. The collected samples were brought to lab alive as far as possible but those which were already dead or died during travelling were dipped in absolute alcohol.

3.2 Collection of Fishes

The sample fishes were collected with the help of fishermen who used various types of nets for fishing and also some local boys who were using fishing rods for catching the fishes.

3.3 Photography

All the collected specimens were photographed with digital camera using scale on white paper sheet so that all the morphological characters were distinctly visualized as shown in figure below:

Figure 3.1: ***Mastacembelus armatus***

Figure 3.2: ***Oreochromis mossambicus***

Photographs were taken from dorsal, ventral as well as lateral view of fishes. The specimens were then given particular code which is used throughout the research.

3.4 Tissue Sampling

The soft tissues of pectoral fins and tail fins of specimens were cut using sterilized sharp scissors and forceps. After washing with 100% ethanol, the tissues were stored into 2 ml eppendorf tubes having absolute alcohol and labeled with specific codes. The tubes were stored at -20˚C for future use. Those tubes with pectoral fins were brought to Paul Hebert Center of DNA Barcoding and Biodiversity Studies, BAMU, Aurangabad, India. Before taking to the lab, ethanol was changed again so as to avoid the dilution of alcohol due to water resulting from the tissue dehydration. Same thing was done for the tail fins too. The sampled tissues which were brought to the lab were stored at -55˚C for further processing.

Fishes were preserved in absolute alcohol as voucher specimens in the jars so that they can be used in the relative works. For these whole fishes also once the ethanol was changed. The preservation was not done in formaldehyde as it is usually done because in DNA barcoding, formaldehyde can deteriorate the DNA (Ward RD *et.al*).

3.5 Fish Identification Methods

The fishes were identified relying on the book: “Fishes, Fishing Implements and Methods of Nepal” by Jeevan Shrestha. Also various reliable electronic databases like Wikipedia, Fishbase etc. were also used for identification purpose.

3.6 DNA EXTRACTION USING CTAB METHOD

The alcohol dipped samples were air dried on a tissue paper and transferred into the microfuge tubes each. The pre-warmed CTAB (600 µl), 3 µl of β- Mercaptoethanol and 10 µl of 20% SDS were added to each tubes. The samples were then crushed finely with the help of the scissors. After crushing, 3 µl of Proteinase K was added to each tube. They were then vortexed for 5 min vigorously. The tubes were incubated at 55˚C overnight. Vortexing was carried out at the interval of 2-3 hours after incubation (if possible). After cooling the samples to room temperature, they were centrifuged at 14,000 rpm for 10 min. Then equal volume of Phenol: Chloroform: Isoamyl alcohol (25:24:1) were added to each. After mixing them well by vortexing, they were centrifuged at 14,000 rpm for 10 min. Then the supernatants were taken in a fresh microfuge tube discarding the pellet. 600 µl of Chloroform: Isoamyl alcohol (24:1) were added in each tube and mixed well by vortexing. The samples were then centrifuged at 12,000 rpm for 10 min. The supernatants were taken in new tubes and the debris was discarded. Then 400 µl of chilled Isopropanol was added to the collected supernants and

mixed slowly until white flakes appeared. The tubes were then kept at -35˚C or -50˚C in deep freeze for an hour. The samples were brought to the room temperature and again centrifuged at 10,000 rpm for 10 min. After decanting off the supernatant, 400 µl of 70% chilled ethanol and 100 µl of Ammonium acetate were added to the pellet for washing. They were then centrifuged at 10,000 rpm for 10 min. Again the supernatants were decanted and 400 µl of absolute alcohol was added to each tube which was followed by centrifugation at 10,000 rpm for 10 min. The supernatants were decanted and the pellets were dried at RT. Lastly, the dried pellets were dissolved in 25 µl of TE.

3.7 Quantification of DNA

The quantification of DNA was done on Nanodrop ND1000 Spectrophotometer by using ND 1000 V371 software. Firstly, initialization of spectrophotometer was done by placing 1.5 µl of D/W or millipore water. Then, after swabbing it with tissue paper, 1.5 µl TE was kept on same pore to set blank. It was wiped again and 1.5 µl of sample DNA dissolved in TE was kept. Absorbance was taken at 260 nm for calculating the DNA concentration as given below:

DNA Concentration (ng/µl) = OD_{260} * 50

The ratio of OD at 260 and 280 nm was used to find out the purity of DNA and RNA or protein contamination too.

Good quality DNA	260/280 = ~1.8
DNA with RNA contamination	260/280 = >1.8
DNA with protein contamination	260/280 = <1.8

The ratio of OD at 260 and 230 nm was used to judge/check the contamination of Phenolic compounds.

Finally, DNA were diluted to the final stock concentration 100 ng/µl and stored at -20˚C for further use.

3.8 Qualitative analysis of DNA

1% agarose gel (Axygen) was prepared in TBE buffer to check the DNA samples by electrophoresis. For this, 0.5X TBE buffer stained with Ethidium bromide (500µg/µl) 3µl per 45 ml of gel was used. 3 µl of the DNA sample was mixed with 2µl of gel loading dye

(Himedia) and was loaded into the wells in gel. The samples were run for 15 min with constant current of 75 mA and were then visualized under UV trans-illuminator system (Gel Documentation System, Biorad, Inc. USA). Gel images were taken using Quantity One Software and saved for further use. The samples were used for further process based on the band quality of DNA in the gel.

3.9 PCR Amplification of Gene

Mitochondrial region Cytochrome Oxidase subunit-1 gene (COI gene) was selected for DNA barcoding as the standard gene. The universal cocktail primer set C_FishF1t1–C_FishR1t1 was selected for amplification of COI gene for fishes.

The reaction mixture for the Polymerase Chain Reaction (PCR) was composed as shown in the table:

Table: 3.1: PCR reagents composition and reaction volume

Particulars	Concentration	Volume/reaction
Nuclease free water (NFW)	-	17 µl
PCR reaction buffer (B)	10X	2.50 µl
$MgCl_2$	25Mm	0.4 µl
dNTPs	2.5Mm	2 µl
Forward Primer (FP)	10 pM	1 µl
Reverse Primer(RP)	10 pM	1 µl
Kappa Taq Polymerase	5 Units/µl	0.1 µl
Template DNA	100 ng/µl	1 µl
Total Reaction Volume		**25 µl**

Following programme was set up in the Thermal Cycler (ABI Verity, USA):

Table 3.2: PCR conditions for COI gene of fishes

Stage	Process	Temperature	Time	Cycles
I	Initial	94˚C	2 min	1 cycle
II	Denaturation	94˚C	30 sec	

	Annealing	52˚C	40 sec	35 cycle
	Extension	72˚C	1 min	
III	Final Extension	72˚C	7 min	1 cycle
Hold		4˚C	∞	

After the preparation of all the reaction mixtures, the PCR tubes were spun for 10 min at 100 centrifugal force to mix everything and bring each components together.

3.10 PCR Amplicon Check up

1.5% agarose gel was prepared in 45 ml of 0.5× TBE buffer along with 3µl of Ethidium bromide (500 ng /µl) in order to check the quality of PCR products. After electrophoresis for 10 min at 75 mA current along with a size standard or marker of 1 kb, the bands were checked on Gel Documentation system. A single band at about 650 bp length indicated positive PCR amplicons. Remaining ones which showed non-specific bands were discarded and were re-amplified using various PCR conditions such as lowering annealing temperature, increasing the template concentration, decreasing the $MgCl_2$ concentration etc.

3.11 PCR Clean up

The products obtained after PCR amplification were cleaned up in order to remove un-incorporated dNTPs and residual primers. Exo-SAP was carried out for cleaning up the PCR products. Omission of this step leads to degradation in sequencing results for the 50 or so bp. When the PCR product is put for bi-direction sequencing, such degradation is of little concern. But when the PCR product is sequenced in just a single direction, it is needed to clean up them as well as precipitate by ethanol washing.
For Exo-SAP, 0.25 µl of Exo-I and 0.5 µl of SAP with 1 µl of 10 × SAP buffer were added to a PCR tube and then 2.5 µl of the template i.e. PCR product was added which was followed by centrifugation. Then the mixtures were incubated at 2 different conditions - 37° C for 45 min and 80° C for 15 min to degrade the left over primers and nucleotides in the reaction mixtures and to inactivate the enzymes Exo-I and SAP respectively. The reaction mixture was then ready for the cycle sequencing.

3.12 Cycle Sequencing Reaction

The cycle sequencing reaction composition is as follow in the table 3.3 below (Hajibabei *et.al*, 2005).

Table 3.3: Cycle sequencing reagent concentration

Reagents	Concentration	Volume/Reaction
Ready Reaction Mix	2.5×	0.50 µls
Dilution Buffer	5×	1.75 µl
Primers	1.00 pM	2.00 µl
Milli Q (NFW)	_	4.75 µl
Template DNA		1.00µl
Final Volume		**10.00µl**

Cycle Sequencing Primers (Messing, 1983):

M13F (-21): 5'-TGTAAAACGACGGCCAGT-3'

M13R (-27): 5'-CAGGAAACAGCTATGAC-3'

As M13 tailed primers were used for PCR amplification, for cycle sequencing also M13 primers were used for high throughput sequencing.

The PCR condition for cycle sequencing is a shown in the table 3.4 below.

Table 3.4: Cycle sequencing PCR Reaction

Process	Temperature	Time	Cycles
Initial Denaturation	96 °C	3 min	1 cycle
Denaturation	96 °C	30 sec	35 CYCLE
Annealing	50°C	15 sec	
Extension	60°C	4 min	
Hold	4°C	∞	

3.13 Ethanol Wash of cycle sequenced products

Master mix I (MMI) and Master mix II (MMII) were prepared as:

Table 3.5: Master mix composition for Cycle sequencing product washing

MMI		**MMII**	
125 mM EDTA	– 2 µl	3M Sodium acetate (pH 4.6)	– 2 µl
Milli Q	– 10 µl	Absolute alcohol	–50 µl
Total:	12 µl/reaction		52 µl/reaction

12 µl of MMI and 52 µl of MMII were added to each cycle sequencing PCR product and kept at RT for 15 min. The tubes were then inverted several times before centrifuging at 5000 rpm for 40 min in a 24˚C cooling centrifuge. The supernatant was discarded at 100 g for 1 min. Then 100 µl of 70% ethanol was added to the tubes and centrifuged for 10 min at the same conditions. It was repeated for 3 times. After the final discard, the tubes were let open and kept at RT for an hour. When the tubes dried, they were checked for crystals.

3.14 Sequencing

In each tube, 15 µl of HiDye formamide was added carefully as it is highly hazardous to health. All the tubes were centrifuged for 1 min at 100g. Then they were snap-chilled. For Snap-chilling, tubes were placed in thermocycler set at 95^{o}C for 3 min. 3-4 sec before completion of set time period; the tubes were taken out and immediately kept in the ice-bucket for quick chilling to avoid reannealing of the DNA strands. The samples were then ready for sequencing. The sequencer machine (ABI, USA) having 4 capillary of 50 cm length was used. The readied samples were loaded into the wells finally for obtaining sequences bidirectionally.

3.15 DNA Sequence Alignment

The sequence trace files were assembled using Codon code aligner software along with the standard reference sequence. Using this program, ends were trimmed from the raw sequences referring to the standard sequence. After trimming, forward and reverse sequences for each specimen were assembled. Each assembled pair was examined and edited manually, and each sequence was checked for stop codons. The edited individual

contigs for each species were aligned with Muscle to produce consensus sequences representing each species. Finally the consensus sequence from each contig was aligned using Clustal W program and exported in a FASTA format.

3.16 Deposition of Data

The generated COI sequences were submitted to BOLD along with all the requirements needed such as specimen data sheet including species name, voucher data, collection record, identifier name, PCR primers used to generate the amplicon and trace files. All COI sequences were also deposited in GenBank.

3.17 Data Analysis

Sequence divergences were calculated using Tamura Nei distance model. To provide a graphic representation of it, the mid-point rooted NJ tree was created. Bootstrap values for NJ tree was estimated using searches with 1000 pseudoreplicates. To infer phylogenetic relationships between the sample species from matrix of sequences, ML and MP analysis were conducted using MEGA 5.2 (available from: www.megasoftware.net). The robustness of trees was assessed by bootstrapping 1000 times. The aligned sequences were also subjected for nucleotide BLAST search to verify the sequence similarity to previously identified COI fish sequences and to further strengthen our results. Ranking system was enforced to the sequences using BOLD. Similarly, barcode gap analysis was also conducted. Nucleotide and amino acid composition in the sequence data were analyzed including GC content and substitution pattern using MEGA tools. The percent identity and pairwise distance were also determined.

CHAPTER 4

RESULTS

4.1 Morphological Classification

In total, about 30 fishes were collected and out of them 14 species were identified. They were found to be belonging to 7 orders comprising 5 families, 13 genera and 14 species. The family with the greatest number of species was Cyprinidae (8), followed by Mastacembelidae (1), Clariidae (1), Bagridae (1), Belonidae (1), Channidae (1) and Cichlidae (1). All the fish specimens were of the class Actinopterygii.

Table 4.1: Family wise number of individuals studied.

Class	Order	Family	Individuals Studied
Actinopterygii	Synbranchiformes	Mastacembelidae	1
Actinopterygii	Siluriformes	Clariidae Bagridae	2
Actinopterygii	Beloniformes	Belonidae	1
Actinopterygii	Cypriniformes	Cyprinidae	8
Actinopterygii	Perciformes	Channidae Cichlidae	2

4.2 DNA Processing Results

After the morphological identification of fishes, they were digitally photographed for digital records. A complete database was then uploaded to Barcode of Life Data System. The pectoral fin tissue samples were then used for DNA extraction. The isolated DNA was examined on agarose gel for quality assurance as shown in figure 4.1.

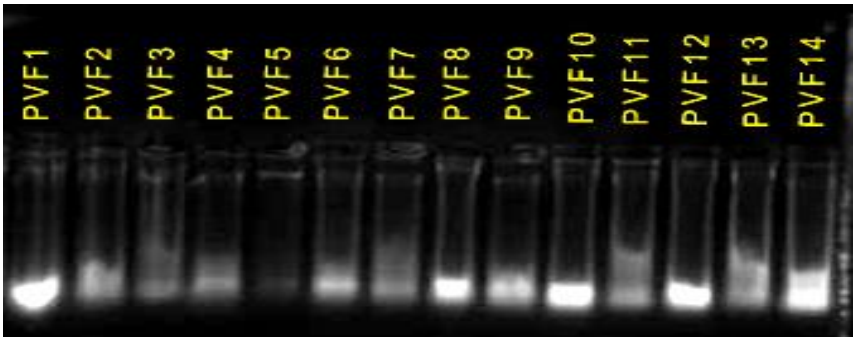


Figure 4.1: Quality check for Genomic DNA (1% agarose gel)

All DNA samples were then screened for amplification of COI gene with the fish cocktail primer set C_FishF1t1–C_FishR1t1:

VF2_t1 : 5'-TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC-3'

FishF2_t1: 5'-TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC-3'

FishR2_t1: 5'-CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA-3'

FR1d_t1 : 5'-CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAYCARAA-3'

The nucleotides which are underlined suggest the M13 tails (Ivanova et al., 2007). The figure 4.2 shows the PCR amplified product along with the standard DNA ladder of 1 KB.

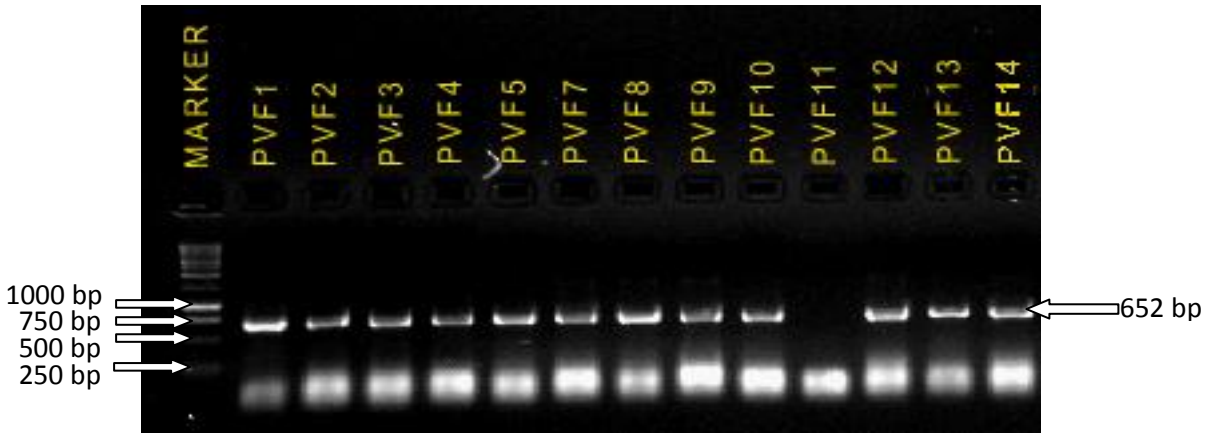


Figure 4.2: PCR amplified product along with standard DNA ladder of 1 KB.

The samples were sequenced and the sequences obtained were then compared with standard reference file using the software Codon code aligner. NUMTs and stop codons were also examined. Every single base of each of the sequence aligned was checked in the software carefully. The subsequent contig assemblies generated using the sequences had no stop codons and Numts.

4.3 COI Gene Profiles

Out of 14 specimens studied for COI based identification, 13 specimens' sequences were successfully obtained though the PCR amplification resulting good amplicons. Only one sequencing of Sample no. 6 failed. Table no. 4.2 portrays the fish species employed for DNA based studies in the research. The read length in majority (93%) was more than 600 bp long, while the remaining one sequence had 549 bp. So, the average sequence length achieved was >600 bp. Since there was no any degeneracy in the sequence anywhere in the middle, it can be confirmed that no any insertions/deletions were observed. The lack of stop codons is displayed by the fact that all the amplified sequences were functional mitochondrial COI sequences. All these sequences were larger than 600 bp in average, signaling the absence of NUMTs too. Besides, NUMTs do not appear to be a concern in

fish barcoding. Only those sequences with >500 bp size and of better quality were taken for phylogenetic analysis.

Among 652 bp considered for present analysis, 367 characters were conservative, 285 were variable and 218 were phylogenetically informative under parsimony.

Table 4.2: Details of fishes studied for their molecular taxonomy using DNA barcoding

Code	Order	Family	Name of Fish
PVF1	Synbranchiformes	Mastacembelidae	*Mastacembelus armatus*
PVF2	Siluriformes	Clariidae	*Clarias batrachus*
PVF3	Beloniformes	Belonidae	*Xenentodon cancila*
PVF4	Cypriniformes	Cyprinidae	*Tor putitora*
PVF5	Siluriformes	Bagridae	*Mystus cavasius*
PVF6	Cypriniformes	Cyprinidae	*Puntius conchonius*
PVF7	Cypriniformes	Cyprinidae	*Cirrhinus mrigala*
PVF8	Cypriniformes	Cyprinidae	*Chagunius chagunio*
PVF9	Perciformes	Channidae	*Channa orientalis*
PVF10	Perciformes	Cichlidae	*Oreochromis mossambicus*
PVF11	Cypriniformes	Cyprinidae	*Hypopthalmichthys nobilis*
PVF12	Cypriniformes	Cyprinidae	**Pethia ticto/Puntius ticto*
PVF13	Cypriniformes	Cyprinidae	*Barilius vagra*
PVF14	Cypriniformes	Cyprinidae	*Labeo rohita*

*Puntius ticto or Pethia ticto are synonymous names.

In accordance with the Fish BOL campaign, all sequences and collateral specimen information were deposited within the BOLD, where this data can be inquired by users, annotated and curated in light of new information. All the sequences have been deposited in GenBank also and accession number for the barcodes, specimen and collection data, sequence trace files and primer details are available within the DBFFN project files in BOLD. The respective BIN no. and NCBI accession numbers of all the submitted sequences are shown in table 4.3.

Table 4.3: BOLD ID and NCBI accession no. of the submitted sequences of the species studied.

Code	Name of organism	BIN	NCBI accession no.
PVF1	*Mastacembelus armatus*	BOLD:AAJ1660	KF742431
PVF2	*Clarias batrachus*	BOLD:AAM1926	KF742432
PVF3	*Xenentodon cancila*	BOLD:ABU9035	KF742433
PVF4	*Tor putitora*	BOLD:ACE8967	KF742434
PVF5	*Mystus cavasius*	BOLD:ABX1815	KF742435
PVF7	*Cirrhinus mrigala*	BOLD:AAE2831	KF742436
PVF8	*Chagunius chagunio*	BOLD:AAZ5053	KF742437
PVF9	*Channa orientalis*	BOLD:ACH0185	KF742438
PVF10	*Oreochromis mossambicus*	BOLD:AAA8511	KF742439
PVF11	*Hypopthalmichthys nobilis*	BOLD:AAN0845	KF742440
PVF12	*Pethia ticto /Puntius ticto*	BOLD:AAZ3076	KF742441
PVF13	*Barilius vagra*	BOLD:ACH0169	KF742442
PVF14	*Labeo rohita*	BOLD:AAC9904	KF742443

4.3.1 Species Identification

The resulting sequences were exported individually and nucleotide BLAST was carried out to find out how much the species were similar to those which have been submitted to GenBank. This aided in confirmation of morphological identification. The table no. 4.3 depicts the similarity % of the specimen to its closest relative in GenBank, obtained by doing BLAST and received from the BOLD-IDS search engine.

Table 4.4: Percentage similarity of COI gene of the specimen obtained from BLAST and BOLD

Sample Code	Name	% Similarity of COI from BLAST	% Similarity of COI from BOLD
PVF1	*Mastacembelus armatus*	99	98.5
PVF2	*Clarias batrachus*	99	98
PVF3	*Xenentodon cancila*	97	98.5
PVF4	*Tor putitora*	100	100
PVF5	*Mystus cavasius*	99	99
PVF7	*Cirrhinus mrigala*	99	98
PVF8	*Chagunius chagunio*	99	99.5
PVF9	*Channa orientalis*	97	98
PVF10	*Oreochromis mossambicus*	99	98
PVF11	*Hypopthalmichthys nobilis*	100	98.5
PVF12	*Pethia ticto/Puntius ticto*	99	99
PVF13	*Barilius vagra*	98	97
PVF14	*Labeo rohita*	99	97.5

4.3.2 Grading for Taxonomic Reliability

There exist chances of disagreements in the DNA barcodes arrays due to various reasons such as taxonomic uncertainty and operational shortcomings. To find out whether the sequences we have submitted to the BOLD are reliable or not, we can rank them using BIN Discordance analysis tool from BOLD. The species' sequences were queried in order to find if DNA barcode sequence data from multiple independent observers produce congruent and unambiguous matches for the given species. Out of 13 sequences, 6 were concordant with other barcodes clusters in the same BINs, 6 were discordant and 1 of the sequence was the only one present in the BINs. But the result was reviewed in the BIN page separately for all the species one by one. After this, 76.92% of the species were found to be concordant externally. 7.96% of the species were discordant, while for 15.38% of the species no sufficient data was available in BOLD. Grade A signifies concordant species, grade E signifies discordant species and grade D signifies singletons. The table below shows the rankings of our studied species.

Table 4.5: Attribution of grades (A to E) to DNA barcodes of 13 fish species

Species	Grade	Species	Grade
M. armatus	D	*C. orientalis*	A
C. batrachus	A	*O. mossambicus*	A
X cancila	A	*H. nobilis*	A
T. putitora	A	*P. ticto*	A
M cavasius	A	*B. vagra*	D
C. mrigala	E	*L. rohita*	A
C. chagunio	A		

4.3.3 Selection of the Best Fit Model for Analysis

The hierarchical likelihood ratio tests indicated that the best-fit model for subsequent analysis was Tamura-Nei model with invariable sites and gamma shape parameter (T93+G+I). The evolutionary model T93+G+I was employed for each gene and for the concatenated genes data set to describe the substitution pattern the best on the basis of lowest BIC scores (Bayesian Information Criterion). A discrete gamma distribution (+G) with 5 rate categories was used to model the non-uniformity of evolutionary among sites, also by assuming that a certain fraction of sites are evolutionarily invariable (+I). This Tamura-Nei model justified the use of 5 parameters for maximum likelihood analysis.

4.3.4 Nucleotide Composition

The nucleotide sequence analysis based on the mtDNA COI sequence showed the domination of C:T as a whole. Considering all the codon position, average nucleotide frequencies in cox1 nucleotide of all the taxa were 27.47% cytosine, 25.60% adenine, 17.55% guanine and 29.42% thymine. The different nucleotide composition at all the codon positions ranges from 8.1% guanine at 3rd codon position to 42% thymine at 2nd codon position. In the first codon position, the proportion of guanine (30.5%) is highest and thymine (19%) is lowest. Similarly, in the second codon position, the rate of thymine is highest (42%) and guanine proportion (14.1%) is least. Finally, in third codon position the proposition of adenine (36.3%) is highest whereas, the proposition of guanine (8.1%) is least. There was G bias at 3rd codon position with the average of 8.1%, whereas in 1st codon position and 2nd codon position G was 30.5% and 14.1% respectively.

As expected, the rate of synonymous substitutions (dS) was much higher than the rate of nonsynonymous substitutions (dN). As all the dN-dS values were negative, it signified the absence of overabundance of nonsynonymous substitutions. Figure 4.3 illustrates the synonymous and on-synonymous substitution per site.

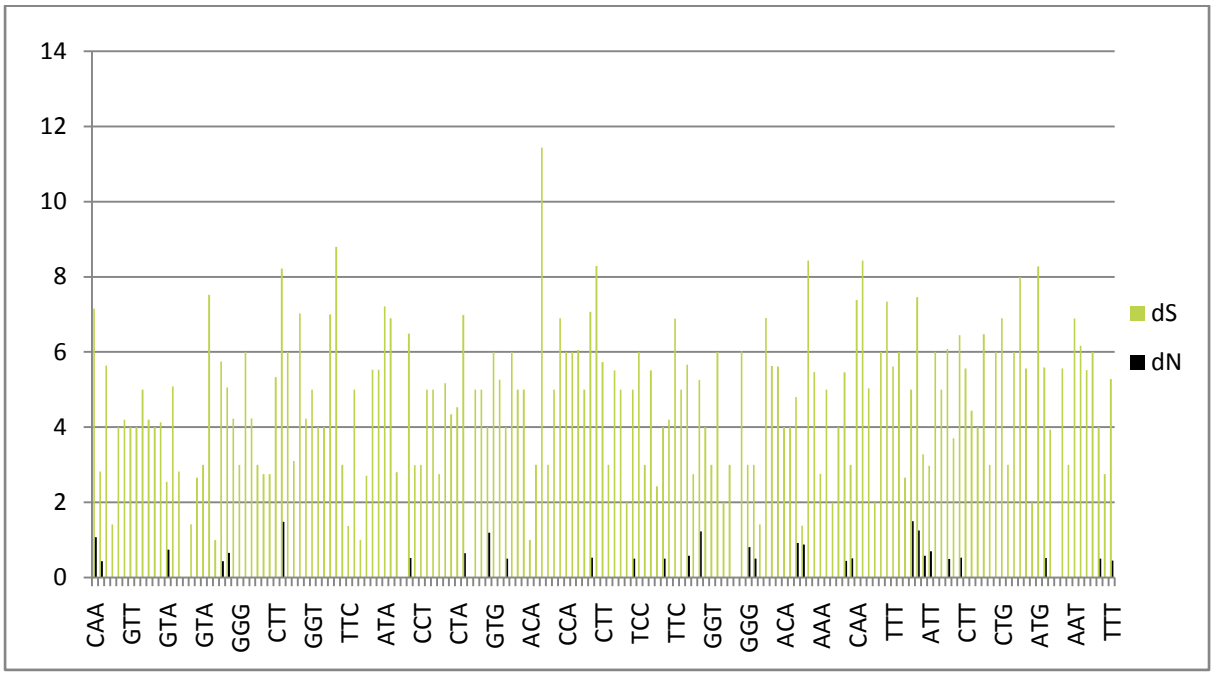


Figure 4.3: Synonymous and nonsynonymous substitutions per site

4.3.4.1 GC Content

Here GC content is considered. The overall GC content in the sequences was found to be 44.97. At 1st codon position, it was highest i.e. 56.01%, while least at codon position 3 with 35.58% of GC as shown in table 4.6.

Table 4.6: Nucleotide composition (%) of the COI sequences under study (13 sequences)

	Min	Mean	Max	SE
G%	16.17	17.52	18.88	0.218
C%	26.07	27.45	29.23	0.282
A%	22.50	25.60	27.80	0.383
T%	27.45	29.41	31.75	0.330
GC%	43.10	44.97	48.11	0.393
GC% Codon Pos 1	53.83	56.01	58.90	0.491
GC% Codon Pos 2	42.08	42.83	43.71	0.127
GC% Codon Pos 3	32.03	35.58	46.80	1.148

The Appendix 4 shows the composition of GC in the certain frequencies of species studied. For instance, in case of 1st bar diagram, 30% of the species have GC content of 44-45%, 15% of the species have GC content of 46-47%, and approximately 8% of the species had GC content of 49%. Similar type of interpretation can be done for rest of the diagrams 4.5 below.

4.3.5 Amino acid Composition

The composition of aminoacids in the sequences are shown in the bar graph below (Fig.4.5). On average, 210 aminoacids were obtained after translation of the nucleotides. The amino acids frequencies of all the species ranged from 0.55% lysine to 16.167% leucine. In the same way the concentration of alanine, aspartic acid, glutamic acid, phenylalanine, glycine, histidine, isoleucine, lysine, methionine, proline, glutamine, asparagine, serine, threonine, valine, tryptophan and tyrosine were 10.39%, 3.072%, 0.914%, 6.22%, 9.071%, 1.83%, 7.5%, 5.08%, 4.43%, 7.06%, 2.23%, 1.54%, 6.22%, 6.33%, 7.02%, 2.23% and 2.16% respectively. One amino acid cysteine was not found to be in any protein sequences of COI gene of the fishes.

4.3.5.1 Amino acid Variability

Multiple sequence alignment of the nucleotide as well as protein sequences was carried out using a new MSA tool applying HMM-HMM profiles, suitable for medium-large alignments. The aligned sequences of the mt. COI region was obtained which was then used for carrying out various analysis such as phylogenetic analysis to assess shared evolutionary origins, protein families characterization, identification of shared region of homology and determination of consensus sequence of several aligned sequences. The MSA of the nucleotide generated using Clustal W is shown in the Appendix 6, while that

of protein is given in the pages to follow. The star (*) represents the conserved sites and the dot (.) represents the variable sites in the sequences. The variable sites detected in the fragment of COI region are marked with green color and the sites which varied despite being conserved are highlighted red.

There were 174 conserved sites, 43 variable sites, 14 sites were most parsimony informative and 29 sites were singleton among the 210 analyzed. Around the positions 24, 27, 42, 188 and 225 the regions were most variable. Site 24 was dominated by Alanine, an aliphatic amino acid but is substituted by other aliphatic amino acid Valine frequently and a hydroxyl amino acid Serine in one case. Position 27 is dominated by Glycine but Arginine, is also present at times along with Serine. At position 42, Serine is most common, which is substituted by Alanine and Asparagine rarely. Position 188 is dominated by an aliphatic amino acid Valine, but is substituted by other aliphatic amino acids, Isoleucine and Leucine on two occasions. At 225th position, Glycine is most dominant and is substituted by Proline and Arginine each in two protein sequences.

Out of 16 positions within the sequenced region which are known to have important functions, only 12 were completely conserved over all fish species. As it can be seen, Tyrosine at 19 is not conserved in *M. cavasius* and *X. cancila*. Tyrosine, an aromatic amino acid is substituted by other amino acids Leucine (aliphatic) and Methionine (hydroxyl). Position 40 should be Glutamic acid, an acidic amino acid conserved but Lysine, a basic amino acid is present in case of *C. batrachus*. Similarly, position 50 is Aspartic acid conserved site but is replaced by Asparagine, a hydroxyl amino acid in *C. orientalis.* Position 221 is a conserved site for Aspartic acid in fishes but is found to be substituted by a basic amino acid, Lysine in *C. orientalis,* while in *O. mossambicus* there is presence of an unspecified amino acid. This kind of changes in the conserved sites can hamper the metabolism in fishes.

4.3.6 Transition/Transversion Bias

The average evolutionary divergence over all sequence pairs was found to be 0.230 (SE. 0.015). It shows the number of base substitutions per site from averaging over all sequence pairs with the 1000 bootstrapping procedure. Each entry shows the probability of substitutions from one base in rows to another base in the column. The rates of different transitional substitutions are underlined and those of transversional substitutions are without underline as shown in table 4.7 below. The transition/transversion rate ratios are k_1= 2.371 (purines) and k_2= 2.991 (pyrimidines). The overall transition/transversion bias was R= 1.429, calculated using the following formula: $[A*G*k_1 + T*C*k_2]/[(A+G)*(T+C)]$

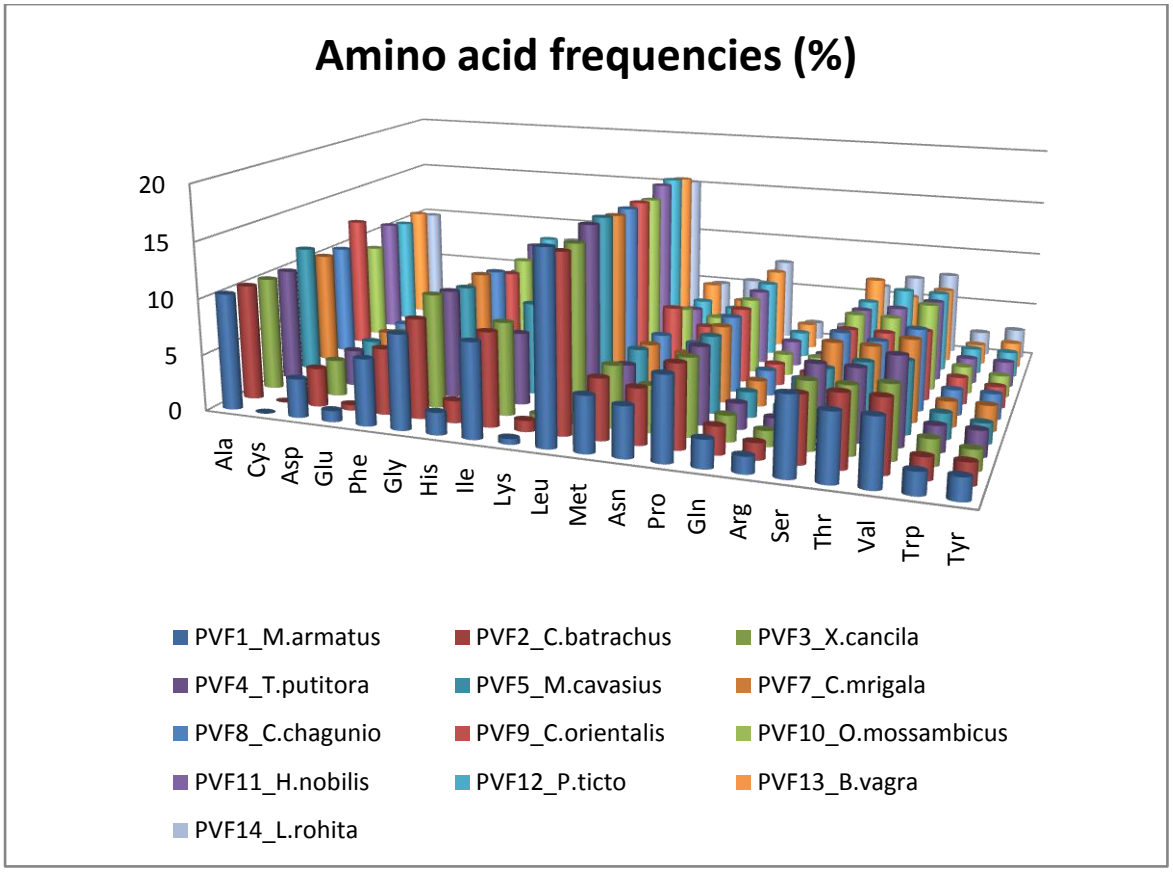


Figure 4.4: Amino acid composition

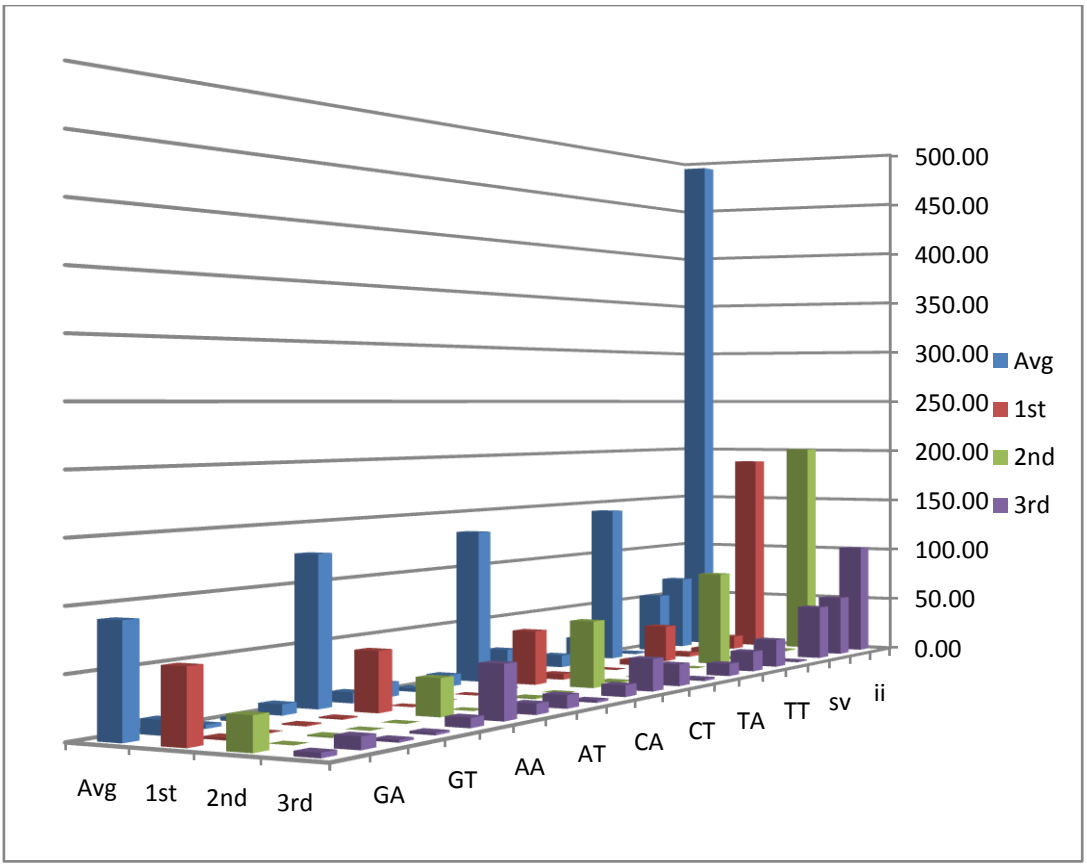


Figure 4.5: Nucleotide frequency of all the taxa studied.

Table 4.7: MCL transition / transversion bias.

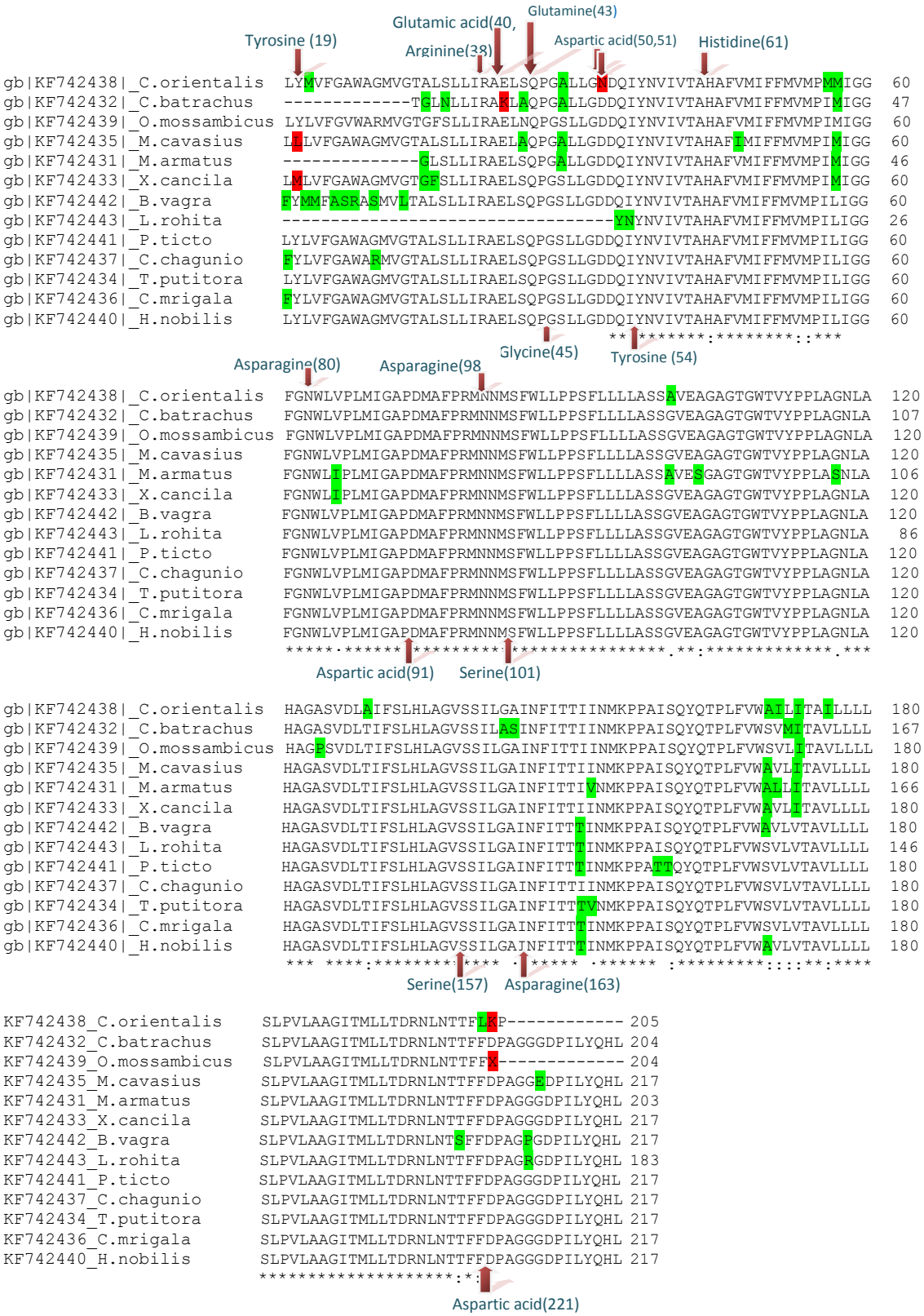
From/To	A	T	C	G
A	-	6.5	5.74	7.99
T	5.54	-	17.17	3.37
C	5.54	19.43	-	3.37
G	13.13	6.5	5.74	-

As shown in the fig. 4.6, the nucleotide substitution at codon position 2nd is lowest of all. The si/sv ratio was only 0.62. At codon position 1st, the substitution was higher than 2nd while less than 3rd with the si/sv ratio 3.43. The highest nucleotide substitution was found at 3rd codon position and si/sv ratio was 1.12. The transition/transversion ratio shows that most transition occurred at 1st codon position and least at 2nd codon position, while most transversion also occurred at 3rd codon position and least at 2nd codon position again.

The nucleotide frequencies within all taxa were counted as displayed in the table below. The frequencies varied to greater extent. In 1st codon position, GG content was 59 while TA, CA, CG, AT, AC, GC pair was zero. The concentration of TT was highest of all i.e. 85 at 2nd codon position whereas, TC, TA, TG, CA, AT, AC, AG, GT and GA was not present. On the contrary, at the 3rd codon position, all pairs of nucleotide were present with AA (45) being highest and GA (1) being the lowest. On average, TT (143) was present more commonly followed by CC (133), AA (126) and GG (90). The least occurring were TG, CG, GT and GC overall as illustrated in the figure 4.7.

Table 4.8: Nucleotide frequency at various positions.

	TT	TC	TA	TG	CT	CC	CA	CG	AT	AC	AA	AG	GT	GC	GA	GG
1stPos	33	5	0	1	5	47	1	0	0	0	50	1	1	0	2	59
2ndPos	85	0	0	0	1	58	0	1	0	0	31	0	0	1	0	27
3rdPos	25	18	11	2	19	28	10	2	11	9	45	8	2	2	1	4
Avg.	143	23	11	3	24	133	10	3	12	9	126	9	3	3	12	90


CLUSTAL O(1.2.0) multiple sequence alignment
Tyrosine (19)
Glutamic acid(40,
Arginine(38)
Glutamine(43)
Aspartic acid(50,51)
Histidine(61)
Glycine(45)
Tyrosine (54)
Asparagine(80)
Asparagine(98
Aspartic acid(91)
Serine(101)
Serine(157)
Asparagine(163)
Aspartic acid(221)

4.4 Pairwise Distances

Pairwise distances of COI gene are shown in Table 4.8. The bootstrap consensus tree inferred from 1000 replicates was taken to represent the evolutionary history of taxa analyzed. All positions containing gaps and missing data were eliminated. There were a total of 509 positions in the final dataset. The pairwise distance of COI sequences among the 13 fish species revealed the shortest genetic distance (0.150) between *Tor putitora* and *Cirrhinus mrigala.* The longest genetic distance (0.304) exists between *Channa orientalis* and *Oreochromis mossambicus*.

Table 4.9: Pairwise distance of COI sequences of fishes.

Taxon	1	2	3	4	5	6	7	8	9	10	11	12	13
M. armatus	-												
C. batrachus	0.244												
X. cancila	0.233	0.208											
T. putitora	0.219	0.236	0.256										
M. cavasius	0.262	0.236	0.233	0.253									
C. mrigala	0.247	0.259	0.247	0.150	0.250								
C. chagunio	0.253	0.236	0.216	0.192	0.265	0.216							
C. orientalis	0.265	0.268	0.253	0.283	0.280	0.286	0.286						
O. mossambicus	0.222	0.239	0.242	0.224	0.230	0.283	0.265	0.250					
H. nobilis	0.219	0.219	0.230	0.171	0.247	0.194	0.192	0.265	0.224				
P. ticto	0.230	0.238	0.236	0.176	0.253	0.165	0.202	0.289	0.280	0.213			
B. vagra	0.268	0.283	0.256	0.227	0.256	0.202	0.216	0.289	0.233	0.178	0.191		
L. rohita	0.250	0.244	0.247	0.153	0.222	0.126	0.213	0.304	0.253	0.200	0.184	0.213	-

4.5 Barcode Gap Analysis

Neither conspecific nor congeneric distances were calculable as the species were singleton. So, the nearest neighbor distances were considered in order to find out the genetic distance between a species and its closest congeneric relative. As mentioned in the table 4.10, *Cirrhinus mrigala* and *Labeo rohita* were most closely related neighbors with NND of 13.02, whereas *Channa orientalis* and *Oreochromis mossambicus* were closely related with highest NND of 24.22. *C. chagunio* is more near to *H. nobilis* than *B. vagra*, while *H. nobilis* is more near to *Tor putitora*. *X. cancila* was nearest to *Clarias batrachus* and vice versa with NND of 21.75. The confamilial K2P distance within 7 families was calculated, which ranged from 13.02%-23.42% with the mean distance of 18.7% (S.E 0.13).

The histogram (fig. 4.8) below plots the distribution of the nearest neighbor distances for each species. The K2P distances among nearest neighbors ranged from 13.02-24.22.

The mean distance was found to be 18.49 with 0.28 S.E. 15% of the species had the divergence between 13-15%. Around 8% of species was found to have divergence between 15-17% and 19%. And about 55% of total species studied had divergence value above 19%.

Table 4.10: Species with their nearest neighbor and distance to them.

Order	Family	Species	Nearest Species	Distance to NN
Beloniformes	Belonidae	Xenentodon cancila	Clarias batrachus	21.75
Cypriniformes	Cyprinidae	Barilius vagra	Hypophthalmichthys nobilis	19.29
Cypriniformes	Cyprinidae	Chagunius chagunio	Hypophthalmichthys nobilis	18.29
Cypriniformes	Cyprinidae	Cirrhinus mrigala	Labeo rohita	13.02
Cypriniformes	Cyprinidae	Hypophthalmichthys nobilis	Tor putitora	15.17
Cypriniformes	Cyprinidae	Labeo rohita	Cirrhinus mrigala	13.02
Cypriniformes	Cyprinidae	Puntius ticto	Tor putitora	16.47
Cypriniformes	Cyprinidae	Tor putitora	Cirrhinus mrigala	14.15
Perciformes	Channidae	Channa orientalis	Oreochromis mossambicus	24.22
Perciformes	Cichlidae	Oreochromis mossambicus	Tor putitora	21.8
Siluriformes	Bagridae	Mystus cavasius	Labeo rohita	20.3
Siluriformes	Clariidae	Clarias batrachus	Xenentodon cancila	21.75
Synbranchiformes	Mastacembelidae	Mastacembelus armatus	Hypophthalmichthys nobilis	21.2

4.6 Percent Similarity

Clustal w tool was used to calculate the similarities between the species with respect to their sequences. The most similarity (87.98%) was found between *Cirrhinus mrigala* and *Labeo rohita*. After them, Tor putitora and Cirrhinus mrigala were most similar (87.42%). In the same way the least similar pair was of Barilius vagra and Clarias batrachus (73.81%). Overall, Channa orientalis and Clarias batrachus were ~78% similar to all other fishes, meanwhile Hypopthalmichthys nobilis was ~83% similar to all other species which was highest similarity percentage in average.

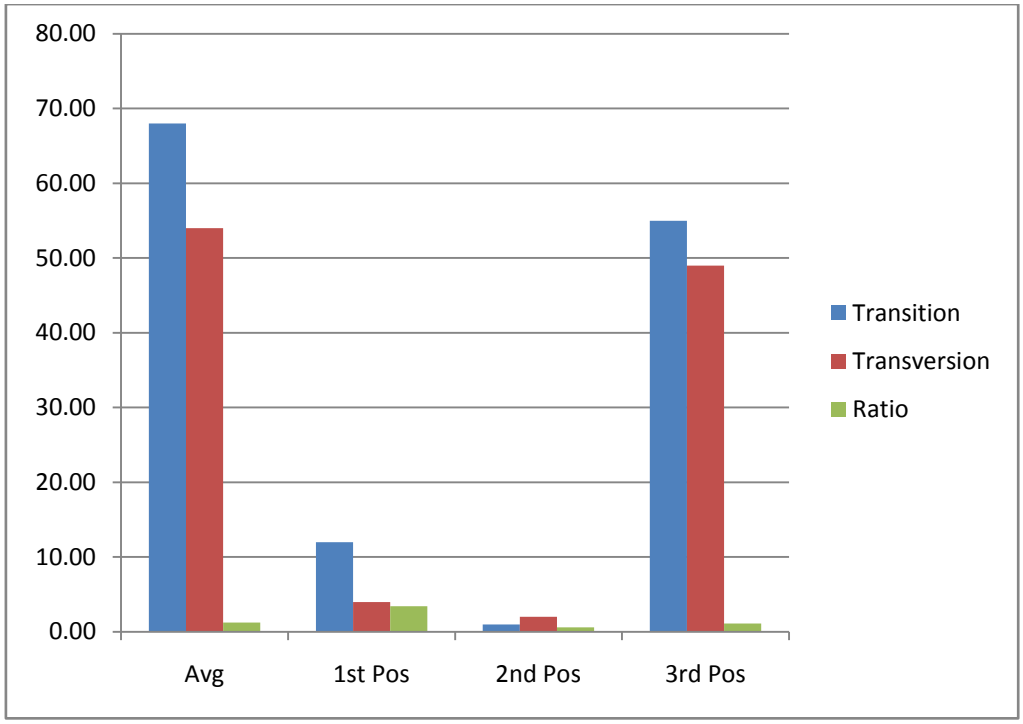


Figure 4.6: COX1 substitution plots. Number of transition and transversion at different codon positions.

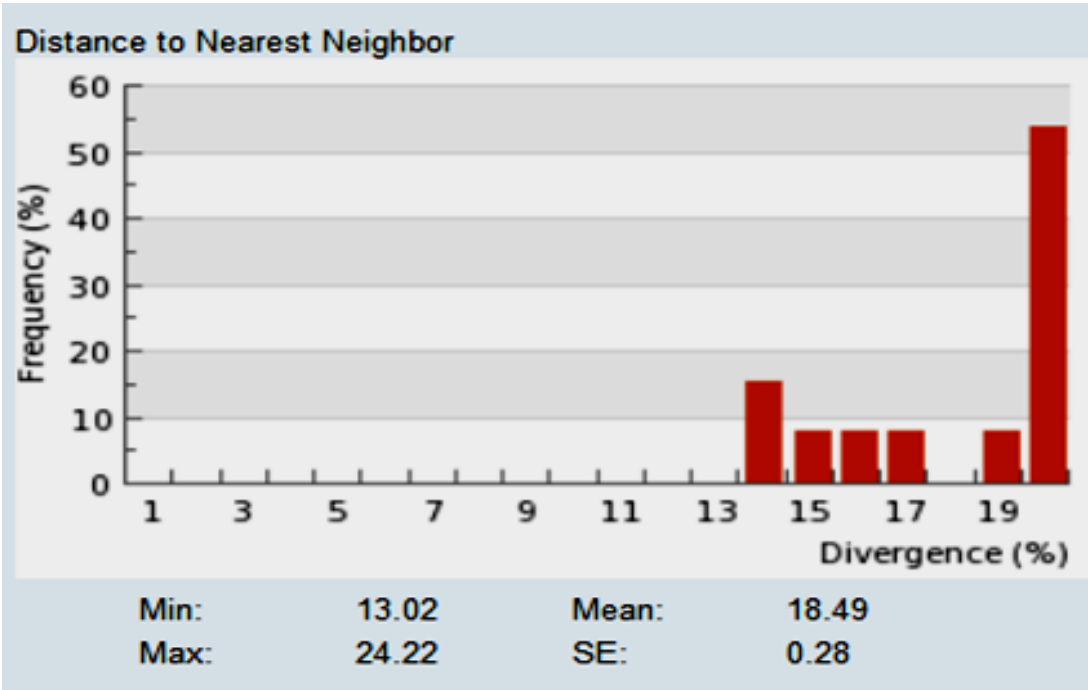
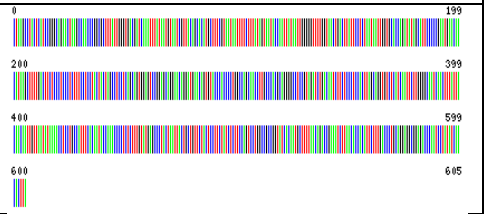
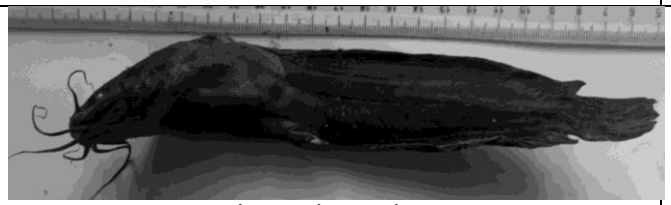
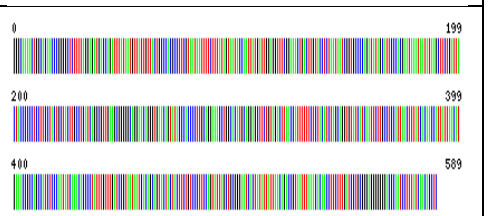
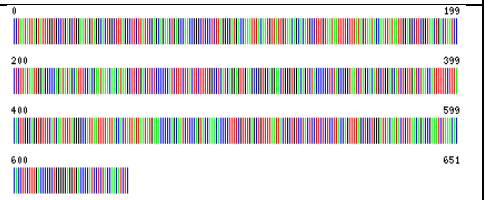
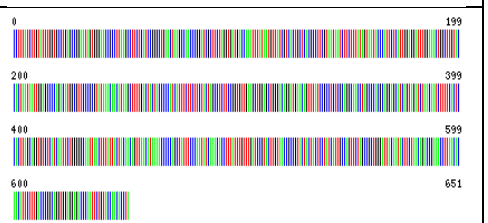
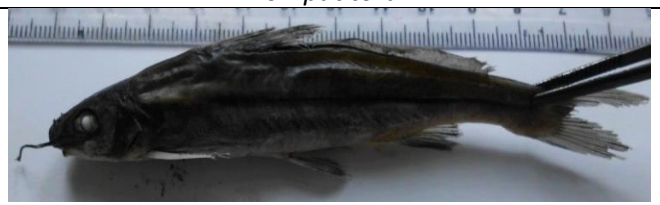
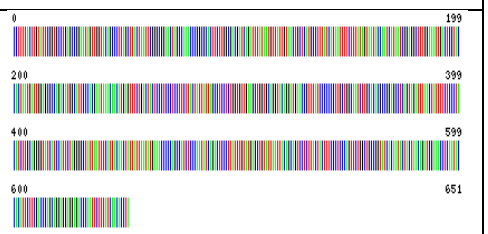
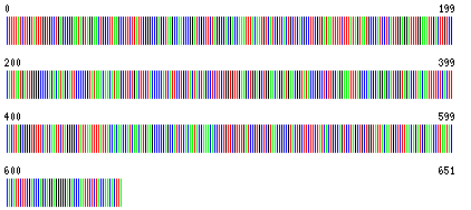
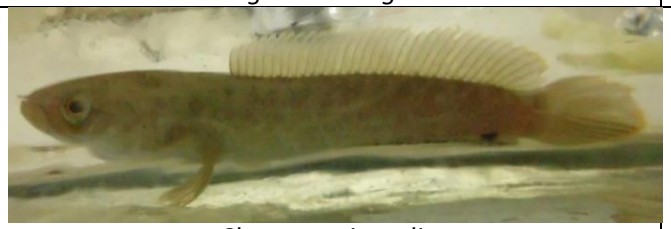
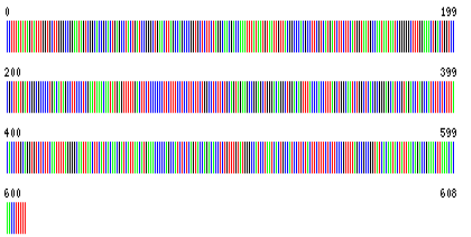
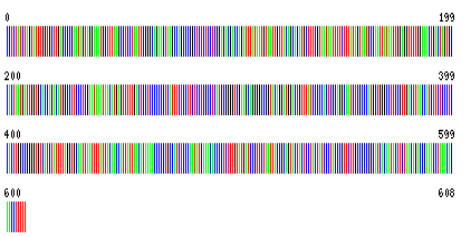
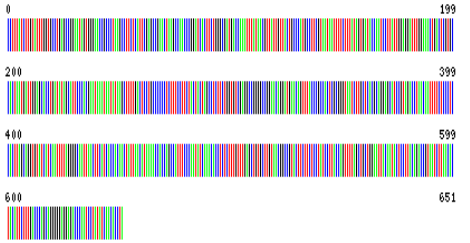
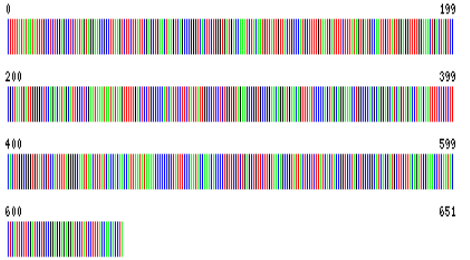
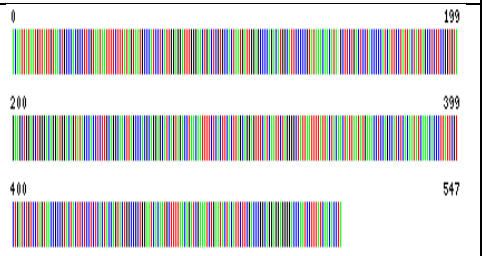


Figure 4.7: Histogram of Barcode gap analysis

Table 4.11: Species along with the sample code and their sequences and barcodes.

Code	Sample Picture	Sequence (5'-3')	Barcode
PVF1	 *Mastacembelus armatus*	GGGCCTAAGCCTACTCATCCGGGCAGAACTAAGCCAACCCGGCGCTTTATTGGGTGACGATCA AATTTATAATGTAATCGTTACAGCACATGCTTTCGTAATAATTTTCTTTATAGTAATACCAATTAT AATTGGGGGTTTTGGTAACTGACTTATTCCTCTAATAATTGGTGCACCAGATATAGCATTCCCCC GAATGAACAACATAAGCTTTTGACTTCTCCCTCCCTCTTTCCTCCTACTACTAGCCTCCTCAGCAG TGGAGTCCGGAGCAGGAACAGGATGAACAGTGTACCCACCCCTGGCCAGCAACCTTGCCCACG CAGGGGCCTCCGTAGATCTGACAATCTTCTCCCTCCACTTAGCAGGTGTCTCATCAATTCTTGGG GCAATCAACTTTATTACAACAATTGTTAATATAAAACCTCCTGCCATTTCACAATACCAAACCCCT CTCTTTGTATGAGCCCTACTAATTACTGCTGTTCTACTCCTTCTATCTCTTCCAGTCCTGGCCGCC GGTATCACAATGCTTTTAACAGACCGAAATCTTAACACCACATTCTTTGACCCTGCAGGAGGCG GAGACCCAATCCTATATCAACACTTA	
PVF2	 *Clarias batrachus*	TACGGGCCTAAACTTACTAATCCGGGCAAAACTGGCACAACCCGGGGCTCTTTTAGGAGATGA CCAGATTTATAATGTTATTGTTACTGCCCACGCCTTCGTAATAATTTTCTTTATAGTAATACCAAT TATGATTGGAGGTTTCGGAAACTGACTTGTACCTCTAATAATCGGTGCCCCAGATATAGCATTC CCACGAATAAATAATATAAGCTTCTGATTACTACCCCCCTCCTTCCTACTGCTACTTGCCTCATCA GGCGTTGAAGCGGGGGCAGGAACAGGGTGAACAGTATACCCACCCCTTGCAGGAAACCTGGC ACATGCAGGAGCTTCTGTAGACTTAACCATTTTTTCTCTACATCTAGCAGGTGTATCATCAATTC TTGCCTCCATTAACTTTATCACAACCATTATTAACATGAAACCGCCAGCCATCTCCCAATATCAA ACACCCCTATTTGTTTGATCCGTAATAATCACAGCAGTACTACTACTTCTGTCCCTTCCAGTATTA GCTGCGGGAATCACTATATTATTAACAGACCGTAATTTAAACACAACCTTCTTTGATCCTGCGG GAGGGGGGGACCCAATCCTTTATCAACACCTC	
PVF3	 *Xenentodon cancila*	CCTAATATTAGTATTTGGTGCCTGAGCTGGAATAGTAGGCACTGGCTTTAGCCTTCTTATTCGAGCAGAA CTGAGCCAACCGGGCTCCCTCTTAGGCGATGACCAAATTTACAATGTTATTGTAACAGCACATGCCTTTG TAATAATTTTCTTTATAGTAATACCAATTATAATTGGAGGCTTCGGTAACTGACTAATCCCTCTTATAATT GGAGCCCCAGATATAGCATTTCCACGAATAAACAACATAAGCTTCTGACTTCTCCCCCCATCTTTCTTACT TCTTTTGGCCTCATCAGGAGTTGAAGCCGGTGCAGGAACTGGGTGGACAGTCTACCCCCCTCTAGCCGG TAATCTAGCACATGCTGGAGCATCTGTTGATCTAACCATTTTTTCTTTACACTTGGCAGGTGTTTCATCAA TTTTAGGGGCCATTAATTTTATTACCACTATTATTAATATGAAACCCCCAGCAATCTCCCAGTACCAAACA CCCCTTTTCGTTTGAGCTGTTTTAATTACTGCTGTCCTACTCCTTCTCTCCTTACCAGTTTTAGCTGCTGGG ATTACAATACTCTTAACAGACCGAAATCTAAACACCACCTTCTTTGACCCCGCTGGGGGAGGTGACCCCA TCCTCTACCAACATCTC	
PVF4	 *Tor putitora*	CCTTTATCTTGTATTTGGTGCCTGAGCCGGAATAGTGGGAACCGCCTTAAGCCTTCTCATCCGG GCTGAACTAAGCCAACCCGGATCGCTTCTAGGTGATGACCAAATTTATAATGTTATCGTCACTG CTCACGCCTTCGTAATAATCTTCTTTATAGTAATACCCATTCTCATTGGAGGATTTGGAAACTGG CTTGTACCACTAATAATTGGAGCCCCAGACATGGCGTTCCCGCGTATAAACAACATAAGCTTTT GACTACTACCCCCATCATTTCTTCTTCTACTAGCCTCTTCCGGTGTTGAAGCTGGAGCTGGGACA GGATGAACAGTTTATCCGCCTCTCGCAGGCAACCTAGCCCACGCAGGAGCATCAGTAGATCTA ACAATCTTTTCACTCCACTTAGCAGGTGTCTCATCAATTCTGGGGGCAATTAACTTTATTACTAC AACAGTTAACATAAAACCCCCAGCTATTTCCCAATATCAAACACCTTTATTTGTTTGATCCGTACT TGTAACCGCCGTACTACTACTCCTATCATTGCCAGTTCTAGCCGCTGGGATTACAATACTTCTAA CAGACCGAAACCTTAACACAACATTCTTTGACCCGGCAGGTGGAGGAGACCCAATTCTGTACCA ACACCTA	
PVF5	 *Mystus cavasius*	CCTTTTACTTGTATTCGGTGCCTGAGCCGGAATAGTTGGTACAGCCCTTAGCTTACTAATTCGGG CGGAACTAGCCCAACCCGGCGCACTTCTTGGCGACGATCAGATTTATAATGTTATTGTAACTGC TCATGCCTTTATCATAATTTTCTTTATAGTAATACCAATCATAATCGGAGGCTTTGGAAATTGAC TCGTACCACTAATGATTGGAGCGCCAGACATGGCATTCCCACGAATAAACAACATGAGTTTCTG ACTACTCCCTCCCTCCTTCTTACTATTATTAGCCTCCTCTGGTGTTGAAGCTGGTGCAGGTACAG GATGAACCGTTTATCCCCCCCTTGCTGGCAATCTTGCACACGCCGGAGCCTCAGTAGATTTAACT ATTTTCTCCCTACATCTTGCAGGGGTATCATCCATTCTAGGGGCTATTAATTTCATTACAACTATT ATTAATATGAAACCCCCAGCCATCTCCCAATATCAGACTCCCTTATTTGTATGAGCTGTATTAATT ACAGCAGTCCTCTTACTACTTTCCCTCCCAGTTCTGGCTGCCGGTATCACAATACTACTAACAGA TCGAAACCTCAATACCACATTCTTCGACCCAGCAGGAGGAGAGGACCCAATTCTTTATCAACAC CTA	

PVF7	 *Cirrhinus mrigala*	CTTTTATCTTGTATTTGGTGCCTGAGCCGGAATAGTAGGAACTGCCTTAAGCCTTCTTAT TCGGGCCGAGCTAAGCCAACCCGGATCGCTTCTAGGCGACGACCAAATTTACAATGTCA TCGTCACTGCTCACGCCTTCGTGATAATTTTCTTTATAGTAATGCCCATCCTCATTGGAGG ATTTGGAAATTGACTTGTCCCATTAATGATTGGGGCCCCAGACATAGCATTCCCCCGTAT AAACAACATAAGCTTCTGACTTCTACCCCCATCATTCCTGCTACTACTAGCCTCTTCTGGT GTTGAAGCTGGAGCTGGAACAGGATGAACGGTATACCCGCCTCTTGCAGGAAATTTAG CCCACGCAGGAGCATCGGTAGACCTAACAATTTTCTCACTTCACTTAGCGGGTGTTTCAT CAATTCTAGGGGCTATTAATTTTATTACCACAACCATCAACATGAAGCCCCCAGCCATCT CACAATACCAAACACCTCTGTTCGTTTGATCCGTGCTCGTAACCGCCGTACTGCTTCTTCT ATCACTGCCAGTCCTAGCTGCTGGTATTACAATGCTCTTAACAGATCGAAACCTTAATAC CACATTCTTCGACCCAGCAGGAGGGGGAGACCCAATTCTCTACCAACACTTA	
PVF8	 *Chagunius chagunio*	CTTTTATCTTGTATTTGGTGCCTGAGCCCGGATAGTAGGAACTGCTTTAAGTCTCCTCATT CGAGCCGAACTGAGCCAACCCGGATCACTTCTAGGCGACGATCAAATCTACAATGTAAT CGTTACCGCCCATGCTTTCGTAATAATTTTCTTTATAGTTATACCCATTCTTATTGGAGGC TTTGGAAACTGATTAGTACCCCTCATAATTGGAGCCCCCGATATAGCATTCCCACGAATA AACAATATAAGCTTTTGACTATTACCCCCTTCATTCTTACTGCTTTTAGCCTCCTCTGGTGT TGAAGCCGGAGCCGGAACAGGATGGACAGTATATCCGCCTCTGGCAGGTAACCTAGCC CACGCAGGGGCATCCGTCGACCTAACCATCTTTTCTTTACACTTAGCTGGTGTTTCATCA ATTCTGGGAGCAATTAATTTCATCACCACAATTATTAATATGAAACCTCCAGCTATCTCCC AATATCAAACACCCTTATTTGTGTGATCTGTGCTTGTAACTGCCGTGCTGCTTCTTTTATC CCTTCCAGTTTTAGCCGCAGGAATTACAATACTTCTAACAGATCGTAACCTCAACACCAC ATTCTTCGACCCGGCAGGGGGTGGGGATCCAATTTTATATCAACACCTG	
PVF9	 *Channa orientalis*	CCTTTATATAGTATTTGGTGCTTGGGCTGGAATAGTCGGCACCGCACTGAGCCTACTGAT CCGGGCTGAACTTAGCCAGCCCGGTGCTCTTCTAGGCAACGACCAAATTTATAATGTAA TTGTTACGGCCCACGCCTTCGTCATGATCTTCTTCATGGTAATGCCAATAATAATCGGGGG GCTTTGGAAACTGACTGGTCCCGCTTATGATCGGCGCCCCTGACATAGCCTTCCCTCGAA TAAACAATATGAGTTTTTGACTTCTCCCCCCTTCTTTCCTCCTTCTTCTGGCCTCTTCTGCA GTAGAAGCCGGAGCTGGGACAGGCTGGACAGTTTACCCACCTTTAGCTGGCAATCTGG CTCACGCGGGAGCATCCGTAGACCTAGCCATCTTCTCTTTACACCTTGCAGGTGTCTCTT CAATTTTAGGGGCAATTAACTTCATCACAACGATCATTAACATGAAACCCCCAGCCATCT CTCAGTACCAGACACCTCTGTTTGTATGGGCCATCCTAATCACTGCCATCCTTCTACTTCT TTCTCTTCCCGTTTTAGCCGCCGGTATCACAATACTATTAACAGACCGAAACTTAAACAC AACCTTTTTGAAACCG----------	
PVF10	 *Oreochromis mossambicus*	CCTCTATCTAGTATTTGGTGTCTGAGCCCGAATAGTAGGAACTGGGTTTAGCCTCCTAAT TCGGGCAGAACTAAACCAGCCCGGCTCTCTCCTCGGAGACGACCAGATTTATAATGTAA TTGTTACAGCACATGCTTTCGTAATAATTTTCTTTATAGTAATGCCAATTATAATTGGAGG TTTTGGAAACTGACTAGTGCCACTAATGATTGGTGCACCAGACATGGCCTTCCCTCGAAT AAATAACATGAGTTTTTGACTCCTCCCCCCCTCATTTCTCCTTCTCCTCGCCTCATCCGGG GTCGAAGCAGGGGCCGGTACAGGATGGACTGTTTATCCCCCACTCGCAGGCAATCTCG CCCATGCTGGGCCTTCCGTTGACTTAACCATCTTCTCCCTCCACTTGGCCGGGGTGTCAT CTATTTTAGGTGCAATTAATTTTATTACAACCATTATTAACATAAAACCCCCTGCCATCTC CCAATATCAAACACCCCTCTTTGTATGATCCGTTCTAATTACCGCAGTACTACTCCTACTA TCCCTACCCGTTCTTGCCGCCGGCATCACAATACTTCTAACAGACCGAAACCTAAACACA ACCTTTTTTGA----------	

PVF11	 *Hypopthalmichthys nobilis*	CCTTTATCTTGTATTTGGTGCCTGAGCCGGAATAGTGGGAACCGCCCTAAGCCTTCTCAT TCGAGCCGAACTAAGCCAACCCGGATCACTTCTGGGCGATGACCAAATTTATAACGTTA TTGTTACTGCCCATGCCTTCGTAATAATTTTCTTTATAGTGATACCAATCCTTATTGGAGG ATTTGGAAACTGACTCGTGCCACTAATGATTGGAGCACCTGATATAGCATTCCCACGAA TAAATAATATAAGCTTTTGACTCCTGCCCCCCTCTTTCCTTCTACTACTAGCCTCTTCTGGT GTCGAGGCCGGGGCCGGAACAGGATGAACAGTTTACCCGCCACTCGCGGGTAATCTTG CTCACGCAGGAGCATCCGTAGACCTAACAATTTTCTCCCTCCACTTAGCAGGTGTATCAT CAATTTTAGGGGCAATTAACTTCATCACCACAACTATTAACATAAAACCACCAGCCATTT CCCAATATCAAACACCTCTCTTTGTTTGAGCTGTGCTTGTAACGGCCGTACTTCTTCTCCT ATCCTTACCAGTTTTAGCTGCTGGAATTACAATACTCCTTACAGACCGTAATCTTAACACT ACATTCTTTGACCCAGCAGGGGGAGGAGACCCAATCCTATATCAACACCTA	
PVF12	 *Puntius ticto(Pethia ticto)*	CCTTTATCTTGTATTCGGTGCCTGAGCCGGAATGGTAGGAACCGCCCTGAGCCTCCTTAT CCGAGCCGAACTAAGTCAACCAGGATCACTCCTAGGTGATGATCAAATTTATAATGTAA TCGTCACTGCTCACGCCTTCGTAATAATTTTCTTTATAGTTATGCCCATCCTGATCGGCGG ATTCGGAAACTGACTAGTCCCCTTAATAATCGGAGCCCCCGATATAGCATTCCCACGAAT AAATAACATAAGTTTTTGACTTCTACCACCCTCATTCCTACTATTATTAGCCTCCTCTGGT GTTGAAGCCGGAGCAGGGACAGGGTGAACAGTTTACCCGCCGCTAGCAGGAAACCTG GCCCATGCTGGAGCGTCAGTAGACCTAACAATTTTTTCACTTCACTTAGCAGGTGTTTCA TCAATTCTTGGGGCAATTAACTTTATTACTACAACTATTAATATGAAACCCCCAGCCACTA CCCAGTACCAAACACCACTGTTCGTCTGATCCGTACTTGTAACTGCCGTCCTACTCCTACT ATCACTACCAGTCTTGGCCGCGGGGATTACAATGCTTCTAACAGATCGAAACCTTAATAC CACATTCTTCGACCCGGCAGGGGGAGGAGACCCAATCCTCTATCAACACCTA	
PVF13	 *Barilius vagra*	CTTTTATATAATATTTGCCTCTCGAGCCTCTATAGTATTAACGGCCCTAAGTCTTCTTATTC GAGCTGAACTAAGTCAGCCCGGGTCACTTCTGGGTGATGACCAAATCTACAATGTTATT GTTACTGCCCATGCTTTTGTAATGATTTTCTTTATAGTGATGCCAATTCTTATTGGAGGTT TTGGAAACTGACTAGTCCCGCTAATGATTGGGGCTCCAGACATAGCATTCCCTCGAATA AATAATATAAGTTTTTGACTCCTGCCCCCATCATTCCTTCTATTATTGGCCTCCTCTGGTG TAGAAGCCGGTGCCGGAACAGGATGAACAGTTTATCCCCCACTAGCAGGAAACCTGGC CCACGCAGGAGCATCAGTAGACCTAACAATTTTCTCTCTTCACTTGGCGGGTGTATCGTC CATTTTAGGGGCAATTAACTTTATCACCACAACCATTAATATAAAACCCCCCGCTATTTCC CAATACCAAACACCCCTGTTCGTCTGAGCTGTTCTTGTAACAGCCGTATTACTCCTCTTAT CACTACCGGTCCTAGCTGCCGGCATCACGATGCTTCTTACAGATCGAAACCTCAATACCT CTTTCTTCGATCCTGCCGGACCAGGGGATCCTATCCTTTACCAACACCTA	
PVF14	 *Labeo rohita*	TACAATTATAATGTTATTGTAACTGCCCACGCCTTCGTAATAATTTTCTTTATAGTAATGC CCATCCTCATTGGAGGATTTGGGAACTGACTCGTGCCACTAATGATTGGAGCCCCAGAC ATGGCATTCCCCCGTATAAACAACATAAGCTTCTGACTCCTACCCCCATCATTCCTATTAC TATTAGCCTCTTCCGGTGTAGAAGCTGGAGCTGGGACAGGATGGACAGTATACCCACCT CTTGCAGGCAACTTAGCCCACGCAGGAGCATCAGTAGACCTAACAATTTTCTCACTTCAC TTAGCAGGAGTTTCATCAATTCTAGGGGCTATTAATTTTATTACTACAACTATTAATATGA AACCTCCAGCCATCTCACAATATCAAACACCCTTATTCGTCTGATCTGTCCTAGTAACCGC CGTACTACTTCTCCTCTCACTACCAGTACTGGCCGCTGGAATCACAATGCTTTTAACAGA TCGAAATCTGAATACTACATTCTTCGACCCGGCAGGACGAGGGGACCCAATCCTTTATC AACACCTA	

4.7 Genetic Diversity Analysis Using Phenograms

The phylogenetic tree generated based on K2P/NJ model is shown in figure 4.9. The figure provides overview of sequence divergence between all species of the Begnas Lake. According to the NJ tree, the species in the present study were clustered independently within their corresponding genera. Two clades which consisted of orders Cypriniformes and Siluriformes were identified with bootstrap values of 77% and 20% respectively. The 3rd clade should have consisted of *O. mossambicus* and *C. orientalis* which belonged to order Perciformes. Instead, *M. armatus* of Synbranchiformes family was found to be clustered with *O. mossambicus*. In the 2nd clade, *X. cancila* of Beloniformes family was found to be nearer to *C. batrachus* of Siluriformes as they were clustered together with 44% bootstrap support. Clade 2nd and 3rd were supported weakly by NJ analysis with bootstrap values less than 50%.

The ML and MP analyses were also conducted which showed somewhat varying result. The clade representing Perciformes was separately grouped which was absent in NJ tree. Cypriniformes, Siluriformes separately clustered. The outcome of MP analysis provided the same pattern as obtained by ML analysis. Still no strong bootstrap support was obtained for families Siluriformes and Perciformes.

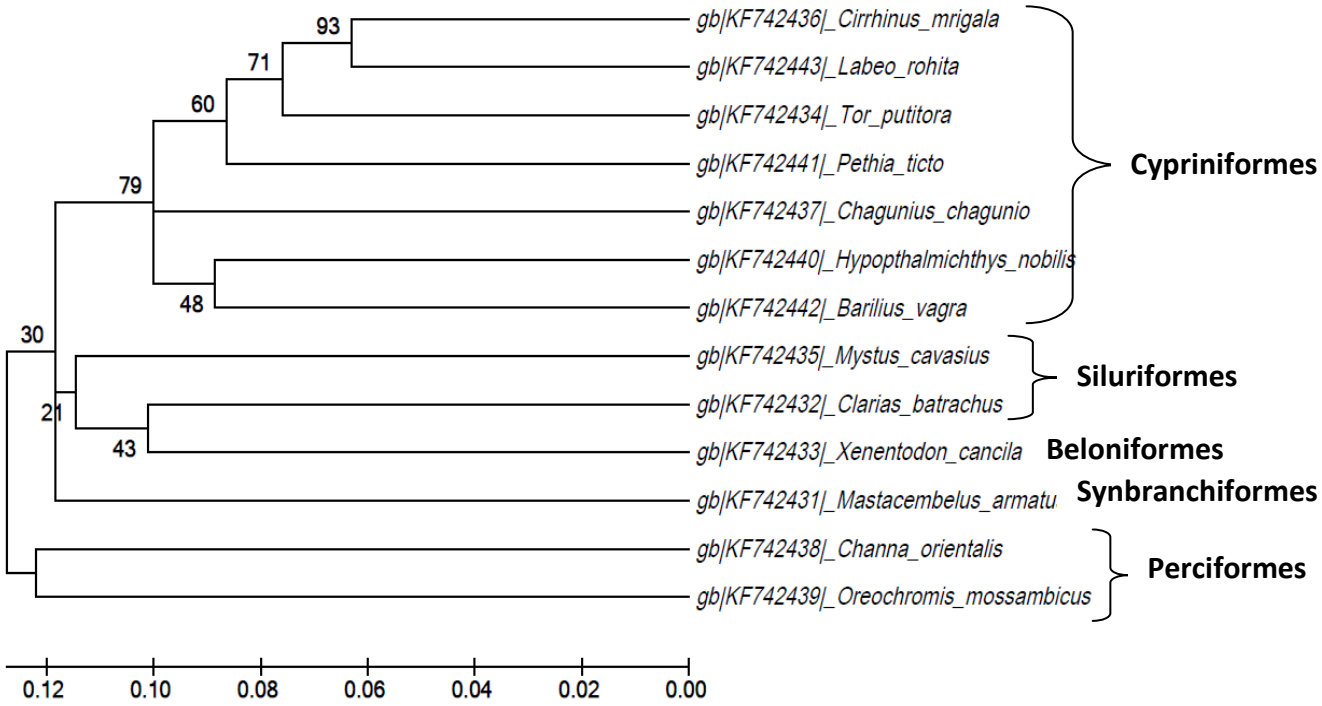


Figure 4.9: Phylogenetic tree inferred using Neighbor joining method based on K2P distance using MEGA5.1 software. Trees were constructed with the barcode fragment of the COI gene sequences. Numbers given at the main branches refer to bootstrap proportions among 1,000 bootstrap replicates.

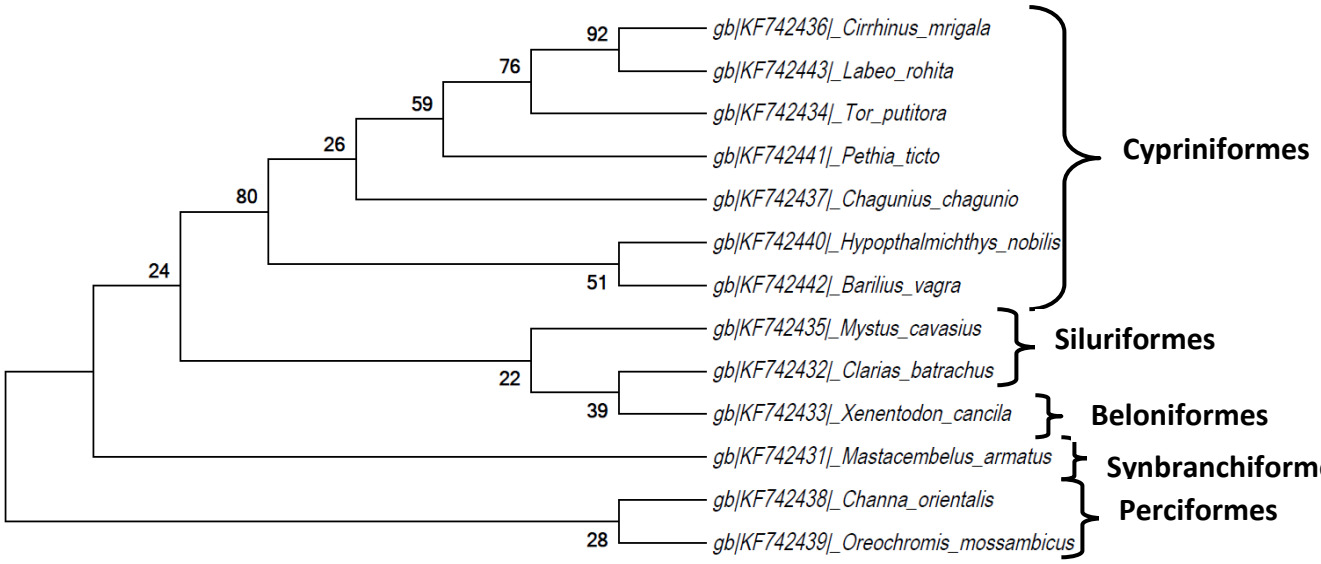


Figure 4.10: Molecular phylogenetic analysis by Maximum Likelihood method based on Tamura-Nei model by using MEGA5 software, version2. The scale bar represents an interval of Tamura-Nei genetic distance for the fishes.. The tree with the highest log likelihood (3364.5301) is shown.

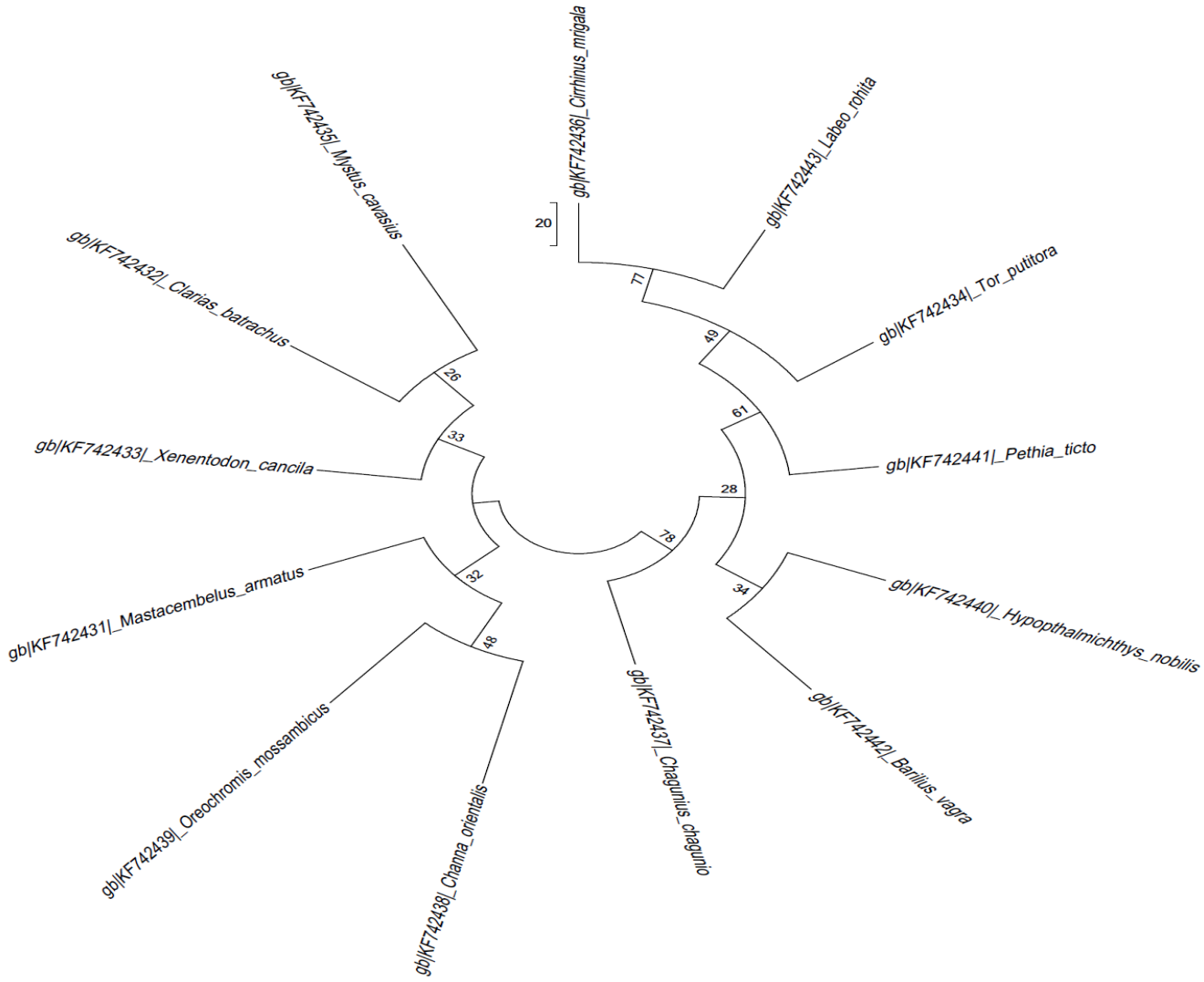


Figure 4.11: Maximum Parsimony tree constructed using Tamura-Nei model and the closest neighbor interchange method of the MEGA 5.2 software package. The numbers show the percentage of bootstrap confidence. This most parsimonius tree's length is 725.

Among the Perciformes, the one studied here, *C. orientalis and O. mossambicus*, clustered with their nearest neighbor with 100% bootstrap support to those from GenBank as shown in figure. The average pairwise distance using K2P among these Perciformes was found to be 0.178 (S.E 0.016). The overall GC content was 47.0%.

Similarly, the Cyprinoides, *L. rohita, C. mrigala, P. ticto, C. chagunio, H. nobilis* clusterd with the respective identical species inferred from GenBank with very strong (100%) bootstrap value. *T. putitora* and *B. vagra* grouped in the respective clades with their closest species taken from GenBank with the support of 98% and 99% respectively. It was shown in the figure below. The average K2P divergence among the order Cypriniformes was 0.142 and the GC content was 44.9%.

Within Siluriformes, *C. batrachus* crowded with other same species retrieved from GenBank with strong support (100%). Same was the case with *M. cavasius* as elaborated in the dendogram. The overall pairwise K2P distance between Siluriformes was 0.172 (S.E 0.013), while GC content was found to be 44.2%.

This shows that all the species were successfully descriminated by barcoding.

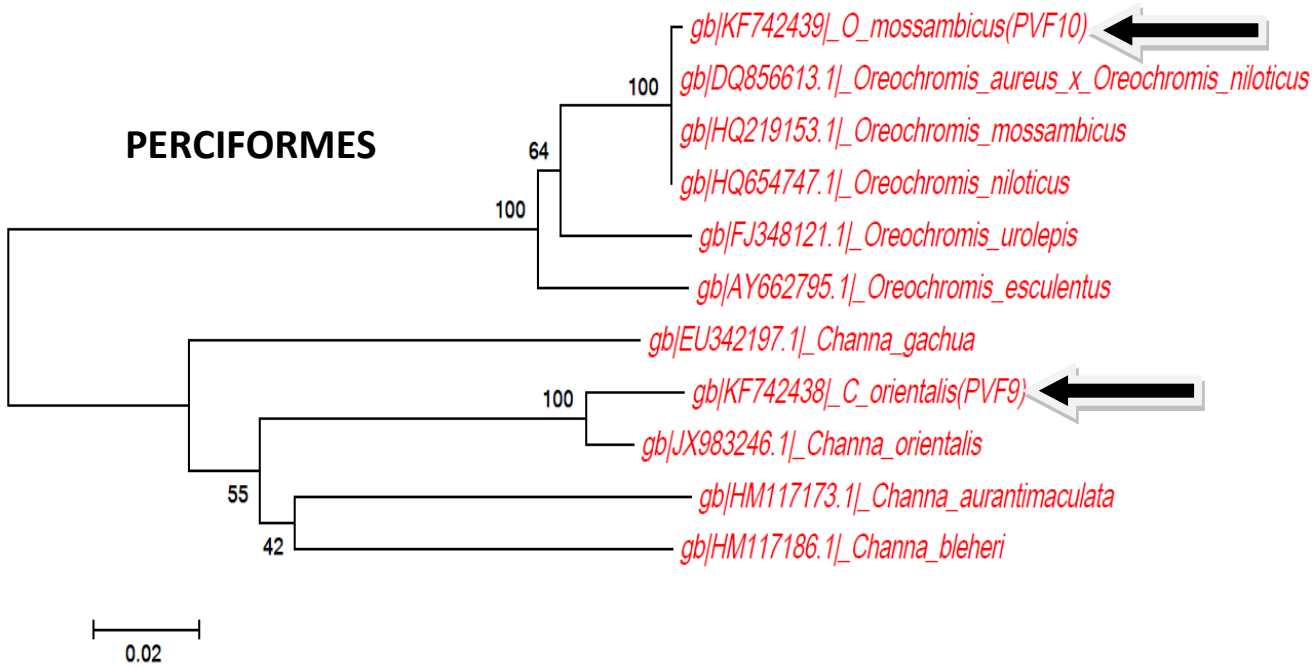


Figure 4.12: An NJ phylogram showing COI barcode divergences in specimens of Order Perciformes analyzed in the present work and of GenBank species. Taxa of interest are pointed with the arrow. Numbers at nodes represent bootstrap values. Arrows indicate the species studied in this research.

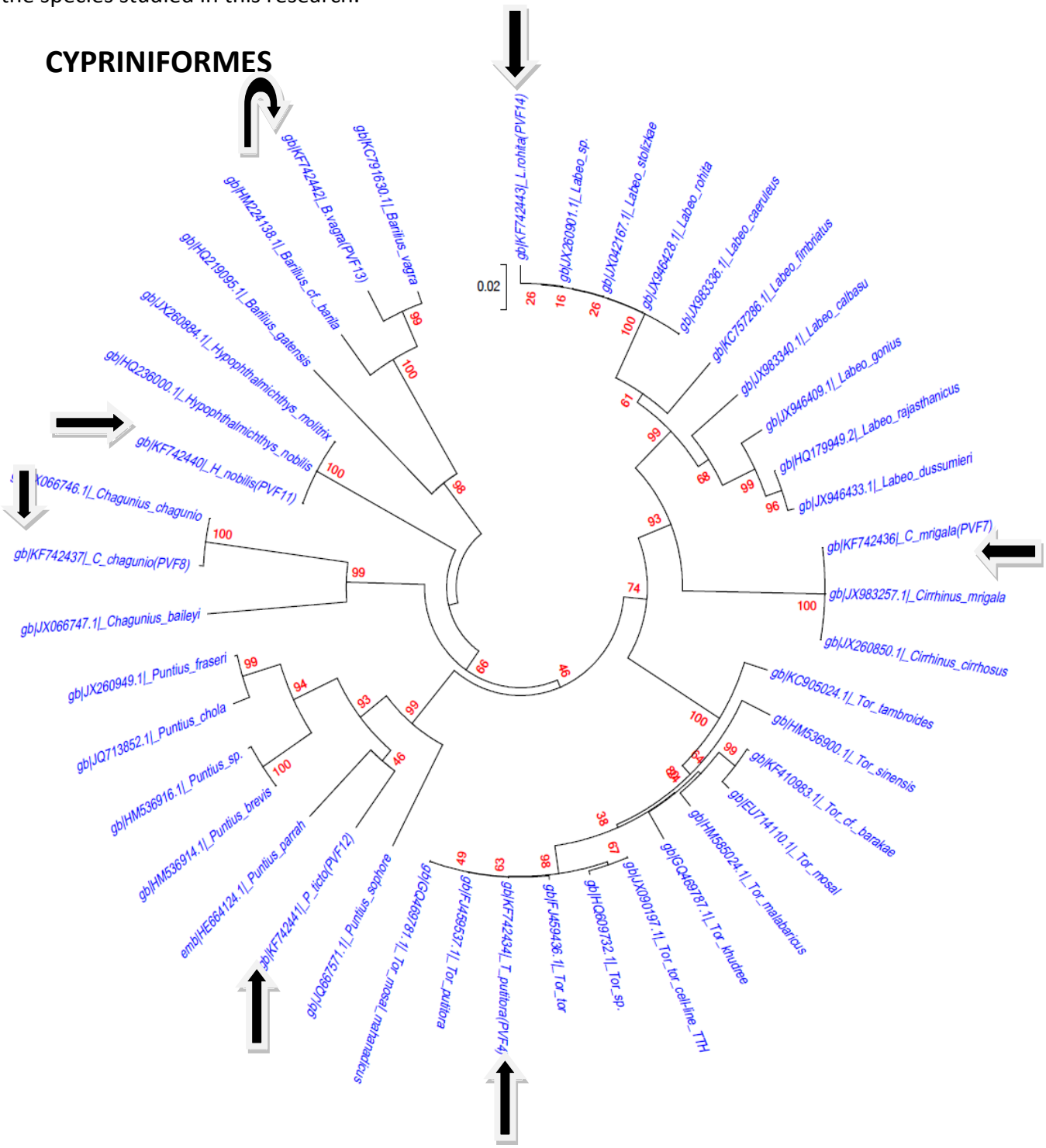


Figure 4.13: K2P distance NJ tree of COI sequences from the species of the Order Cypriniformes analyzed in the present work and of GenBank. Taxa of interest are pointed with the arrow. Numbers at nodes represent bootstrap values. Arrows indicate the species studied in this research. Evolutionary distance divergence scale bar is 0.02.

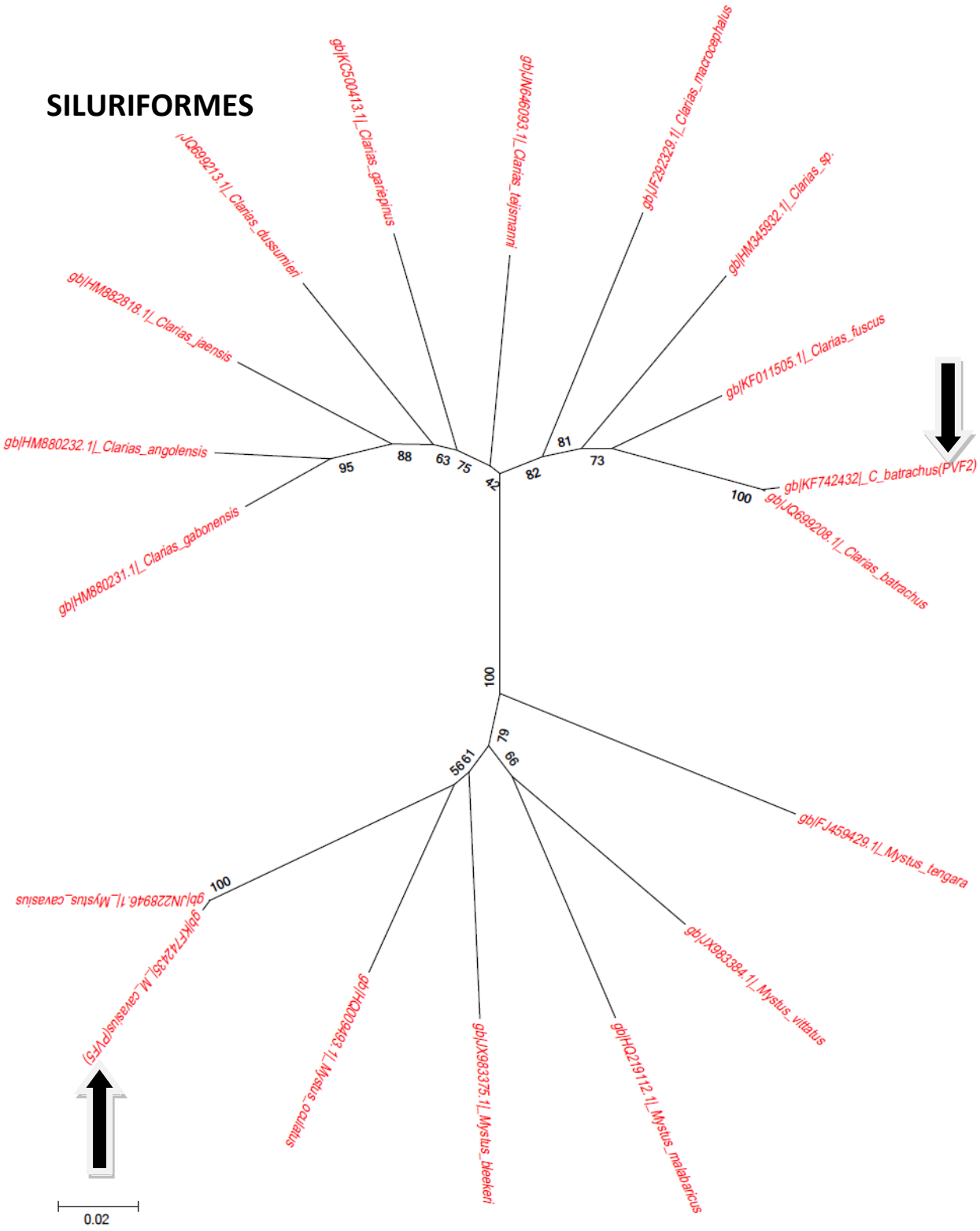


Figure 4.14: NJ tree based on the mitochondrial DNA COI nucleotide sequences of Siluriformes analyzed in the present work and of GenBank species. Taxa of interest are pointed with the arrow. Numbers at nodes are bootstrap values based on 1000 replicates. Evolutionary distance divergence scale bar is 0.02.

CHAPTER 5

DISCUSSION

The present study represents the first molecular survey of Pokhara's freshwater ichthyofauna. Amplification of the COI 5'region (652 bp) was successful for all assayed individuals. A total of 14 freshwater fishes from Lake Begnas were barcoded, out of which 13 were successful. Among the barcoded fishes, 2 were exotic fish i.e. *O. mossambicus* and *H. nobilis* and the rest were indigenous species.

5.1 Mitochondrial COI as Barcode

This study has strongly supported the efficacy of COI barcodes for diagnosing the freshwater fishes since all 13 species examined here represented a single coherent array of barcode sequences which are discrete from any others. As discussed previously, mitochondrial DNA is highly abundant in the cell and lacks recombination apparently, also it is maternally inherited in most taxa and its evolutionary rate is generally faster. Due to these reasons, mtDNA sequences provide excellent and useful markers for reconstructing the systematic assessment and phylogenetics of organisms and the deep-branch taxonomic classification of fishes (Mu X *et al.,* 2012; Ingman M *et al.,* 2000). Thus, mtDNA sequences are also applied in stream community ecology, population genetics, and systematics and taxonomy. DNA barcoding identifies fish products, agricultural pests, and disease vectors (Alexander LC *et al.,* 2009).

The universal primers amplified the target region in all 13 species generating COI barcodes of >600 bp on average, except 1, which was morphologically identified to be *Puntius conchonius*. Electropherograms did not contain sequence ambiguities, and the translated sequences did not contain premature stop codons, features consistent with bona fide mitochondrial sequences and not those of nuclear pseudogenes or NUMTs. In the Actinopterygii, no evidence of existence of NUMTs was reported in a review of their occurrence in plants and animals (Benasson *et al.,* 2001).

5.2 Species Identification Based on BLAST and BOLD

The identification search is done using BOLD and NCBI/GenBank but certain difficulties can be encountered in the process. BOLD-IDS confirms species identification search only if the species in the reference database has at least 3 barcoded specimens and identifies the query sequences if it matches the reference within the conspecific distance of less than 2% or not exceeding 3% (Wong EHK *et al.,* 2008). Indeed, BOLD data records and sequences often lack transparency for almost all species except those which are most

common. As a matter of fact, a large percentage of barcodes available from BOLD publicly are taken from GenBank records where there is high chances of tentative, wrong or low-quality sequences being stored. In addition, using BOLD-IDS there can be mistakes in private submissions and in records gathered from GenBank for species with few records which can lead to incorrect identification of sample sequences. Also, frequent alterations to the records can also change the identification results obtained. So, the person who needs to use the existing database in BOLD or NCBI must be careful so as to avoid errors leading to misidentifications (Wong LL *et al.,* 2011).

5.3 Ranking System

Ranking system can be used to assess the level of taxonomic reliability of species-specific DNA barcode arrays in the reference library. If DNA barcode sequence data from multiple researchers produce congruent and obvious matches for a given species, the reliability of the taxonomy is considered to be greater. Grading ranges from A to E.

Grade A: If a species is externally concordant. It matches with specimens from other BOLD projects or published sequences, with a maximum of 2% sequence divergence.

Grade B: If a species is internally concordant. No matching found through the BOLD-IDS. But it matches with specimens within the dataset. At least 3 specimens of the same species should be available, with 2% sequence divergence at most.

Grade C: If a species is internally concordant regarding genetic structures within species. The requirement is similar as for grade B but here, intraspecific divergence is more than 2%. In this case, BOLD-IDS can indicate monophyletic nearest neighbour of the same species, with more than 2% patristic distance.

Grade D: If there is no sufficient data i.e. low number of species analyzed. Also, no matching sequences are available in BOLD.

Grade E: If a species is discordant. No matching found with the same species in the BOLD.

Regarding discordance, there may be numerous causes. Morphological misidentifications, taxonomic uncertainty, sample processing defects, introgressive hybridization, or recent divergence are some reasons for discordant species assignments. In such cases, species name provided may not be matching accurately with the DNA barcode sequence.

This type of ranking system incorporates empirically-derived estimate of taxonomic congruence and validity. It can be easily implemented by any researcher to its own reference library using easy tools of BOLD (Costa FO *et. al.,* 2012).

A species barcodes assigned with grade B may shift to grade A when matching sequence will be available. Those allotted with grade C might need to be examined with additional markers. The species ranked D will need large number of species to be analyzed. In the same way, a species with E grading might improve its rank in the time interval as many more similar species will be barcoded and uploaded in the BOLD.

5.4 Compositional Analysis of COI Sequences

The overall GC content of the mitochondrial COI region was 44.97% which corresponded to those of Australian fishes (47.1%) (Ward *et al.,* 2005) and as reviewed by Saccone *et al.* (43.2%). We observed significantly more nucleotide changes at the codon position 3rd than 1st, and more at 1st than at 2nd; the study on Australian fishes displayed a similar result. There are more nt. changes at the 3rd codon position than the 1st and more at 1st than the 2nd. Similar kind of study of teleosts showed average nucleotide composition around 25%, with A=22.2%, C=27.7%, G=20.3% and T=29.8% (Miya M *et al.,* 2003) and another study on Channa species showed mean base composition of A=24.7%, G=20.2%, T-29.9% and C=25.2% (Siti-Balkhis AB *et al.,* 2011). Within Schizothorax sp., nucleotide composition showed CT bias (C=28.1%, T=28%, A=25.6% and G=18.2%) (Chandra S *et al.,* 2012).

The Perciformes were found to have higher percentage of GC, while Siluriformes had lower GC percent compared to Cypriniformes. This suggests that there is possibility of higher transition rate in Siluriformes and Cypriniformes than Perciformes. Mostly, Perciformes are carnivores and Siluriformes too. Cypriniformes are herbivores. Thus, it can be said that herbivores fishes show higher rate of nucleotide variation compared to carnivorous fishes. In the study of fishes from Godavari River, similar kind of result was obtained. GC content of carnivores and herbivores were 44.49% and 44.32% respectively (Kalyankar VB, 2012).

The protein coding genes are degenerate and here third codon base usually evolves faster than the first base, which in turn evolves faster than the second base. So, it can be anticipated that it is a 3rd base position variability that provides DNA barcoding its strength of species discrimination (Kumar KS *et al.,* 2011). Our analysis also shows that each third codon position base is highly variable. Variation at the third codon position is bimodally distributed. In the vertebrate mitochondrial code, every amino acid is exemplified by at least two codons, and every amino acid allows some variation in the third base. Amino acids with four or six codons dominate the more variable mode. Leucine and Serine are the amino acids coded by six codons each. Valine, Proline, Threonine, Alanine, Arginine and Glycine are coded by four codons each. While, amino acids with two codons dominate the less variable mode. Phenylalanine, Isoleucine,

Methionine, Tyrosine, Histidine, Glutamine, Asparagine, Lysine, Glutamic acid, Aspartic acid, Cysteine and Tryptophan are the amino acids coded by two codons each. Those amino acids which are encoded by more codons show more flexibility and more variation for third codon base (Ward *et al.,* 2007; Kumar KS *et al.,* 2011).

In our study also, the nucleotide variation was maximum at 3rd codon position and minimum at 2nd codon position. Substitution at 3rd codon position accumulates quickly and become saturated with transversion at a maximum level. At 1st and 2nd positions also, mutations accumulate despite saturation at 3rd position. Transversion gather slowly than transition. Rate of transition is higher among closely related species while, transversion is common among distantly related species (Meyer A, 1993). Transition rate i.e. substitution of purine by purine or pyrimidine by pyrimidine was higher in comparison to transversion rate i.e. substitution of purine by pyrimidine and vice versa. The ratio of non synonymous to synonymous mutations is less than 1 which indicates the strong purifying selection of *cox1* gene. The nucleotide variation reflects synonymous changes mostly, so while translating to protein level there is only little variation. The low rate of non synonymous mutations required for an important functional gene like *cox1.* Mostly, any amino acid is replaced by another amino acid with similar functions when there is change in the amino acid. For instance, an aliphatic amino acid is usually replaced by other aliphatic amino acids, among the most variable sites. But at times basic amino acid is interchanged with basic amino acid and vice versa. This kind of changes may alter the protein function. Amino acid substitutions must be more frequent among physico-chemically similar amino acids than among dissimilar amino acids and same type of thing is seen commonly. This shows the action of strong purifying selection in conserving amino acids and maintaining protein function intact.

Cox1 region holds a significant role as already discussed before. Some sites within it are vital. Seven amino acid residues are important for the d-pathway viz. tyrosine 19, asparagine 80, 98 and 163, aspartic acid 91, and serine 101 and 157. In the same way, the prosthetic group, heme is bounded by histidine 61. Other important functional positions include aspartic acid 51, tyrosine 54 and arginine 38 for H pathway; glutamic acid 40, glutamine 43, glycine 45 for calcium/Sodium binding site; and aspartic acid 50 and aspartic acid 221 for cytochrome c interactions (Ward *et al.,* 2007).

5.4.1 Nucleotide variation vs. Amino acid variation

It is obvious that to analyze the barcode sequences on the basis of nucleotide variations rather than amino acid variations is far reasonable. Most of the protein sequences over all the studied fishes are almost conserved. Since proteins are coded by variable (2 or more) codons in many cases, the changes can't be addressed into the species level. But

when nucleotide sequence variations are considered, the discrimination on species level is more vivid. This implies that amino acid sequence information is insufficient to discriminate the species as strongly as with nucleotide sequence information. For instance, in the current study, there were 43.71% variable sites among fish species when considering nucleotides, while only 24.71% variable sites were found in amino acid sequences among all the studied fishes. This can make huge difference when it is about comparing and identifying closely related species. Hence, it can be said that amino acid sequence diversity is much less than nucleotide sequence diversity and has abstruse species resolution as described by Ward RD and co (Ward *et al*., 2007)

5.5 Sequence Divergence

DNA barcoding works on the principle that inter-species variations are greater than the intraspecies variations. This helps to identify various species using nucleotide sequences (Khan SA *et al.,* 2011). One of the crucial criteria for the success of species identification by DNA barcoding is based on the differences between intra and inter-specific divergence, also known as barcoding gap. The congeneric divergence should be greater than conspecific divergence which means the genetic divergence should increase with increasing taxonomic levels theoretically (Pereira LHG *et al.,* 2013; Wong LL *et al*., 2011; Muchlisin ZA *et al.,* 2013). As expected this was not possible with the presently reported study because the size of sample taken was limited. There was no more than one specimen per species which restricted the conspecific genetic distances. Similarly, the presence of the multiple species within single genera was also not available in our study. Thus, the limited number of species in the study didn't permit further inference to be made on the sequence divergence. Nevertheless, in the previous fish barcoding studies, average K2P distances within species, genera, and families and so on were clearly corroborated as shown in the table 5.1 below:

Table 5.1: Summary of the DNA barcoding surveys of the freshwater fishes highlighting the no. of species, higher taxa, families and genera with multiple species analyzed.

Survey	No. of species analyzed	No. of higher taxa (order)	No. of families and families within multiple species	No. of genera and genera within multiple species (>2)	Mean value of K2P divergence of conspecific/congeneric comparison (%)	Reference
Upper Parana River	254	10	36/20	126/19	1.30/6.80	Pereira LGH *et al.,* 2013

Basin						
Paraiba do Sul River, Brazil	58	5	17/8	40/4	0.13/10.36	Pereira LGH, 2011.
Sao Fransisco RiverBasin, Brazil	101	6	22/11	75/6	0.50/10.61	Carvalho DC *et al.,* 2011.
Canada	190	20	28/15	85/21	0.27/8.75	Hubert N, 2008.
Mexico & Guatemala	61	8	15/5	36/6	0.45/5.10	Valdez-Moreno *et al.*, 2009.
Cuba	27	8	10/4	17/2	0.40/8.00	Lara A *et al.*, 2010.
Taal Lake Philippines	23	9	17/2	21/2	0.60/11.07	Aquilino SVL *et al.,* 2011.
North America	752	24	50/18	178/45	0.73/13.67	April J *et al.*, 2011.
Argentina	36	8	18/3	32/1	0.33/1.68	Ross JJ *et al.,* 2012.
India	25	1	9/4	17/2	-	Bhattacharjee MJ *et al.*, 2012.
Mexico	31	4	8/3	16/4	0.78/6.08	Meija O *et al.*, 2012.

5.6 Phylogenetic Analysis

The sequence similarity of the species barcoded presently with those barcoded earlier was examined through construction of phylogram. It is clear that the closely related species invariably get clustered in same clade. Thus, COI gene sequence can act as universal DNA marker for identification of fishes. By utilizing the advances in electronics and genetics, barcoding is going to be helpful for the researchers to quickly recognize unknown species and to retrieve information about them. Phylogenetic COXI sequence could effectively cluster most congeneric and confamilial species (Ward *et al.,* 2005). This could be observed in prior studies including Australian fishes (Ward et al., 2005), Cuban fresh water fishes (Lara *et al.* 2009), freshwater fishes from Mexico and Guatemala (Valdez-Moreno *et al.*, 2009), Canadian freshwater fishes (Hubert *et al.,* 2008) and Indian carangid fishes (Persis *et al.,* 2009). It is hypothesized that genetic

divergences should increase with the increasing taxonomic levels. But this was unable to be accomplished in our case as lower taxonomic analyses i.e. congeneric and conspecific values were not obtained due to limited sample size as already explained above.

Phylogenetic tree is a branching diagram that represents the evolutionary history of a group of organisms. Such a tree can provide a huge amount of information. For any particular group of animals this tree could identify the ancestors and closest relatives of the group (Hebert PDN *et al.,* 2003). In this study, the NJ, ML and MP trees revealed identical phylogenetic relationship among the species. The phylogenetic relationship among the species was clearly established, and similar species were clustered under same nodes while dissimilar species were clustered under separate nodes with both high and low bootstrap value support. Confamilial species clustered together in the trees. There is phylogenetic signal in COI sequence data although barcode analysis seeks only to delineate species boundaries (Kalyankar VB, 2012). Low bootstrap values in analysis indicates their position within respective families to be uncertain (Creˆte-Lafrenie`re A *et al.,* 2012). When genetic distances are low, the K2P model provides the best metric (Nei and Kumar 2000). Generally a simple NJ algorithm is used because the goal of barcoding is to provide species identification based on sequence similarity rather than to reconstruct deeper phylogenetic relationships accurately. In addition, NJ provides the necessary speed of analysis for the large data sets that are typical of DNA barcoding studies (Ball SL et al., 2005). On the other hand, ML methods are very flexible due to their plasticity—i.e., the possibility to implement and apply complex evolutionary models that account for several biases faced by sequences during evolution. Furthermore, ML methods are theoretically very sound and statistically consistent and have proved to be very efficient in recovering correct phylogenies, even when the sequences analyzed have evolved through very complicated evolutionary pathways (Negrisolo E *et al.,* 2004).

K2P model is generally used in barcoding because data set covers a large range of taxa spanning many orders and mtDNA is subject to mutational saturation at this level. Even though there are several distance models that take into account this issue, K2P is one of the simplest and commonest model used for describing differentiation among species using COI. On the other hand, being K2P the standard model used in barcode studies allows a better comparison with other barcode studies. But in our current study for most of the phylogenetic and COI sequence based studies, Tanura Nei Model has been used as analysis parameter as the best fitting substitution model.

5.7 Molecular Taxonomy Complements Morphological Taxonomy

Previously species used to be established through traditional approaches of taxonomy using phenotypes but now DNA barcoding approaches examines species delineation through COI barcode. Taxonomic identification of fish taxa exclusively based on morphological features can sometimes prove difficult because of the phenotype variation affected by environment (Mu X *et al.,* 2012). When morphological and molecular characteristics are combined, the gap between morphological taxonomy and DNA barcoding can be eliminated. The same idea has been manifested in BOLD construction. The sequences inferred are linked with the images as well as collection information of the species. So, problems related to morphological identification can be solved by searching relevant database in BOLD (Zhang J *et al.,* 2012). Occasionally, some species overlapped to others may be exhibited during barcoding. There are 3 factors which might be responsible for it: i) formation of reciprocal monophyly between two sister species, ii) introgressive hybridization which leads to polymorphisms in taxonomy, and iii) enormous taxonomic designation (Meyer *et al.,* 2005; Lara A *et al.,* 2010).

5.8 Barcoding in Our Perspective

DNA barcoding of fishes has already gained impetus in different parts of the world including Australia, Canada, China, Mexico, North America, India etc. But in Nepali waters, no efforts have been made so far. With the view to pioneering this effort to Nepalese fish diversities; the present study was undertaken to document and barcode freshwater fishes of Begnas Lake. All the species occurring in Nepalese waters have to be barcoded, so that as pointed by CBOL 'any animal, plant, any fungus or any organism can be identified on the spot, in an instant and anywhere by anyone' (Khan SA *et al.,* 2011). It is needed to broaden the collaboration in order to allow the assembly of global database of fish COI sequence. For this species need to be collected at larger extent. Also with multiple specimens per species from widely divergent locations is must as this will make the project more reliable and minimizes the risk of genetic diversity underestimation (Ward RD *et al.,* 2005). Confidence in the species discrimination and phylogenetic reconstruction can be increased by increasing the number of taxa sampled as well the number of characters (Creˆte-Lafrenie`re A *et al.,* 2012). After the construction of reference library of all the species on earth, any new unknown specimen can be processed and then identified using search engine in BOLD. The queried species will lie in the separate cluster with very similar species to itself as illustrated in Appendix 6. *Clarias batrachus* from Lake Begnas is placed in the same cluster with *C. batrachus* species from India as can be seen in the appendix. The NJ tree shows vividly that these two species are most closely related among all the sequences of *C. batrachus* in BOLD.

CHAPTER 6

SUMMARY AND CONCLUSION

6.1 SUMMARY

Under this study, we have studied Begnas Lake and small rivulets fishes from Pokhara sampling station. A total of 30 species were collected, 14 of them were morphologically identified and preserved for DNA isolation. Isolated DNA was amplified using universal primer pairs and sequenced bi-directionally for Cytochrome oxidase I gene. A total of 13 sequences were generated having sequence length of 600 bp in average, ranging from 549-652 bp. 1 sample gave no good sequence, so was discarded. All good sequences were analyzed using bioinformatics tools and then deposited to BOLD as well as submitted to NCBI/GenBank. A total of 13 species, 13 genera, 7 families and 5 orders were found to be contained on the database of Pokhara valley fishes.

After editing the sequences with reference to the standard fish sequence, 509 positions were remained in the final dataset. The overall GC content was found to be 44.97%, with the content higher at 1^{st} codon position (56.01%) and lower at 3^{rd} codon position (35.58%). The composition of 210 amino acids ranged from 0.55% lysine to 16.167% leucine. One of the amino acid, Cysteine was found to be absent. The overall evolutionary divergence and transition/transversion bias were 0.23 and 1.429 respectively. Most cases of transition and transversion occurred at 1^{st} and 3^{rd} codon position while least cases were at 2^{nd} position.

The average distance between confamilial individual was about 18.7%, whereas, lower taxonomic levels were not possible to analyze. The distance between nearest neighbor was found to be 18.49% in average. Based on Tamura Nei model, phylogenetic trees (NJ, ML, and MP) were generated which showed 3 clear clusters for Cypriniformes, Siluriformes and Perciformes. Rest of the orders Synbranchiformes and Beloniformes were represented by one species each- *Mastacembelus armatus* and *Xenentodon cancila* respectively. Also, separate NJ trees for Cypriniformes, Perciformes and Siluriformes were created using the sequences of our specimens and those from GenBank. The mean K2P distances between them were calculated as 0.142%, 0.178% and 0.172% respectively using MEGA. Our species correctly grouped with the most similar species from GenBank.

Thus, all 13 species of the present study were determined using COI gene by DNA barcoding method. Morphologically identified specimens were well supported by the phenograms. In the phylogenetic trees also closest relatives were observed to be packed in same groups. Hence, it is proved that DNA barcoding is a genuine tool for species identification using COI gene.

6.2 CONCLUSION

Fishes are one of the highly valued biological diversity of Nepal. Because of the sky rocketing population and increasing pollution, this diversity is led to great risk of extinction. There needs to be some ways for regulatory agencies and fishes managers for species authentication, food safety, conservation management and consumer health and support. DNA barcoding, hence, is emerging as an invaluable tool for the purposes.

Our result suggests that DNA barcodes provide highly effective identification systems for fish species. DNA barcoding is taking great endeavors of biological research genomics, phylogenetics and providing a comprehensive view into the biology, used to screen the large-scale genes, assign unknown individuals and discovery of new species find the phylogenic information and evolutionary relationship to distinguish the species to evaluating the promise and molecular diagnostics of individuals relative to described taxa, and DNA led discovery of new species. Barcoding is a technique that could aid the prompt and accurate identification of species that would be enormously beneficial in the application of molecular taxonomy evidence. Further investigations involving other groups of fresh water species of the area and also increasing the sample size should confirm the feasibility and establish the reliability of the technique for routine application in identification cases and other circumstances featuring fishes of applied importance.

In summary, DNA barcoding is worthy of appreciation. In fact, from the premises of molecular phylogenetics to assembling the tree of life, DNA sequences in environmental sampling and reconstruction of phylogenetic trees to place sequences into an evolutionary context have been used in several inventories of cryptic biodiversity (e.g. soil bacteria or marine/freshwater micro-organisms) (Hebert et al., 2003). Furthermore, deposition of barcode sequences in a public database, along with primer sequences, trace files, and associated quality scores, will make this species identification technique widely accessible. The assembly of a DNA barcode library for fishes will not only aid species recognition, but will also lead to the development of an automated identification system, which would be particularly valuable for law enforcement and allow conservation officials to identify poachers and smugglers. However, the present study only investigated a small proportion of fish species, and our specimens were mainly collected from Lake Begnas, Pokhara. For further studies, more comprehensive taxonomic samples, as well as populations from other geographical regions, are needed. Nevertheless, barcoding can be more valuable when used as an additional tool combined with the well-established tenets of ecology, biology and taxonomy (Dapkey T, 2008).

RECOMMENDATIONS

Obviously, there are some measures to be followed in order to get more advantages from barcoding in the future:

1. In our research we lacked in searching the conspecific and congeneric divergences. In order to provide an accurate measure of diversity, the nature and extent of intraspecific and interspecific divergence need to be quantified.
2. For more reliable taxonomic assessments, additional gene regions/markers, particularly of nuclear origin should be used along with mitochondrial genes. This can also help to understand evolutionary patterns more correctly.
3. The sampling has to be done on the basis of geographical conditions/ natural habitats of the fishes to be studied. This will make the study more specific and scientific for analysis.
4. Museum specimens, which currently lack an authority, can be barcoded so that they can regain taxonomic identification.
5. Further research aiming at molecular genetics, phylogeny and DNA-barcoding needs to be conducted to fishes all over the country employing much developed PCR and sequencing techniques.
6. It is expected to use DNA barcoding as a robust tool for tracking exotic invasive species, controlling smuggling of valuable fishes and checking adulteration in the fish products.

REFERENCES

Abdo Z, Golding GB (2007) A Step Toward Barcoding Life: A Model-Based, Decision-Theoretic Method to Assign Genes to Preexisting Species Groups. *Syst.Biol.* **56**(1):44-56

Alexander LC, Delion M, Hawthorne DJ, Lamp WO, Funk DH (2009) Mitochondrial lineages and DNA barcoding of closely related species in the mayfly genus Ephemerella (Ephemeroptera: Ephemerellidae). *J. N. Am. Benthol. Soc*. **28**(3):584–595

Aliabadian M, Kaboli M, Nijman V, Vences M (2009) Molecular Identification of Birds: Performance of Distance-Based DNA Barcoding in Three Genes to Delimit Parapatric Species. *PLoS ONE.* **4**(1): e4119

Altschul SF, Gish W, Miller W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol.* **215**(3): 403-410

Amaral AR, Sequeira M, Coelho M (2007) A first approach to the usefulness of cytochrome *c* oxidase I barcodes in the identification of closely related delphinid cetacean species. *Mar Freshwater Res*. **55**: 505-510

Antunes A, Ramos MJ (2005) Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. *Genomics* **86**: 708 – 717

April J, Mayden RL, Hanner RH, Bernatchez L (2011) Genetic calibration of species diversity among North America's freshwater fishes. *Proc Natl Acad Sci USA*. **108**(26): 10602-10607

Aquilino SVL, Tango JM, Fontanilla IKC, Pagulayan RC, Basiao ZU, Ong PS, Quilang JP (2011) DNA barcoding of the ichthyofauna of Taal Lake, Philippines. *Mol Ecol Res*. **11**(4):612-619

Austerlitz F, David O, Schaeffer B, Bleakley K, Olteanu M, Leblois R, Veuille M, Laredo C (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics.* **10**(10)

Avise JC (2004) Molecular Markers, Natural History and Evolution. *Chapman and Hall*, *New York*: 511

Ball SL, Hebert PDN, Burian SK, Webb JM (2005) Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *J. N. Am. Benthol Soc.* **24**(3): 508-524.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Mol Ecol.* **13**: 729–744

Becker S, Hanner R, Steinke D (2011) Five years of FISH-BOL: Brief status report. *Mitochondrial DNA*, **22**(S1): 3-9

Bensasson D, Zhang D, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evolut*. **16**(6).

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. *Philosophical T*rans R Soc Lond B. **360**: 1935–1943.

Brunori M, Antontni G, Malatesta F, Sarti P and Wilson MT (1987) Cytochrome-c oxidase: Subunit structure and proton pumping. *Eur. J. Biochem*. **169***:* 1-8.

Buhay JE (2009) "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. *J Crust Biol.* **29**(1): 96–110.

Carvalho DC, Neto DAP, Brasil B.S.A.F, Oliveira D.A.A (2011) DNA barcoding unveils a high rate of mislabeling in a commercial freshwater catfish from Brazil. *Mitochondrion.* **22**(S1): 97–105.

Chandra S, Barat A, Singh M, Singh BK, Matura R (2012) DNA bar-coding of Indian coldwater fishes genus Schizothorax (family: Cyprinidae) from Western Himalaya. World *J Fish Mar Sci.* **4**(4): 430-435

Chaves PB, Graeff VG, Lion MB, Oliveira LR and Eizirik E (2012) DNA barcoding meets molecular scatology: short mitochondrial DNA sequences for standardized species assignment of carnivore noninvasive samples. *Mol Ecol Res*. **12**: 18-35

Costa FO, Landi M, Martins R, Costa MH, Costa ME, Carneiro M, Alves MJ, Steinke D, Carvalho GR (2012) A ranking system for reference libraries of DNA barcodes: application to marine fish species from Portugal. *PLoS ONE*. **7**(4): e35858

Creˆte-Lafrenie`re A, Weir LK, Bernatchez L (2012) Framing the Salmonidae Family Phylogenetic Portrait: A More Complete Picture from Increased Taxon Sampling. *PLoS ONE.* **7**(10): e46662

Dapkey T (2008) Combining DNA barcoding and macroinvertebrate sampling to assess water quality. *University of Pennsylvania*, *Scholarly Commons.*

Dasmahapatra KK and Mallet J (2006) DNA barcodes: recent successes and future prospects. *Heredity.* **97**: 254–255

David O, Laredo C, Leblois R, Schaeffer B, Vergne N (2012) Coalescent-based DNA barcoding: multilocus analysis and robustness. *J Comput Biol.* **19**(3):271-278

DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Phil. Trans. R. Soc. B.* **360:** 1905–1916

Dasmahapatra KK and Mallet J (2006) DNA barcodes: recent successes and future prospects. *Heredity*. **97**: 254–255

Doyle JJ, Gaut BS (2000) Evolution of genes and taxa: a primer. *Plant Mol Biol* **42**: 1–23

Elmeer K, Almalki A, Mohran KA, AL-Qahtani KN, Almarri M (2012) DNA barcoding of *Oryx leucoryx* using the mitochondrial cytochrome C oxidase gene. *Genet. Mol. Res***. 11** (1): 539-547

Etnier DA, Starnes WC (2002) Fishes of Tennessee. New found press, UK: 53-60pp.

Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**:783-791.

Ferri E, Barbuto M, Bain O, Galimberti A, Uni Shigehiko, Guerrero R, Ferte H, Bandi C, Martin C and Casiraghi M (2009) Integrated taxonomy :traditional approach and DNA barcoding for the identification of filarioid worms and related parasites (Nematoda). Frontiers in Zoology, **6**:1.

Ferro W (1981/82) Limnology of Pokhara Valley Lakes (Himalayan Region, Nepal) and its implication for fishery and fish culture. *J. Nepal Res. Center* **5/6**: 27–52

Ferro W, Swar DB (1978) Bathymetric maps from three lakes in Pokhara Valley (Nepal). J. Inst. Sc. **1**: 177–188

Ferro W, Badagami PR (1980) On the biology of the commercially important species of fish of Pokhara Valley (Nepal). J. Inst. Sc. **3**: 237–250

Frézal L, Leblois R (2008) 4 years of DNA barcoding: current advances and prospects. Infection, Genet Evol **8**(5): 727-36

Godfray HCJ (2007) Linnaeus in the information age. Nature Publishing Group **446**: 259-260

Godfray HCJ, Knapp S (2004) **Introduction to Theme Issue, "Taxonomy for the 21st Century"** Philos Trans R Soc London [Biol] ***359****: 559-570*

Gurung TB, Rai AK, Joshi PL, Nepal A, Baidya A, Bista J(2001) Breeding of pond reared golden Mahaseer (Tor putitora) in Pokhara, Nepal. Paper presented at: The symposium on cold water fishes of trans-Himalayan region, 10-13 July 2001, Kathmandu, Nepal.

Hajibabei M, deWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Mackie PM, Hebert PDN (2005) Critical factors for assembling a high volume of DNA barcodes. *Philos Trans R Soc London* ***360****:1959-1967*

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics, *TIG* **30**(10)

Hajibabaei M, Singer GAC, Hickey DA (2006) Benchmarking DNA barcodes: an assessment using available primate sequences. *Genome* **49**: 851–854

Hanner R, Becker S, Ivanova NV, Steinke D (2011) FISH-BOL and seafood identification: Geographically dispersed case studies reveal systemic market substitution across Canada. *Mitochondrial DNA* **22**(S1): 106-122

Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. *PLoS Genet* **6**(2): e1000834

Hebert DG (2001) Museum natural science and the NRF: crisis times for practitioners of fundamental biodiversity science. *S. Afr. J. Sci.* **97**: 168–172

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proc R Soc Lond [Biol]* **270**: 313–321.

Herrmann PC, Gillespie JW, Charboneau L, Bichsel VE, Paweletz CP, Calvert VS, Kohn EC, Emmert-Buck MR, Liotta LA, Petricoin III EF (2003) Mitochondrial proteome: Altered cytochrome c oxidase subunit levels in prostate cancer*. Proteomics* **3**: 1801-1810

Hocker JM (1989) Cytochrome-c-Oxidase Deficient Cardiomyocytes in the Human Heart- An Age-Related Phenomenon. *Am J Pathol* **134**(5)

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and Using a Plant DNA Barcode. *PLoS ONE* **6**(5): e19254

Hubert N, Hanner R, Holm E, Mandrak NE, Taylor E, Burridge M, Douglas W, Dumont P, Curry A, Bentzen P, Zhang J, April J, Bernatchez L (2008) Identifying Canadian Freshwater Fishes through DNA Barcodes. *PLoS ONE* **3**(6): e2490

Huelsenbeck JP (1995) The performance of phylogenetic methods in simulation. Syst. Biol. Vol. **44**: 17–48

Ingman M, Kaessmann H, Paabo S and Gyllensten U (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* **408**: 708-713

Ivanova NV, Zemlak TS, Hanner RH, Hebert PD (2007*)* Universal primer cocktails for fish DNA barcoding. *Mol Ecol Notes* **7**: 544-548

Kalyankar VB (2012) Molecular taxonomy of freshwater fishes from Godavari riverine system using mitochondrial DNA cytochrome I oxidase gene. Phd thesis submitted to Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India.

Kartavtsev YP, Sharina SN, Goto T, Balanov AA, Hanzawa N (2009) Sequence diversity at Cytochrome oxidase 1 (Co-1) gene among Sculpins (Scorpaeniformes, Cottidae) and some other Scorpion fish of Russia Far East with Phylogenetic and Taxonomic Insights. *Genes and Genomics* **31**(2): 183-197

Khadka UR, Ramanathan AL (2012) Major ion composition and seasonal variation in the lesser Himalayan lake: case of Begnas lake of the Pokhara valley, Nepal. Arab J Geosei

Khan SA, Kumar CP, Lyla PS, Murugan S (2011) Identifying marine fin fishes using DNA barcodes. *Current Science* **101**(9): 1152-1154

Kimura M (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**:111-120

Kizirian D, Donnelly MA (2004) The criterion of reciprocal monophyly and classification of nested diversity at the species level. *Mol Phylogenet Evol* **32**: 1072–1076

Kress WJ and DL Erickson (eds.) DNA Barcodes: Methods and Protocols, *Methods Mol Biol* **858**

Krishnankutty N and Chandrasekaran S (2008) Linnaeus 300: Tips for tinkering morphological taxonomy. *Curr Sci* **94**(5): 565-567

Kumar KS, Goswami UC (2011) Nucleotide sequences variation of *Osteobrama* (Heckel) freshwater fish species of North-East India based on mitochondrial cox1 gene. *Scholars Research Library* **3**(6): 437-442

Lara A, de Leon JLP, Rodriguez R, Casane D, Cote G, Bernatchez L, Garcia-Machado E (2010) DNA barcoding of Cuban freshwater fishes: evidence for cryptic species and taxonomic conflicts. Mol Ecol Res **10**: 421-4

Lefebure T, Douady CJ, Gouy M and Gibert J (2006) Relationship between morphological taxonomy and molecular divergence within Crustacea: Proposal of a molecular threshold to help species delimitation. *Mol Phylogenet Evol* **40:** 435–447

Lillywhite K, Lee DJ (2011) Automated fish taxonomy using evolution-constructed features. Lecture Notes in Computer Science **6938**: 541-550

Lou M (2012) Improving specimen identification: Informative DNA using a statistical Bayesian method. PhD Thesis submitted in McMaster University, Hamilton, Ontario.

Lynn DH, Struder-Kypke MC (2010) Comparative analysis of the mitochondrial cytochrome *c* oxidase subunit I (COI) gene in ciliates (Alveolata, Ciliophora) and evaluation of its suitability as a biodiversity marker. *Syst Biodiversity* **8**(1): 131–148

Manktelow M, History of Taxonomy. Evolutionary Biology Centre, Dept of Systematic Biology.

*Mejía O, León-Romero Y, Soto-Galera E (2012) DNA barcoding of the ichthyofauna of Panuco-Tamesi complex: Evidence for taxonomic conflicts in some groups. Mitochondrial DNA **23**: 471-476*

Meyer A (2005) Evolution of mitochondrial DNA in fishes.

Hochachka aruI Mommsen (eds.), *Biochem Mol Biol Int Fish*, **2**

*Meyer CP, Paulay G (2005) DNA Barcoding: Error rates based on comprehensive sampling. PLOS Biology **3** (12):* e422

*Miya M, Takeshima H, Endo H, Ishiguro NB, Inoue JG, Mukai T, Satoh TP, Yamaguchi M, Kawaguchi A, Mabuchi K, Shirai SM, Nishida M (2003) Major pattern of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. Mol Phylogenet Evol **26**:121-138*

Monaghan MT, Balke M, Pons J and Vogler AP (2006) Beyond barcodes: complex DNA taxonomy of a South Pacific Island radiation. *Proc. R. Soc. B* **273**: 887–893

Muchlisin ZA, Thomy Z, Fadli N, Sarong MA, Siti-Azizah MN (2013) DNA Barcoding of freshwater fishes from lake Laut Tawar, Aceh province, Indonesia. *Acta Ichthyologica et Piscatoria* **43**(1):21-29

Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R (2008). Statistical assignment of DNA sequences using Bayesian phylogenetics. *Syst Biol* **57**: 750–757

Munch K, Boomsma W, Willerslev E, Nielsen R (2008) Fast Phylogenetic DNA barcoding. *Phil. Trans. R. Soc. B*. **363**: 3997–4002

Mu X, Wang X, Song H, Yang Y, Luo D, Gu D, Xu M, Liu C, Luo J, Hu Y (2012) Mitochondrial DNA as effective molecular markers for the genetic variation and phylogeny of the family Osteoglossidae. *Gene* **511**:320-325

Negrisolo E, Minelli A,Valle G (2004) The Mitochondrial Genome of the House Centipede Scutigera and the Monophyly Versus Paraphyly of Myriapods. *Mol. Biol. Evol.* **21**(4):770-780

Nei M and Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Nelson JS (2006) Fishes of the world. 4th edition, John Wiley & Sons, Inc., Hoboken, New Jersey: 1-14pp

Ogurtsov AY, Sunyaev S, Kondrashov AS (2004) Indel-Based Evolutionary Distance and Mouse–Human Divergence. *Genome Res* **14:**1610–1616

Parks DH, MacDonald NJ, Beiko RG (2011) Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics **12*:** 328

Peng Z, Wang J, He S (2006).The complete mitochondrial genome of the helmet catfish Cranoglanis bouderius (Siluriformes: Cranoglanididae) and the phylogeny of otophysan fishes. *Gene* **376**: 290–297

Pereira LHG, Hanner R, Foresti F and Oliveira C (2013) Can DNA barcoding accurately discriminate megadiverse Neotropical freshwater fish fauna? *BMC Genetics* **14**:20.

Pereira LGH, Maia GMG, Hanner H, Foresti F, Oliveira C (2011) DNA barcodes discriminate freshwater fishes from the Paraiba do Sul River Basin, Sao Paulo, Brazil. *Mitochondrial DNA* **22**: 71-79

Persis M, Chandra Sekhar Reddy A, Rao LM, Khedkar GD, Ravinder K, Nasruddin K (2009) COI (cytochrome oxidase-I) sequence based studies of Carangid fishes from Kakinada coast, India. *Mol. Biol. Rep*. **36** (7): 1733-40

Petr T., Cold water fish and fisheries in countries of the high mountain arc of Asia (Hindu Kush-Pamir-Karakoram-Himalayas). A Review

Pires AC, Marinoni L (2010) DNA barcoding and traditional taxonomy unified through Integrative Taxonomy a view that challenges the debate questioning both methodologies. *Biota Neotrop* **10**(2)

Prosdocimi F, de Carvalho DC, de Almedia RN, Beheregaray LB (2011) The complete mitochondrial genome of two recently derived species of the fish genus Nannoperca (Perciformes, Percichthyidae). *Mol Biol Rep*.

Qiongying T, Huanzhang L, Mayden R et al. (2006) Comparison of evolutionary rates in the mitochondrial DNA cytochrome *b* gene and control region and their implications for phylogeny of the Cobitoidea (Teleostei, Cypriniformes). *Mol Phylogenet Evol* **39**: 347 – 357

Radulovici AE, Archambault P, Dufresne F (2010) DNA Barcodes for Marine Biodiversity: Moving Fast Forward? *Diversity* **2**: 450-472

Rai AK (2000) Limnological characteristics of subtropical Lakes Phewa, Begnas and Rupa in Pokhara valley, Nepal. *Limnology* **1**(1):33-46

Rai AK, Shrestha BC, Joshi PL, Gurung TB, Nakanishi M (1995) Bathymetric maps of Lake Phewa, Begnas and Rupa in Pokhara Valley, Nepal. *Mem. Fac. Sci. Kyoto Univ. (Ser. Biol.)* **16**: 49-54

Rajbanshi KJ (2001) *Zoo-geographical distribution and the status of coldwater fish in Nepal*. Paper presented at: The symposium on cold water fishes of trans-Himalayan region, 10-13 July 2001, Kathmandu, Nepal.

Ramadan HAI and Baeshen NA (2012) Biological Identifications Through DNA Barcodes. Biodiversity Conservation and Utilization in a Diverse World. Chapter 5; 109-128

Ratnasingham S, Hebert PDN (2013) A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE* **8**(8): e66213

Ratnasingham S and Hebert PDN (2007) BOLD: the Barcode of Life data system (http://www.barcodinglife.org). *Mol. Ecol. Notes* **7**: 355–364

Ross HA, Lento GM,Dalebout ML,Goode M,Ewing G, McLaren P, Rodrigo AG, Lavery S, Baaker CS (2003) DNA Surveillance: Web-Based Molecular Identification of Whales, Dolphins, and Porpoises. *J Hered* **94** (2):111–114

Ross HA, Murugan S, Li WL (2008) Testing the reliability of genetic methods of species identification via simulation. *Syst Biol* **57**: 216–230

Ross JJ, Mabragana E, Castro MG, Diaz de Astarloa M (2012) DNA barcoding Neotropical fishes: recent advances from the Pampa Plain, Argentina. *Mol Ecol Res* **12**: 999-1011

Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene*. **238**:195–209

Saitou N and Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406-425

Saitou N, Ueda S (1994) Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Mol. Biol. Evol.* **11**(3):504-512

Saraste M (1994) Structure and evolution of cytochrome oxidase. *J Antonie van Leeuwenhoek* **65**(4): 285-287

Schizas NV (2012) Misconceptions regarding nuclear mitochondrial pseudogenes (Numts) may obscure detection of mitochondrial evolutionary novelties. *Aquat Biol* **17**: 91–96

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding Consortium (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *PNAS Early Edition, Microbiology.*

Sevilla R, Diez A, Noren M, Mouchel O, Jerome M et al. (2007) Primers and polymerase chain reaction conditions for DNA barcoding teleost fish based on the mitochondrial cytochrome b and nuclear rhodopsin genes. *Mol Ecol Notes* **7(**5): 730-734

Shrestha J., Taxonomic revision of cold water fishes of Nepal

Shrestha J (2001) Cold water fish and fisheries in Nepal. Paper presented at: The symposium on cold water fishes of trans-Himalayan region, 10-13 July 2001, Kathmandu, Nepal.

Shrestha MK, Batajoo RK, Karki GB (2001) Prospects of fisheries enhancement and aquaculture in lakes and reservoirs of Nepal. Paper presented at: The symposium on cold water fishes of trans-Himalayan region, 10-13 July 2001, Kathmandu, Nepal.

Siti-Balkhis AB, Jamsari AFJ, Hwai TS, Yasin Z, Siti-Azizah MN (2011) Evidence of geographical structuring in the Malaysian Snakehead, Channa striata based on partial segment of the COI gene. *Genet Mol Biol* **34**(3):520-523

Sjodin P, Bataillon T, Schierup MH (2010) Insertion and Deletion Processes in Recent Human History. *PLoS ONE* **5**(1): e8650

Soltis DE and Soltis PS (2003) The role of phylogenetics in comparative genomics. *Plant Physiol* **132**: 1790–1800

Steinke D, Hanner R (2010) The FISH-BOL collaborator's protocol. *Mitochondrial DNA* **22**(S1):10-14

Stoeckle M and Ausubel JH (2003) Barcode of life. Draft Scientific Rationale and Strategy: 'DNA and Taxonomy' conference, 9-12 March 2003, Cold Spring Harbor Laboratory.

Stoeckle M, Waggoner P, Ausubel JH (2004) Barcoding life: Ten reasons. DNA and Taxonomy' conference, Cold Spring Harbor Labotratory.

Strauss RE and Bond CE Taxonomic Methods: Morphology. Chapter 4:109-140pp

Subba BR, Gosh TK (1996) A new record of the pigmy barb, *Puntius phutunio* (Ham.) from Nepal. *Journal of Freshwater Biology,* India **8(3)**: 159-161.

Swar DB, Fernando CH (1980) Some studies on the ecology of limnetic crustacean zooplankton in Lake Begnas and Rupa, Pokhara valley, Nepal. *Hydrobiologia* **70**(3): 235-245

Swar DB, Gurung TB (1988) Introduction and cage culture of exotic carp and their impact on fish harvested in Lake Begnas, Nepal. *Hydrobiologia* **166**(3): 277-283

Swartz ER, Mwale M, Hanner R (2008) A role for barcoding in the study of African fish diversity and conservation. *S Afr J Sci* **104**: 293-298

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. *Mol Biol Evol* **1**:269-285

Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *PNAS* **101**:11030-11035

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* **28**: 2731-2739

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy. *Trends Ecol Evol* **18**(2): 70-74

Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S (1996) The whole structure of the 13-subunit oxidized Cytochrome c Oxidase at 2.8 A. *Science* **272**(5265): 1136-1144

Valdez-Moreno M, Ivanova NV, Elias-Gutierrez M, Contreras-Balderas S, Hebert PDN (2009) Probing diversity in freshwater fishes from Mexico and Guatemala with DNA barcodes. *J Fish Biol* **74**: 377-402

Vallinoto M, Sena L, Sampaio I, Schneider H, and Schneider MP (2000) Mitochondrial DNA-like sequence in the nuclear genome of *Saguinus* (Callitrichinae, Primates): transfer estimation. *Genet Mol Biol* **23**(1): 35-42

Vishwanathan R, Pillai VK (1956) AP ER chromatography in fish taxonomy. *Proc Ind Acad Sci B* **43**(6): 334-339

Ward RD, Holmes BH (2007) An analysis of nucleotide and amino acid variability in the barcode region of cytochrome c oxidase I (cox1) in fishes. *Mol Ecol Notes* **7:** 899-907

Ward RD, Zemlak TS, Innes BH, PR Last, Hebert PDN (2005) DNA barcoding Australia's fish species. *Philos Trans R Soc B* **360**: 1847–1857

Wiens JJ (2007) Species Delimitation: New Approaches for Discovering Diversity. *Syst. Biol.* **56**(6):875–878

Wong EHK, Hanner EH (2008) DNA barcoding detects market substitution in North American seafood. *Food Res Int* **41**: 828-837

Wong LL, Peatman E, Lu J, Kucuktas H, He S, Zhou C, Na-nakorn U, Liu Z (2011) DNA Barcoding of Catfish: Species Authentication and Phylogenetic Assessment. *PLoS ONE* **6**(3): e17812

Zhang AB, Sikes DS, Muster C, Li SQ (2008) Inferring Species Membership Using DNA Sequences with Back-Propagation Neural Networks. *Syst. Biol.* **57**(2):202–215

Zhang J, Hanner R (2012) Molecular approach to the identification of fish in the South China Sea. *PLoS ONE* **7**(2)

Websites

http://ghr.nlm.nih.gov/gene/MT-CO1

www.codoncode.com

www.dnabarcoding101.org/introduction.html

www.fishbase.org

www.megasoftware.net

http://www.barcodeoflife.org

http://ibol.org

http://www.angelfire.com/biz/piranha038/taxon.html

http://www.aquaticcommunity.com/fishtaxonomy/scientificclassification.php

http://www.biologyreference.com/Ta-Va/Taxonomy-History-of.html

http://davesgarden.com/guides/articles/view/2051/

http://www.reasons.org/articles/status-update-the-latest-on-neanderthals

http://www.shmoop.com/taxonomy/taxonomy-history.html

http://www.thekingdomofnepal.com/

http://www.worldfishcenter.org/fishbase

http://davesgarden.com/guides/articles/view/2051/

APPENDICES

Appendix 1: Composition of various reagents used.

CTAB Buffer (10X)

For 50 ml

Tris HCl (1M) – 5 ml

0.5 M EDTA – 2 ml

5M NaCl – 17.5 ml

10% CTAB – 10 ml

d/w – 15.5 ml to make volume

pH – 7.5-8.0

All the components are added into a measuring cylinder. Only 45 ml vol. is maintained. The pH is maintained at 7 and autoclaved. After autoclaving, again pH is maintained around 7.7-7.8 and final volume is made 50 ml by addition of autoclaved d/w.

TBE Buffer (5X)

Tris Base – 54 g

Boric acid – 27.5 g

0.5 M EDTA (pH-8.0) – 20 ml

Final Vol. up to 1 liter (final pH-8.0)

Gel loading dye (6X)

10mM Tris (pH-8.0)

0.03% Bromophenol blue

60% Glycerol

60 mM EDTA

TE Buffer

10 mM Tris HCl (pH-8.0)

10 mM EDTA (pH-8.0)

APPENDIX 2: Classification of fishes studied in the research along with the IUCN conservation status.

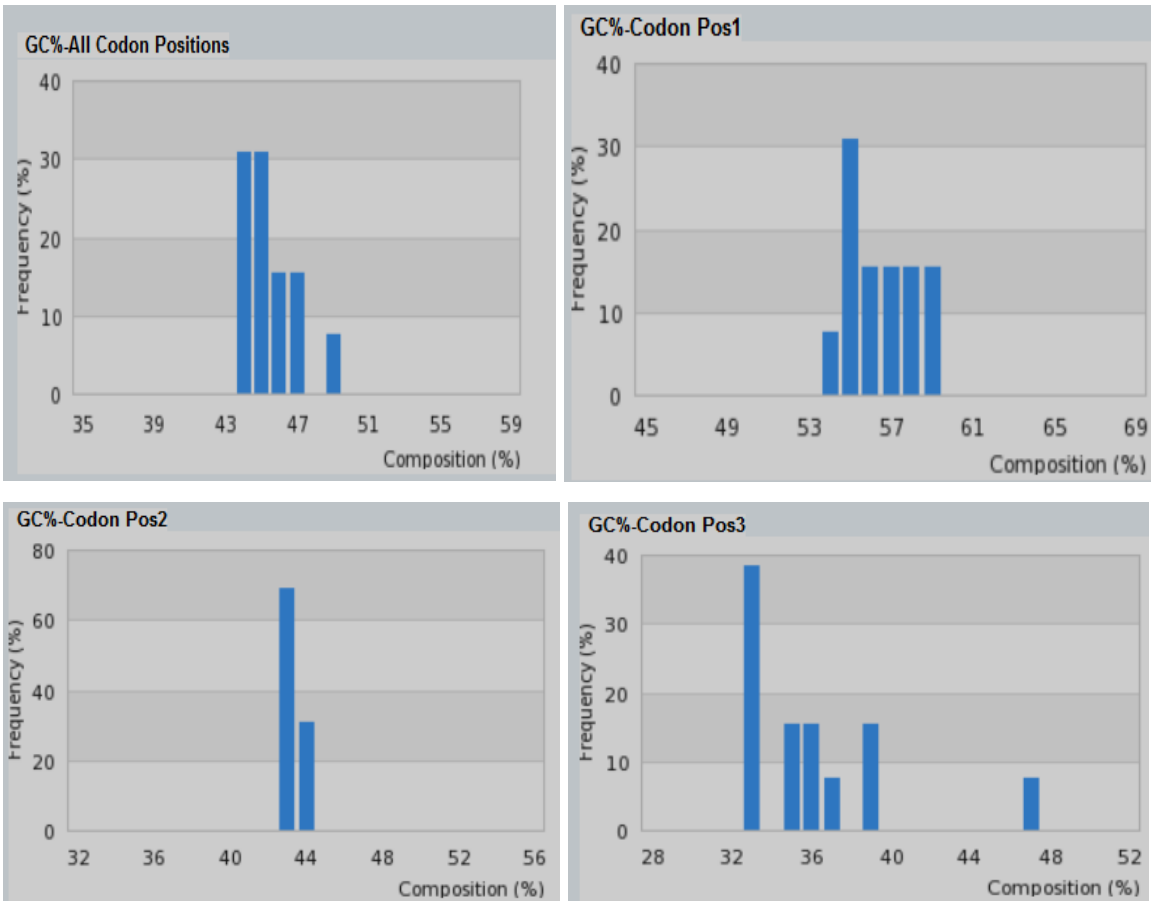
Sample code	Classification	Local name	IUCN Conservation status
PVF1	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order :Synbranchiformes Family :Mastacembelidae Genus : *Mastacembelu*s Species : *M. armatus*	Raj Bam	Least concern
PVF2	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Siluriformes Family : Clariidae Genus : *Clarias* Species : *C. batrachus*	Mahur	Least concern
PVF3	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Beloniformes Family : Belonidae Genus : *Xenentodon* Species : *X. cancila*	Dunge/Chuche Bam	Least concern
PVF4	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Cypriniformes Family : Cyprinidae Genus : *Tor* Species : *T. putitora*	Sahar	Endangered
PVF5	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Siluriformes Family : Bagridae Genus : *Mystus* Species : *M. cavasius*	Junge	Least concern
PVF7	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Cypriniformes Family : Cyprinidae Genus : *Cirrhinus* Species : *C. mrigala*	Naini	Least concern
PVF8	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Cypriniformes	Rewa	Least concern

	Family : Cyprinidae Genus : *Chagunius* Species *: C. chagunio*		
PVF9	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Perciformes Family : Channidae Genus : *Channa* Species : *C. orientalis*	Bhoti	Not accessed in IUCN list yet
PVF10	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Perciformes Family : Cichlidae Genus : *Oreochromis* Species : *O. mossambicus*	Tilapia	Near threatened
PVF11	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Cypriniformes Family : Cyprinidae Genus : *Hypopthalmichthys* Species : *H. nobilis*	Big Head	Not accessed in IUCN list yet
PVF12	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Cypriniformes Family : Cyprinidae Genus : *Pethia/Puntius* Species : *P. ticto*	_	Least concern
PVF13	Kingdom : Animal Phylum : Chordata Class : Actinopterygii Order : Cypriniformes Family : Cyprinidae Genus : *Barilius* Species : *B. vagra*	Faktar/Poti	Least concern
PVF14	Kingdom : Animal Phylum : Chordata Class :Actinopterygii Order : Cypriniformes Family : Cyprinidae Genus : *Labeo* Species : *L. rohita*	Rohu	Least concern

Appendix 3: Concentration and absorbance of the extracted DNA, along with the dilution to be made for amplification.

Sample ID	Conc. (ng/µl)	A260	A280	260/280	Conc. reqd.	260/230	Vol. to make	Original conc. taken	Diluent added
PVF1	1396.79	27.936	13.770	2.03	100	2.32	25	1.79	23.21
PVF2	2582.35	51.647	26.013	1.99	100	2.28	25	0.968	24.032
PVF3	658.31	13.166	6.526	2.02	100	2.24	25	3.798	21.202
PVF4	554.52	11.090	5.529	2.01	100	2.27	25	4.509	20.491
PVF5	219.26	4.385	2.183	2.01	100	2.29	25	11.399	13.601
PVF6	225.34	4.507	2.145	2.10	100	2.9	25	11.096	13.904
PVF7	391.84	7.837	3.929	1.99	100	2.28	25	6.38	18.62
PVF8	1734.64	34.693	16.722	2.07	100	2.31	25	1.44	23.56
PVF9	842.05	16.841	8.362	2.01	100	2.39	25	2.969	22.031
PVF10	920.73	18.415	9.238	1.99	100	2.37	25	2.715	22.285
PVF11	87.77	1.755	0.772	2.27	100	2.31	25	28.484	-3.484
PVF12	249	4.980	2.271	2.19	100	2.25	25	10.04	14.96
PVF13	250.19	5.004	2.430	2.06	100	2.08	25	9.992	15.008
PVF14	92.03	0.860	0.343	2.51	100	2.26	25	27.165	-2.165

Appendix 4: Percent GC content in fishes studied.



Appendix 5: Similarities of the species with one another in percentage using Clustal W.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1: PVF9_Channa	100.00	78.59	77.70	76.79	75.32	76.46	76.46	78.25	77.76	76.95	75.83	74.96	77.92
2: PVF10_Oreochromis	78.59	100.00	80.53	79.90	78.43	78.59	77.12	80.72	80.88	77.94	79.17	76.54	79.74
3: PVF1_Mastacembelus	77.70	80.53	100.00	78.52	77.21	78.20	80.33	81.48	81.15	79.34	79.78	78.98	79.18
4: PVF5_Mystus	76.79	79.90	78.52	100.00	76.38	78.22	79.29	80.37	79.45	79.75	81.97	76.94	78.99
5: PVF13_Barilius	75.32	78.43	77.21	76.38	100.00	79.60	81.13	82.67	79.45	80.67	81.97	73.81	76.38
6: PVF8_Chagunius	76.46	78.59	78.20	78.22	79.60	100.00	82.67	84.05	83.44	82.52	81.79	76.44	80.06
7: PVF12_Puntius	76.46	77.12	80.33	79.29	81.13	82.67	100.00	83.59	85.28	85.12	84.52	77.10	79.75
8: PVF11_Hypopthalmichthys	78.25	80.72	81.48	80.37	82.67	84.05	83.59	100.00	86.50	84.51	83.24	78.25	80.37
9: PVF4_Tor	77.76	80.88	81.15	79.45	79.45	83.44	85.28	86.50	100.00	87.42	86.34	77.27	78.68
10: PVF7_Cirrhinus	76.95	77.94	79.34	79.75	80.67	82.52	85.12	84.51	87.42	100.00	87.98	75.62	79.29
11: PVF14_Labeo	75.83	79.17	79.78	81.97	81.97	81.79	84.52	83.24	86.34	87.98	100.00	80.15	79.23
12: PVF2_Clarias	74.96	76.54	78.98	76.94	73.81	76.44	77.10	78.25	77.27	75.62	80.15	100.00	79.08
13: PVF3_Xenentodon	77.92	79.74	79.18	78.99	76.38	80.06	79.75	80.37	78.68	79.29	79.23	79.08	100.00

Appendix 5: CLUSTAL O (1.2.0) multiple sequence alignment

```
PVF9_C.orientalis        ------CCTTTATATAGTATTTGGTGCTTGGGCTGGAATAGTCGGCACCGCACTGAGCCT      54
PVF10_O.mossambicus      ------CCTCTATCTAGTATTTGGTGTCTGAGCCCGAATAGTAGGAACTGGGTTTAGCCT      54
PVF1_M.armatus           ------------------------------------------------GGGCCTAAGCCT      12
PVF5_M.cavasius          ------CCTTTTACTTGTATTCGGTGCCTGAGCCGGAATAGTTGGTACAGCCCTTAGCTT      54
PVF13_B.vagra            ------CTTTTATATAATATTTGCCTCTCGAGCCTCTATAGTATTAACGGCCCTAAGTCT      54
PVF8_C.chagunio          ------CTTTTATCTTGTATTTGGTGCCTGAGCCCGGATAGTAGGAACTGCTTTAAGTCT      54
PVF12_P.ticto            ------CCTTTATCTTGTATTCGGTGCCTGAGCCGGAATGGTAGGAACCGCCCTGAGCCT      54
PVF11_H.nobilis          ------CCTTTATCTTGTATTTGGTGCCTGAGCCGGAATAGTGGGAACCGCCCTAAGCCT      54
PVF4_T.putitora          ------CCTTTATCTTGTATTTGGTGCCTGAGCCGGAATAGTGGGAACCGCCTTAAGCCT      54
PVF7_C.mrigala           ------CTTTTATCTTGTATTTGGTGCCTGAGCCGGAATAGTAGGAACTGCCTTAAGCCT      54
PVF14_L.rohita           ------------------------------------------------------------      0
PVF2_C.batrachus         TACGGGCCTAAACTTACTAATCCGGGCAAAACTGGCA-CAACCCGGGGCTCTTT------      53
PVF3_X.cancila           ------CCTAATATTAGTATTTGGTGCCTGAGCTGGAATAGTAGGCACTGGCTTTAGCCT      54

PVF9_C.orientalis        ACTGATCCGGGCTGAACTTAGCCAGCCCGGTGCTCTTCTAGGCAACGACCAAATTTATAA      114
PVF10_O.mossambicus      CCTAATTCGGGCAGAACTAAACCAGCCCGGCTCTCTCCTCGGAGACGACCAGATTTATAA      114
PVF1_M.armatus           ACTCATCCGGGCAGAACTAAGCCAACCCGGCGCTTTATTGGGTGACGATCAAATTTATAA      72
PVF5_M.cavasius          ACTAATTCGGGCGGAACTAGCCCAACCCGGCGCACTTCTTGGCGACGATCAGATTTATAA      114
PVF13_B.vagra            TCTTATTCGAGCTGAACTAAGTCAGCCCGGGTCACTTCTGGGTGATGACCAAATCTACAA      114
PVF8_C.chagunio          CCTCATTCGAGCCGAACTGAGCCAACCCGGATCACTTCTAGGCGACGATCAAATCTACAA      114
PVF12_P.ticto            CCTTATCCGAGCCGAACTAAGTCAACCAGGATCACTCCTAGGTGATGATCAAATTTATAA      114
PVF11_H.nobilis          TCTCATTCGAGCCGAACTAAGCCAACCCGGATCACTTCTGGGCGATGACCAAATTTATAA      114
PVF4_T.putitora          TCTCATCCGGGCTGAACTAAGCCAACCCGGATCGCTTCTAGGTGATGACCAAATTTATAA      114
PVF7_C.mrigala           TCTTATTCGGGCCGAGCTAAGCCAACCCGGATCGCTTCTAGGCGACGACCAAATTTACAA      114
PVF14_L.rohita           ------------------------------------------------TACAATTATAA      11
PVF2_C.batrachus         ---------------------------------------TAGGAGATGACCAGATTTATAA      75
PVF3_X.cancila           TCTTATTCGAGCAGAACTGAGCCAACCGGGCTCCCTCTTAGGCGATGACCAAATTTACAA      114
                                                                 * *: ** **

PVF9_C.orientalis        TGTAATTGTTACGGCCCACGCCTTCGTCATGATCTTCTTCATGGTAATGCCAATAATAAT      174
PVF10_O.mossambicus      TGTAATTGTTACAGCACATGCTTTCGTAATAATTTTCTTTATAGTAATGCCAATTATAAT      174
PVF1_M.armatus           TGTAATCGTTACAGCACATGCTTTCGTAATAATTTTCTTTATAGTAATACCAATTATAAT      132
PVF5_M.cavasius          TGTTATTGTAACTGCTCATGCCTTTATCATAATTTTCTTTATAGTAATACCAATCATAAT      174
PVF13_B.vagra            TGTTATTGTTACTGCCCATGCTTTTGTAATGATTTTCTTTATAGTGATGCCAATTCTTAT      174
PVF8_C.chagunio          TGTAATCGTTACCGCCCATGCTTTCGTAATAATTTTCTTTATAGTTATACCCATTCTTAT      174
PVF12_P.ticto            TGTAATCGTCACTGCTCACGCCTTCGTAATAATTTTCTTTATAGTTATGCCCATCCTGAT      174
PVF11_H.nobilis          CGTTATTGTTACTGCCCATGCCTTCGTAATAATTTTCTTTATAGTGATACCAATCCTTAT      174
PVF4_T.putitora          TGTTATCGTCACTGCTCACGCCTTCGTAATAATCTTCTTTATAGTAATACCCATTCTCAT      174
PVF7_C.mrigala           TGTCATCGTCACTGCTCACGCCTTCGTGATAATTTTCTTTATAGTAATGCCCATCCTCAT      174
PVF14_L.rohita           TGTTATTGTAACTGCCCACGCCTTCGTAATAATTTTCTTTATAGTAATGCCCATCCTCAT      71
```

```
PVF2_C.batrachus          TGTTATTGTTACTGCCCACGCCTTCGTAATAATTTTCTTTATAGTAATACCAATTATGAT      135
PVF3_X.cancila            TGTTATTGTAACAGCACATGCCTTTGTAATAATTTTCTTTATAGTAATACCAATTATAAT      174
                           ** ** ** ** ** ** ** ** .* **.** ***** **.** **.**.** .* **

PVF9_C.orientalis         CGGGGGCTTTGGAAACTGACTGGTCCCGCTTATGATCGGCGCCCCTGACATAGCCTTCCC      234
PVF10_O.mossambicus       TGGAGGTTTTGGAAACTGACTAGTGCCACTAATGATTGGTGCACCAGACATGGCCTTCCC      234
PVF1_M.armatus            TGGGGGTTTTGGTAACTGACTTATTCCTCTAATAATTGGTGCACCAGATATAGCATTCCC      192
PVF5_M.cavasius           CGGAGGCTTTGGAAATTGACTCGTACCACTAATGATTGGAGCGCCAGACATGGCATTCCC      234
PVF13_B.vagra             TGGAGGTTTTGGAAACTGACTAGTCCCGCTAATGATTGGGGCTCCAGACATAGCATTCCC      234
PVF8_C.chagunio           TGGAGGCTTTGGAAACTGATTAGTACCCCTCATAATTGGAGCCCCCGATATAGCATTCCC      234
PVF12_P.ticto             CGGCGGATTCGGAAACTGACTAGTCCCCTTAATAATCGGAGCCCCCGATATAGCATTCCC      234
PVF11_H.nobilis           TGGAGGATTTGGAAACTGACTCGTGCCACTAATGATTGGAGCACCTGATATAGCATTCCC      234
PVF4_T.putitora           TGGAGGATTTGGAAACTGGCTTGTACCACTAATAATTGGAGCCCCAGACATGGCGTTCCC      234
PVF7_C.mrigala            TGGAGGATTTGGAAATTGACTTGTCCCATTAATGATTGGGGCCCCAGACATAGCATTCCC      234
PVF14_L.rohita            TGGAGGATTTGGGAACTGACTCGTGCCACTAATGATTGGAGCCCCAGACATGGCATTCCC      131
PVF2_C.batrachus          TGGAGGTTTCGGAAACTGACTTGTACCTCTAATAATCGGTGCCCCAGATATAGCATTCCC      195
PVF3_X.cancila            TGGAGGCTTCGGTAACTGACTAATCCCTCTTATAATTGGAGCCCCAGATATAGCATTTCC      234
                           ** ** ** ** ** **. * .* **  * **.** ** ** ** ** **.** ** **

PVF9_C.orientalis         TCGAATAAACAATATGAGTTTTTGACTTCTCCCCCCTTCTTTCCTCCTTCTTCTGGCCTC      294
PVF10_O.mossambicus       TCGAATAAATAACATGAGTTTTTGACTCCTCCCCCCCTCATTTCTCCTTCTCCTCGCCTC      294
PVF1_M.armatus            CCGAATGAACAACATAAGCTTTTGACTTCTCCCTCCCTCTTTCCTCCTACTACTAGCCTC      252
PVF5_M.cavasius           ACGAATAAACAACATGAGTTTCTGACTACTCCCTCCCTCCTTCTTACTATTATTAGCCTC      294
PVF13_B.vagra             TCGAATAAATAATATAAGTTTTTGACTCCTGCCCCCATCATTCCTTCTATTATTGGCCTC      294
PVF8_C.chagunio           ACGAATAAACAATATAAGCTTTTGACTATTACCCCCTTCATTCTTACTGCTTTTAGCCTC      294
PVF12_P.ticto             ACGAATAAATAACATAAGTTTTTGACTTCTACCACCCTCATTCCTACTATTATTAGCCTC      294
PVF11_H.nobilis           ACGAATAAATAATATAAGCTTTTGACTCCTGCCCCCCTCTTTCCTTCTACTACTAGCCTC      294
PVF4_T.putitora           GCGTATAAACAACATAAGCTTTTGACTACTACCCCCATCATTTCTTCTTCTACTAGCCTC      294
PVF7_C.mrigala            CCGTATAAACAACATAAGCTTCTGACTTCTACCCCCATCATTCCTGCTACTACTAGCCTC      294
PVF14_L.rohita            CCGTATAAACAACATAAGCTTCTGACTCCTACCCCCATCATTCCTATTACTATTAGCCTC      191
PVF2_C.batrachus          ACGAATAAATAATATAAGCTTCTGATTACTACCCCCCTCCTTCCTACTGCTACTTGCCTC      255
PVF3_X.cancila            ACGAATAAACAACATAAGCTTCTGACTTCTCCCCCCATCTTTCTTACTTCTTTTGGCCTC      294
                           **:**.** ** **.** ** *** *  * ** ** ** **  *  *  *  * *****

PVF9_C.orientalis         TTCTGCAGTAGAAGCCGGAGCTGGGACAGGCTGGACAGTTTACCCACCTTTAGCTGGCAA      354
PVF10_O.mossambicus       ATCCGGGGTCGAAGCAGGGGCCGGTACAGGATGGACTGTTTATCCCCCACTCGCAGGCAA      354
PVF1_M.armatus            CTCAGCAGTGGAGTCCGGAGCAGGAACAGGATGAACAGTGTACCCACCCCTGGCCAGCAA      312
PVF5_M.cavasius           CTCTGGTGTTGAAGCTGGTGCAGGTACAGGATGAACCGTTTATCCCCCCCTTGCTGGCAA      354
PVF13_B.vagra             CTCTGGTGTAGAAGCCGGTGCCGGAACAGGATGAACAGTTTATCCCCCACTAGCAGGAAA      354
PVF8_C.chagunio           CTCTGGTGTTGAAGCCGGAGCCGGAACAGGATGGACAGTATATCCGCCTCTGGCAGGTAA      354
PVF12_P.ticto             CTCTGGTGTTGAAGCCGGAGCAGGGACAGGGTGAACAGTTTACCCGCCGCTAGCAGGAAA      354
PVF11_H.nobilis           TTCTGGTGTCGAGGCCGGGGCCGGAACAGGATGAACAGTTTACCCGCCACTCGCGGGTAA      354
PVF4_T.putitora           TTCCGGTGTTGAAGCTGGAGCTGGGACAGGATGAACAGTTTATCCGCCTCTCGCAGGCAA      354
PVF7_C.mrigala            TTCTGGTGTTGAAGCTGGAGCTGGAACAGGATGAACGGTATACCCGCCTCTTGCAGGAAA      354
PVF14_L.rohita            TTCCGGTGTAGAAGCTGGAGCTGGGACAGGATGGACAGTATACCCACCTCTTGCAGGCAA      251
PVF2_C.batrachus          ATCAGGCGTTGAAGCGGGGGCAGGAACAGGGTGAACAGTATACCCACCCCTTGCAGGAAA      315
PVF3_X.cancila            ATCAGGAGTTGAAGCCGGTGCAGGAACTGGGTGGACAGTCTACCCCCCTCTAGCCGGTAA      354
                           ** *  ** **. * ** ** ** **:** **.** ** ** ** **  * ** .* **

PVF9_C.orientalis         TCTGGCTCACGCGGGAGCATCCGTAGACCTAGCCATCTTCTCTTTACACCTTGCAGGTGT      414
PVF10_O.mossambicus       TCTCGCCCATGCTGGGCCTTCCGTTGACTTAACCATCTTCTCCCTCCACTTGGCCGGGGT      414
PVF1_M.armatus            CCTTGCCCACGCAGGGGCCTCCGTAGATCTGACAATCTTCTCCCTCCACTTAGCAGGTGT      372
PVF5_M.cavasius           TCTTGCACACGCCGGAGCCTCAGTAGATTTAACTATTTTCTCCCTACATCTTGCAGGGGT      414
PVF13_B.vagra             CCTGGCCCACGCAGGAGCATCAGTAGACCTAACAATTTTCTCTCTTCACTTGGCGGGTGT      414
PVF8_C.chagunio           CCTAGCCCACGCAGGGGCATCCGTCGACCTAACCATCTTTTCTTTACACTTAGCTGGTGT      414
PVF12_P.ticto             CCTGGCCCATGCTGGAGCGTCAGTAGACCTAACAATTTTTTCACTTCACTTAGCAGGTGT      414
PVF11_H.nobilis           TCTTGCTCACGCAGGAGCATCCGTAGACCTAACAATTTTCTCCCTCCACTTAGCAGGTGT      414
PVF4_T.putitora           CCTAGCCCACGCAGGAGCATCAGTAGATCTAACAATCTTTTCACTCCACTTAGCAGGTGT      414
PVF7_C.mrigala            TTTAGCCCACGCAGGAGCATCGGTAGACCTAACAATTTTCTCACTTCACTTAGCGGGTGT      414
PVF14_L.rohita            CTTAGCCCACGCAGGAGCATCAGTAGACCTAACAATTTTCTCACTTCACTTAGCAGGAGT      311
PVF2_C.batrachus          CCTGGCACATGCAGGAGCTTCTGTAGACTTAACCATTTTTTCTCTACATCTAGCAGGTGT      375
PVF3_X.cancila            TCTAGCACATGCTGGAGCATCTGTTGATCTAACCATTTTTTCTTTACACTTGGCAGGTGT      414
                            * ** ** ** **. * ** ** **  *..* ** ** **  * **  * ** ** **

PVF9_C.orientalis         CTCTTCAATTTTAGGGGCAATTAACTTCATCACAACGATCATTAACATGAAACCCCCAGC      474
PVF10_O.mossambicus       GTCATCTATTTTAGGTGCAATTAATTTTATTACAACCATTATTAACATAAAACCCCCTGC      474
PVF1_M.armatus            CTCATCAATTCTTGGGGCAATCAACTTTATTACAACAATTGTTAATATAAAACCTCCTGC      432
PVF5_M.cavasius           ATCATCCATTCTAGGGGCTATTAATTTCATTACAACTATTATTAATATGAAACCCCCAGC      474
PVF13_B.vagra             ATCGTCCATTTTAGGGGCAATTAACTTTATCACCACAACCATTAATATAAAACCCCCCGC      474
PVF8_C.chagunio           TTCATCAATTCTGGGAGCAATTAATTTCATCACCACAATTATTAATATGAAACCTCCAGC      474
PVF12_P.ticto             TTCATCAATTCTTGGGGCAATTAACTTTATTACTACAACTATTAATATGAAACCCCCAGC      474
PVF11_H.nobilis           ATCATCAATTTTAGGGGCAATTAACTTCATCACCACAACTATTAACATAAAACCACCAGC      474
PVF4_T.putitora           CTCATCAATTCTGGGGGCAATTAACTTTATTACTACAACAGTTAACATAAAACCCCCAGC      474
PVF7_C.mrigala            TTCATCAATTCTAGGGGCTATTAATTTTATTACCACAACCATCAACATGAAGCCCCCAGC      474
PVF14_L.rohita            TTCATCAATTCTAGGGGCTATTAATTTTATTACTACAACTATTAATATGAAACCTCCAGC      371
PVF2_C.batrachus          ATCATCAATTCTTGCCTCCATTAACTTTATCACAACCATTATTAACATGAAACCGCCAGC      435
PVF3_X.cancila            TTCATCAATTTTAGGGGCCATTAATTTTATTACCACTATTATTAATATGAAACCCCCAGC      474
                           ** ** *** * *   * ** ** ** ** ** ** *  .* ** **.**.** ** **

PVF9_C.orientalis         CATCTCTCAGTACCAGACACCTCTGTTTGTATGGGCCATCCTAATCACTGCCATCCTTCT      534
PVF10_O.mossambicus       CATCTCCCAATATCAAACACCCCTCTTTGTATGATCCGTTCTAATTACCGCAGTACTACT      534
PVF1_M.armatus            CATTTCACAATACCAAACCCCTCTCTTTGTATGAGCCCTACTAATTACTGCTGTTCTACT      492
PVF5_M.cavasius           CATCTCCCAATATCAGACTCCCTTATTTGTATGAGCTGTATTAATTACAGCAGTCCTCTT      534
```

```
PVF13_B.vagra           TATTTCCCAATACCAAACACCCCTGTTCGTCTGAGCTGTTCTTGTAACAGCCGTATTACT      534
PVF8_C.chagunio         TATCTCCCAATATCAAACACCCTTATTTGTGTGATCTGTGCTTGTAACTGCCGTGCTGCT      534
PVF12_P.ticto           CACTACCCAGTACCAAACACCACTGTTCGTCTGATCCGTACTTGTAACTGCCGTCCTACT      534
PVF11_H.nobilis         CATTTCCCAATATCAAACACCTCTCTTTGTTTGAGCTGTGCTTGTAACGGCCGTACTTCT      534
PVF4_T.putitora         TATTTCCCAATATCAAACACCTTTATTTGTTTGATCCGTACTTGTAACCGCCGTACTACT      534
PVF7_C.mrigala          CATCTCACAATACCAAACACCTCTGTTCGTTTGATCCGTGCTCGTAACCGCCGTACTGCT      534
PVF14_L.rohita          CATCTCACAATATCAAACACCCTTATTCGTCTGATCTGTCCTAGTAACCGCCGTACTACT      431
PVF2_C.batrachus        CATCTCCCAATATCAAACACCCCTATTTGTTTGATCCGTAATAATCACAGCAGTACTACT      495
PVF3_X.cancila          AATCTCCCAGTACCAAACACCCCTTTTCGTTTGAGCTGTTTTAATTACTGCTGTCCTACT      534
                         *  :* **.** **.** **  * ** ** **. *  *  * .* ** ** .*  *  *

PVF9_C.orientalis       ACTTCTTTCTCTTCCCGTTTTAGCCGCCGGTATCACAATACTATTAACAGACCGAAACTT      594
PVF10_O.mossambicus     CCTACTATCCCTACCCGTTCTTGCCGCCGGCATCACAATACTTCTAACAGACCGAAACCT      594
PVF1_M.armatus          CCTTCTATCTCTTCCAGTCCTGGCCGCCGGTATCACAATGCTTTTAACAGACCGAAATCT      552
PVF5_M.cavasius         ACTACTTTCCCTCCCAGTTCTGGCTGCCGGTATCACAATACTACTAACAGATCGAAACCT      594
PVF13_B.vagra           CCTCTTATCACTACCGGTCCTAGCTGCCGGCATCACGATGCTTCTTACAGATCGAAACCT      594
PVF8_C.chagunio         TCTTTTATCCCTTCCAGTTTTAGCCGCAGGAATTACAATACTTCTAACAGATCGTAACCT      594
PVF12_P.ticto           CCTACTATCACTACCAGTCTTGGCCGCGGGGATTACAATGCTTCTAACAGATCGAAACCT      594
PVF11_H.nobilis         TCTCCTATCCTTACCAGTTTTAGCTGCTGGAATTACAATACTCCTTACAGACCGTAATCT      594
PVF4_T.putitora         ACTCCTATCATTGCCAGTTCTAGCCGCTGGGATTACAATACTTCTAACAGACCGAAACCT      594
PVF7_C.mrigala          TCTTCTATCACTGCCAGTCCTAGCTGCTGGTATTACAATGCTCTTAACAGATCGAAACCT      594
PVF14_L.rohita          TCTCCTCTCACTACCAGTACTGGCCGCTGGAATCACAATGCTTTTAACAGATCGAAATCT      491
PVF2_C.batrachus        ACTTCTGTCCCTTCCAGTATTAGCTGCGGGAATCACTATATTATTAACAGACCGTAATTT      555
PVF3_X.cancila          CCTTCTCTCCTTACCAGTTTTAGCTGCTGGGATTACAATACTCTTAACAGACCGAAATCT      594
                         **  * **  * ** **  * ** ** ** ** ** **. *  *:***** **:**  *

PVF9_C.orientalis       AAACACAACCTTTTTGAAACCG-------------------------------------      616
PVF10_O.mossambicus     AAACACAACCTTTTTTGA-----------------------------------------      612
PVF1_M.armatus          TAACACCACATTCTTTGACCCTGCAGGAGGCGGAGACCCAATCCTATATCAACACTTA        610
PVF5_M.cavasius         CAATACCACATTCTTCGACCCAGCAGGAGGAGAGGACCCAATTCTTTATCAACACCTA        652
PVF13_B.vagra           CAATACCTCTTTCTTCGATCCTGCCGGACCAGGGGATCCTATCCTTTACCAACACCTA        652
PVF8_C.chagunio         CAACACCACATTCTTCGACCCGGCAGGGGGTGGGGATCCAATTTTATATCAACACCTG        652
PVF12_P.ticto           TAATACCACATTCTTCGACCCGGCAGGGGGAGGAGACCCAATCCTCTATCAACACCTA        652
PVF11_H.nobilis         TAACACTACATTCTTTGACCCAGCAGGGGGAGGAGACCCAATCCTATATCAACACCTA        652
PVF4_T.putitora         TAACACAACATTCTTTGACCCGGCAGGTGGAGGAGACCCAATTCTGTACCAACACCTA        652
PVF7_C.mrigala          TAATACCACATTCTTCGACCCAGCAGGAGGGGGAGACCCAATTCTCTACCAACACTTA        652
PVF14_L.rohita          GAATACTACATTCTTCGACCCGGCAGGACGAGGGGACCCAATCCTTTATCAACACCTA        549
PVF2_C.batrachus        AAACACAACCTTCTTTGATCCTGCGGGAGGGGGGGACCCAATCCTTTATCAACACCTC        613
PVF3_X.cancila          AAACACCACCTTCTTTGACCCCGCTGGGGGAGGTGACCCCATCCTCTACCAACATCTC        652
                           ** ** :* ** ** .*
```

Appendix 6: Species identification using reference library from BOLD system.

Appendix 7: List of all the specimens belonging to 3 families Cypriniformes, Perciformes and Siluriformes extracted from GenBank for analysis with the accession numbers.

Cypriniformes:

Name of species	NCBI Accession no.	Name of species	NCBI Accession no.
Tor putitora	FJ459537.1	*B. cf. barila*	HM224138.1
T. tor cell line TTH	JX090197.1	*B. gatensis*	HQ219095.1
T. mosal mahanadicus	GQ469781.1	*Labeo caeruleus*	JX983336.1
T. khudree	GQ469787.1	*L. rohita*	JX946428.1
T. tor	FJ459436.1	*L. stolizkae*	JX042167.1
T. cf. barakae	KF410983.1	*L. fimbriatus*	KC757286.1
T. tambroides	KC905024.1	*L. gonius*	JX946409.1
T. sinensis	HM536900.1	*L. rajasthanicus*	HQ179949.2
T. malabaricus	HM585024.1	*L. calbasu*	JX983340.1
T. tor sp.	HQ609732.1	*L. sp.*	Jx260901.1
T. mosal	EU714110.1	*L. dussmieri*	JX946433.1
Cirrhinus mrigala	JX983257.1	*Puntius brevis*	HM536914.1
C. cirrhosus	JX260850.1	*P. sp.*	HM536916.1
Chagunius chagunio	JX066746.1	*P. chola*	JQ713852.1
C. baileyi	JX066747.1	*P. parrah*	HE664124.1
Hypopthalmichthys nobilis	HQ236000.1	*P. sophore*	JQ667571.1
H. molitrix	JX260884.1	*P. fraseri*	JX260949.1
Barilius vagra	KC791630.1		

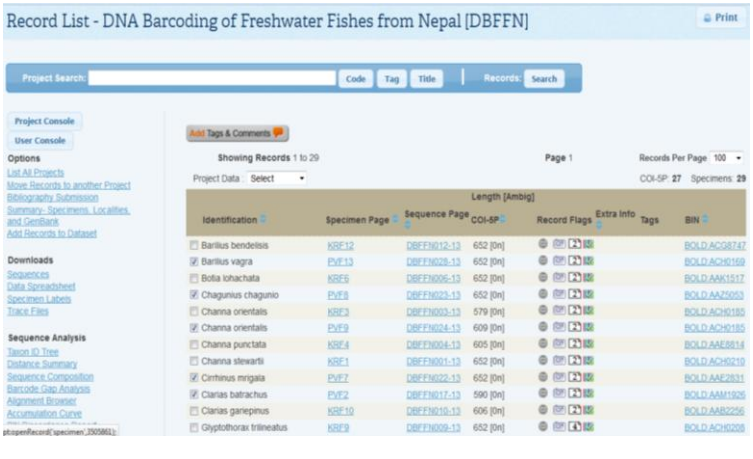
Perciformes:

Name of species	NCBI Accession no.	Name of species	NCBI Accession no.
Channa orientalis	JX983246.1	*O. aureus X O. niloticus*	DQ856613.1
C. aurantimaculata	HM117173.1	*O. niloticus*	HQ654747.1
C. bleheri	HM117186.1	*O. esculentus*	AY662795.1
C. gachua	EU342197.1	*O. urolepsis*	FJ348121.1
Oreochromis mossambicus	HQ219153.1		

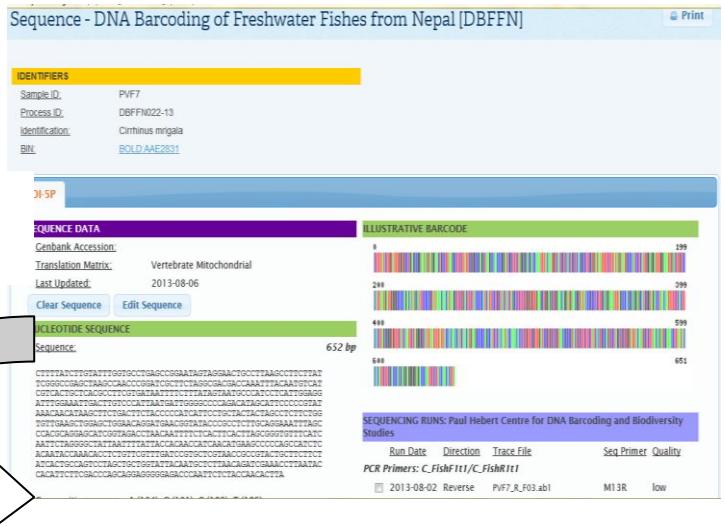
Siluriformes:

Name of species	NCBI Accession no.	Name of species	NCBI Accession no.
Clarias batrachus	JQ699208.1	*C. sp.*	HM345932.1
C. fuscus	KF011505.1	*C. gariepinus*	KC500413.1
C. macrocephalus	JF292329.1	*Mystus cavasius*	JN228946.1
C. gabonensis	HM880231.1	*M. bleekeri*	JX983375.1
C. teijsmanni	JN646093.1	*M. oculatus*	HQ009493.1
C. jaensis	HM882818.1	*M. vitattus*	JX983375.1
C. angolensis	HM880232.1	*M. malabaricus*	HQ219112.1
C. dussumieri	JQ699213.1	*M. tengara*	FJ459425.1

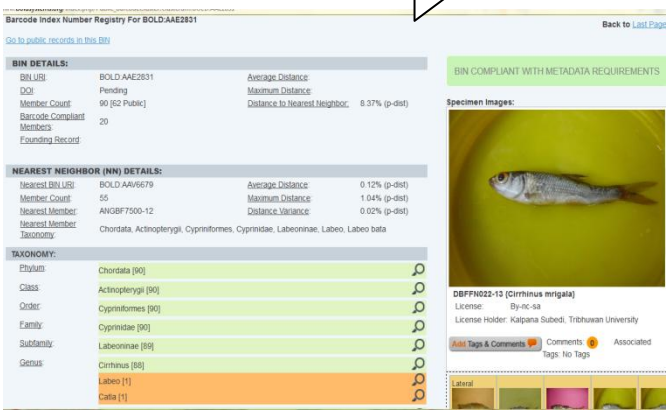
Appendix 8: Quick view of BOLD System:v3



Project page



Sequence Page



Specimen Page

Trace Viewer Page

Appendix 9: Specimen Data Sheet of the studied fishes in BOLD.

	Collection Info							
Sample ID	**Collectors**	**Collection Date**	**Continent/ Ocean**	**Country**	**Zone**	**District**	**Sector**	**Exac**
PVF1	Kalpana Subedi	27/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF2	Kalpana Subedi	27/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF3	Kalpana Subedi	27/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF4	Kalpana Subedi	26/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF5	Kalpana Subedi	27/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF6	Kalpana Subedi	27/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF7	Kalpana Subedi	28/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Small r
PVF8	Kalpana Subedi	28/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF9	Kalpana Subedi	28/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Small r
PVF10	Kalpana Subedi	27/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF11	Kalpana Subedi	26/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF12	Kalpana Subedi	27/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF13	Kalpana Subedi	27/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas
PVF14	Kalpana Subedi	26/6/2013	Asia	Nepal	Gandaki	Kaski	Pokhara	Small r