DNA BARCODING AND PHYLOGENETIC ANALYSIS OF FISHES OF BEGNAS LAKE

M.Sc. Thesis

2016

Submitted to

Central Department of Biotechnology

Tribhuvan University, Kathmandu, Nepal

Submitted by

Pradip Paudel

M.Sc. Biotechnology

Supervisor

Prof. Dr. Tilak R. Shrestha

Central Department of Biotechnology

T.U. Reg no. 5-2-242-11-2007

CERTIFICATE

Dr. Babasaheb Ambedkar Marathwada University, Aurangabad - 431 004

Paul Hebert Centre for DNA Barcoding and Biodiversity Studies

G. D. Khedkar Ph.D.
Director

Tele (off.) : (Direct)+91-240-2403216

Fax : +91-240-2403335.
E. Mail : gdkhedkar@gmail.com
www.phcdbs.org

Ref No.: PHCDBS/ Certificate-Pradip/26/01 **Date:** 30/06/2015

CERTIFICATE

This is to certify, that **Mr. Pradip Paudel**, student of Central Department of Biotechnology, Tribhuvan University, kirtipur, Nepal has pursued his research work entitled "***DNA Barcoding and Phylogenetic Analysis of Fishes of Begnas Lake***" at our centre during **2 to 14 Feb, 2015.**

30/6/2015

Director,
PHCDBS

DIRECTOR
Paul Hebert Centre For DNA
Barcoding And Biodiversity Studies.
Babasaheb Ambekar Marathwada University.
Aurangabad.

RECOMMENDATIONS

There are some measures to be followed in order to get more advantages from barcoding of fishes of Nepal in the future:

1. 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.
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. Museum specimens, which currently lack an authority, can be barcoded so that they can regain taxonomic identification.
4. 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.
5. Fishes sampling should be accompanied with GPS reading which wasn't done at present reading due to the technical difficulties.

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.

Acknowledgement

First of all, I would like to express the deepest appreciation to my supervisor, Dr. Tilak R. Shrestha, Professor of Central Department of Biotechnology, Tribhuvan University, Nepal, who has guided me in every step from inception of idea in my mind to completion of work.

My special acknowledge Dr. Gulab D Khedkar, director of Paul Hebert DNA Barcoding Centre and Biodiversity studies (PHDBS) at Dr. Babashaeb Ambedkar Marathwada University, Aurangabad, India for Providing Laboratory and other technical support toward the completion of this work. I am highly indebted to Dinesh Nalage, Mahesh singre, Sajid khan, Amol Kalyankar, Krishna kholae and all lab members of that institution. I am highly indebted to local fisherman for their help in sample collection and recognizing their local names. They were so excited by briefing about my study that they provided me different fish's species which was very hard to catch in those cold climates.

I would like to forward sincere gratitude to the scientists who have developed the universal fish barcoding technique, viz Natalia V Ivannova and coworkers for COI amplification primers and PCR condition, Mehrdad Hajibabei for optimizing cycle sequencing reaction Joachim Messing for sequencing Primers.

I would like to p resent my sincere thanks to head of department Prof. Dr. Rajani Malla of my home institution Central Department of Biotechnology, TU along with all the faculties and staffs. I present my hearty thanks to my dear friends who encourage me. Finally, I would thank everyone who directly or indirectly helped me for the completion of work whose name might not have been mentioned but I will always be grateful to them.

ABSTRACT

Approximately 650-bp sequences of mitochondrial DNA (mtDNA) have been designated as ‘‘DNA barcodes’’ and have become one of the most contentious and animated issues in the application of genetic information to global biodiversity assessment and species identification. Mitochondrial COI region has become well established marker for animal DNA barcoding. However, we suggest that the utility of barcodes is suspect and vulnerable to technical challenges that are particularly pertinent to mtDNA. The region possess high Taxonomic conservation as well as enough Variability for species discrimination. Although the Barcode Library for all organism present in the Earth Primarily aimed at Facilitating identification, its application in forensic science, identification of mislabeling, ecological studies, food web study, phylogenetic study and population genetics has been proved by ongoing intensive studies.

Nepal is rich in Biodiversity with abundant endemic flora and fauna, this type of research will be a milestone in the genetically diverse country like Nepal. Overexploitation of freshwater ecosystem possess serious threat to fish and fisheries DNA barcoding aid the accurate estimation of Biodiversity and its conservation.

Keywords: DNA Barcoding, COI-5P, K2P model, transition/transversionsubstitution, Phylogenetic

TABLE OF CONTENTS

CHAPTER 3: METHODOLOGY

CHAPTER 4: RESULTS

CHAPTER 5: DISCUSSION

CHAPTER 6: SUMMARY AND CONCLUSION

LIST OF TABLES, FIGURES AND APPENDICES

[A] TABLES **Page no.**

[B] FIGURES

List of Abbreviations and acronyms

A/ T/ G/ C	Adenine/Thymine/Guanine/cytosine
ATP	Adenosine triphosphate
BIN	Barcode Index Number
BOLD	Barcode of life Database
BLAST	Basic local alignment search tool
bp	Base pair
°C	Degree Centigrade
CBOL	Consortium of barcode
COI	Cytochrome c Oxidase subunit 1
COI-5P	Cytochrome c Oxidase subunit 1 gene
Cytb	Cytochrome b gene
DNA	Deoxyribonucleic acid
dNTP	Deoxy ribonucleotide triphosphate
Exo-SAP	Exonuclease shrimp Alkaline Phosphate
iBOL	International Barcode of life
indels	Insertion and deletion
ITS	Internal Transcribe spacer
IUCN	International Union for conservation of Nature
FAO	Food and Agriculture Organization
FISH-BOL	Fish Barcode of life Initiative
FDA	Food and Drug Administration
Kb	Kilo base
mA	Milliampere
matK	Megakaryocyte associated tyrosine kinase gene
MEGA	Molecular Evolutionary Genetics Analysis
Min	Minutes
mM	Milli molar
mtDNA	Mitochondrial Deoxyribonucleic Acid
NCBI	National Centre for Biotechnology Information
NFW	Nuclease Free Water
NJ	Never Joining
ng	Nano gram
NUMTS	Nuclear Mitochondrial DNA
PCR	Polymerase Chain Reaction
pM	Picomolar
psbk-psbl	Intergenic spacer between gene coding for small photosystem II components (PSII-K) and PSII-L
RAG	Recombinant activating Gene
Rbcl	Ribulosebisphospate carboxylase gene
rDNA	ribosomal DNA
rpoB	Gene coding beta subunit of RNA Polymerase
rpoC1	Gene coding Gamma subunit of RNA Polymerase
S	Svedberg Unit
TBE	Tris/Borate/EDTA Buffer
µg/ µl	Microgram/ Microlitre

CHAPTER1: INTRODUCTION

1.1 Background studies

Nepal is one of world largest source of fresh water resources arising from high hills and snowy mountains which give rise to big river and big fresh water lakes in the hilly areas. 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 (T. Petr and Deep Bahadur Swar 2002). 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).These fresh water resources are the best habitat for fresh water fishes where they grow develop and breed in the proper seasons. In Nepal, traditionally capture fishery is being done as a main profession subsistence by community like Majhi and Tharu which is the most delicious and good nutritional value. Beside food value, fish have cultural, religious and medicinal value in traditional medicine in Nepal and worldwide. The use of fishing for recreational purposes as sports fishing and for decoration in aquarium is also increasing Nepal is rich in fresh water fishes Diversity. Shrestha (2001) has reported Nepal for a total of 182 fish species belonging to 92 Genera under 31 families and 11 orders. 76 indigenous cold water fish has been compiled and reported by Rajbanshi (2001) in 'The symposium on cold water fishes of Trans-Himalayan Region ', 10-13 July 2001, Kathmandu. 9 endemic fishes of Nepal have been reported out of which, the following five fishes *Myersglanis blyrhii* (Day, 1952), *Psilorhynchus pseudecheneis* (Menon and Dutta, 1962), *Schizothorax macrophthalmus* (Terashima, 1984), *S. nepalensis* (Terashima, 1984) and *S. raraensis* (Terashima, 1984) inhabit cold Water. Accurate assessment of species diversity remains a major challenge for systematic ichthyology because of drastic morphological shifts encountered across developmental stages and sometimes sexes, and perhaps more subtle shifts across Geographic ranges (Hanner *et al.,* 2011). It took over ten Centuries for taxonomists to Describe 1.7 million species but this figure is still a gross under estimate of the true biological diversity on earth (Blaxter, 2003; Wilson, 2003) . There are following limitations of morphology based taxonomy (Hebert *et al*., 2003) phenotypic plasticity in the character employed for species recognition lead to incorrect identification.

- Morphologically cryptic species are often overlooked.
- There is lack of taxonomic keys to identify immature species of many species and

- Traditional Taxonomy requires high levels of expertise in any given group and is therefore restricted to specialists

As a means to revitalize traditional Taxonomy and help it rise above the taxonomic crisis, alternative and complementary approaches have been put forward, for example; molecular taxonomy (Hebert *et al.*, 2003; Tautz *et al*., 2003). DNA barcoding has been successful in the identification and delimitation of new species from Various groups (Hebert *et al*., 2004; ward *et al*., 2005; Cywinska *et al*., 2006; Hajibabaei *et al.,* 2006; smith *et al.,* 2007; Borisenko *et al.,* 2008; Kerr *et al.,* 2009). These studies have established that DNA barcoding is efficient for identification of Lepidoptera, fishes, mosquitos, mammals and birds. After many studies conducted since 2003 using DNA barcoding criticisms and skepticism towards barcoding have changed. This method has proven fast, reliable and cheap both in the discovery and identification of biodiversity (Radulovici *et al.,* 2010). DNA barcoding is the unique identification system based on the sequence analysis of short stretch of DNA. Mitochondrial gene cytochrome oxidase I (COI) can be used as global identification system for animals (Hebert *et al.,* 2003). This proposal is based on the two known important advantages of COI. First, the availability of very robust Universal primers for the recovery of 5' end of COI from representatives of most, if not all animal phyla (Folmer *et al.,* 1994; Zang and Hewit *et al.,* 1997). Second, COI appears to possess a greater range of phylogenetic signal than any other mitochondrial gene. In common with other Protein coding gene, its third position nucleotide show high incidence of base substitutions, leading to a rate of molecular evolution that is about three times greater than of 12S or 16S rDNA (Knowlton and Weigt, 1998). In fact, the evolution of this gene is rapid enough to allow the discrimination of not only closely allied species, but also phylogeographic groups within a single species (Cox and Hebert, 2001; Wares and Cunningham, 2001) With; recent researches, some other mitochondrial genes or nuclear ribosomal DNA Fragments have been proposed as alternatives for species identifications. There is report describing a set of 21 PCR primers and amplification conditions developed to barcode any teleost fish's species according to their mitochondrial cytochrome b and nuclear rhodopsin gene sequence (Sevilla *et al.,* 2007). COI is usually more conserved than cytb. So, COI barcode region is a more suitable species marker across wide range of taxa and cytb for intra species identification. Where ever there is low resolution from the COI gene alone, the combination of other molecular marker such as cytb, 16S and 18S rDNA can help to solve this problem, (Zang and Hanner *et al.,* 2013). Cytb

and COI are equally well suited for the species identification of fishes, 16S has drawbacks in the discriminating closely related species (Kochzius *et al*., 2010).

1.1.1 Pokhara Valley

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 Begnas Lake

Begnas Lake is the second largest lake of Pokhara Valley. Tucked away at an altitude of 650 meters, this lake is located in the Siswa village on the eastern part of Pokhara and is 13 km away from the city of Pokhara. Spread across a total area of 3 square km, the lake has a capacity of 29.05 million cubic meters. The northern and western parts of the Begnas Lake are relatively deeper than the eastern and southern parts. Begnas Lake 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). 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). 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 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. Begnas Lake was stocked with exotic carp species- *Hypophthalmichthys molitrix*, *Aristichthys nobilis, Ctenopharyngodon idellus, Cyprinus carpio* and *Labeo rohita* (Swar DB. *et al.,* 1988).

Figure 1.1: Sample site map. Map of Kaski district showing Pokhara valley with Begnas Lake (Source: http://www.thekingdomofnepal.com/)

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 Fish Base, 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

The first scientific record of Nepalese fish is done by Hamilton in "Fishes of Ganges" in 1822. Later a number of ichthyologists, including McClelland (1839), Gunther (1861), Beaven (1877) and Day (1889) studied on the fishes of the Asian continent which were deposited in Indian Museum, British Museum, Natural History Museum Standford and Field Museum of Natural History Chicago, USA. Boulenger *et al.,* (1907) studied some of the fishes from Nepal. Hora (1921-1952) had made admirable researche on the fishes from Nepal (Koshi, Bagmati, Gandaki, Rapti and Karnali). Dr. Ripley's expedition team collected fishes from Koshi and Karnali and deposited the in Indian Museum, Calcutta. Leviton, Mayers and Swan (1955) listed many fishes during the course of California Himalayan Expedition. Taft (1955) prepared checklist of the fishes containing 95 species representing 13 families. De witt elaborated the checklist of Taft including 102 species representing 21 families. In the 20th Century, Regan (1907) studied seven fish species send to him by Dr. Annendalei, India, out of which five species were reported from Nepal collected from Kathmandu and adjacent areas like Sundarijal and Pharping. Out of the reported fish species, one species *Diptichus annandalei* sp. nov. Was found then the new to Science (Synonymous to *Schizothorax*

richardsonii). Hora (1937) also studied 158 fish speies from Hulchok Mugling, Nagarkot and Sundariajal collected on his request by the Resident, British Legation in Nepal. Menon (1949) listed 52 fish species from the river Koshi, Eastern Nepal. Thapa Rajbanshi (1968) presented a paper in Hill stream fishes of Nepal. Majpuria and Shrestha (1968) published a paper on fresh water fishes of Nepal. Bhatt and Shrestha (1973) studied fish and wildlife of Suklaphata. Shrestha (1975) studied structure and seasonal changes in gonads of *Schizothorax plagiostomus* and *Garra gotyla*. Shrestha J. (1978) studied the fish fauna of Nepal and reported 118 fish species in which she described two new species and one sub species (*Barillus jalkapoorie* sp. nov, *Lepidocephalichthys nepalensis* sp. nov., and *Pseudeutropius murius batarensis* sub sp. nov.,). There has been report of 82 fish species from downstream of the river Bagmati (Shrestha et al., 1979). Ferro and Badagami (1980) reported 22 fish species from Begnas lake and Rupa in Pokhara Valley whereas during same time period 62 fish species were reported from Gandak river system of Chitwan Valley (McGladdery *et al.,* (1980) . Jayaram (1989) reported 106 fish species under 61 Genera, 21 families and 8 other. The first complication of the reported fish fauna for the Central or Nepal Himalaya within the boundary of the kingdom of Nepal for the period 1783-1982 was prepared which include 171 fish species of which 164 were indigenous and 7 exotic (Rajbanshi,1982). Terashima (1984) reported three new species of *Schizothorax* endemic to the Mahendra Lake (Rara Lake). Edds (1985) has further reported a list of 111 and 113 native fish species from the river kali Gandaki /Narayani River and the waters of the Royal Chitwan National park, Chitwan respectively. Jha and Shrestha (1989) has studied fish and Fauna of the river karnali and have reported 57 fish species under 38 genera, 19 families and 9 orders from the rivers Rapti and the river Narayani. Shrestha (1990) has recorded 108 fish species from the river koshi 102 fish species from Gandak 74 species from the karnali 82 species from Bagmati (downstream near Karmaiya) 34 fish species from the Trishuli and 69 species from the Mahakali, 96 fish species representing 19 Families and 5 orders from Nepal have been reported in book "Inland fishes of India and adjacent countries"(Talwar and Jhingran, 1991). Shrestha, J. (1994) has reported a total of 188 freshwater fish's species, out of which 173 indegenous and 9 are exotic. Shrestha, T.K (1995) recorded a total of 183 species, out of which 173 are indigenous and 10 exotic fish species. Out of the reported exotic fishes, two species – *Oncorhynchus rhodurus* (Jordan *et al.,) Mcgrego*r and *Salmo trutta L.* do not exist presently in the country. Subba (1995) has reported a new record on

the occurance o the hill stream fish *Olyra logicaudata* from a tributary of the River Trijuga, at tributary of the Koshi River, Saptari district, Eastern Nepal. Environment Impact Assessment (EIA) studies carried during different hydroelectricity projects have reported fish species of the water system under study.

1.1.5 Taxonomy

Taxonomy is the science of defining groups of biological organisms on the basis of shared characteristics and giving names to those groups. It is defined as the science related to discovery, recognition, definition, and naming of groups of organisms. 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 D.A. *et al.,* 2002). 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 basic taxonomy scheme divides living organism into Domain, Kingdom, Phylum or Division, Classes, Order, Family, Genera and Species. Sometimes Subphylum /Subdivision, Subclasses, Suborder and Sub-families are also used. 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, and Tournefort who contributed in plant classification system. Among all, the Swedish botanist Carlus Linnaeus (1707–1778) is regarded as the father of taxonomy, as he developed a system known as Linnaean classification for categorization of organisms and binomial nomenclature for naming organisms in his book "The System of Nature (Systema Naturae)". Linnaeus used five ranks: class, order, genus, species, and variety. It provides opinions on species boundaries, and on the phylogenetic relationship between species. Well before Linnaeus, plants and animals were considered separate Kingdoms. Linnaeus used this as the top rank, dividing the physical world into the plant, animal and mineral kingdoms. As advances in microscopy made classification of microorganisms possible, the number of kingdoms increased, five and six-kingdom systems being the most common. Four kingdom (Monera, Protista, Plantae and Animalia) of classification was put forwarded by Copeland; 1938) similarly Five kingdom (Monera, Protista, Plantae, fungi and Animalia) of classification was proposed by Whittaker (1969) published by him is remarkable for an overall framework of classification. Erasmus

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.

1.1.7 Morphological Taxonomy

Morphological taxonomy is the primary tool used by taxonomists to classify the organism by observing its phenotypic characteristic to categorize it into Hierarchical system of nomenclature. It is the most primitive methodology which relies on morphological characteristics like shapes, sizes, pigmentation patterns, disposition of fins, and other external features that aid in recognition, identification, and classification. The whole concept of morphological taxonomy is contribution of ‘The father of Taxonomy’, Carlous 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 which is being practiced from hundreds of years. Till now most of the organisms in the earth have been explained on the basis of morphological taxonomy. This method has less advantages than more disadvantages. It is easy fast but often criticized for its reliability and specificity and highly trained persons are needed which isn't feasible all time. The construction of morphological keys or trait is difficult and laborious. 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. These obstacles of morphological taxonomy can be subdued by using DNA barcoding to classify living organisms (Hebert *et al*., 2003). 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

Fish are the member of a paraphyletic group that have gill bearing aquatic craniate that lack limbs with digits including hagfish, lampreys, and cartilaginous and bony fishes well as various extinct related groups (Nelson, 2006). They can be found in nearly all aquatic environments from high mountain streams to the abyssal and even hadal depths of the deepest oceans. It exists greater diversity than other vertebrates. All the traditional taxonomist and phylogenetic study among fishes was based upon morphological characters. Ichthyologist used several approaches to study distinction and relationship among fishes, including "traditional "morphological methods, biochemical method, chromosomal studies and, most recently, molecular approaches. In morphological studies, meristic data (count that relate to body segments such as scales and fin rays) and other counts (e.g. gills raker, sensory pores and other features) and morphological (body measurements) data are taken into consideration. Classification of fishes have taken evolve for over three centuries, beginning even before the concept published by Linnaeus (1758) in systema nature, hypothesized to be closely related are grouped within genera, related genera within families, families into orders, and order into classes. His taxonomic approach became the systematic approach to study the organism including the fish. Peter Artedi was a Swedish naturalist and is known as the father of Ichthology who recognized five additional orders of fish *Malacopterygii, Acanthoptergii, Branchiostegi, Chondropterygii* and *Plagiuri*. He developed modern method of anatomical feature that are modernly exploited. In nineteenth century Marcus Elieser Bloch of Berlin ang Georges Cuvier of Paris made attempt to consolidate the knowledge of ichthyology. Cuvier summarized all the information in his monumental "Historie naturelle des Poissons" which was published between 1828 and 1849 in a 22-volume series that describes 4,514 species of fishes 2,311 of these were new. Albert Gunther published "Catalogue of the fishes of the British Museum" between 1859 and 1870 describing over 6,800 species and mention 1,700 more. Dr. Francis Buchanan Hamilton (1762-1829) a Scottish physician who significantly contributed to botanist geographic and

majorly to ichthyologist by listing many Indian and Indian sub-continent fishes while living in India. His work was appreciated on Indian fish entitled "An account of the fishes found in the river Ganges and its branches (1822)" which describe over 100 species not formally recognized scientifically. 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*.)

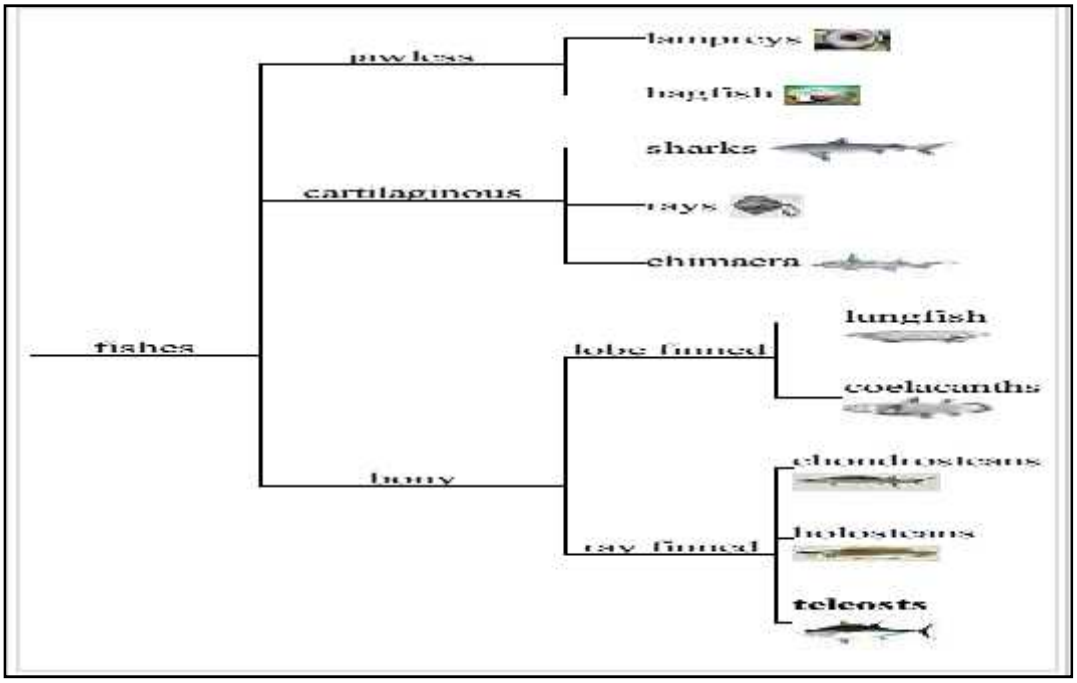


Figure 1.2: Basic taxonomy of fishes (*en.wikipedia.org/wiki/Template Basic fish taxonomy*)

1.2 Current Studies

Nepal has at least 186 fish species, only few species has been barcoded. Nevertheless, some medicinal plants have been barcoded and the data submitted to the BOLD. However, barcoding of many animals including birds, springtails invertebrates, skipper, 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 Begnas Lake of Pokhara being carried out in Paul Hebert Centre of DNA barcoding and Biodiversity Studies(PHCDBS), Aurangabad.

1.3 Hypothesis

DNA barcoding serve as the core of global bio-molecular identification of animals, plants, insects and others based upon mitochondrial Cytochrome c Oxidase subunit I (COI) approximately 652 bp which is used as the comprehensive barcode region. COI-5P is highly conserved up to species level so possess high phylogenetic relation within same species. 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. 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 inter-generic or inter families' species K2P value is greater than 2%. Thus, we hypothesize that integration of DNA barcodes (partial cytochrome c oxidase subunit I sequences) into bio-assessment 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 obtain Fish COI sequence catalogue from Begnas Lake by 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 amplified product.
- Submission of COI Barcode sequences to BOLD and to NCBI for validity and accession number.
- 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 also vital part of ecology and biodiversity. Fishes are considered as good foods with respect to human health being rich source of proteins, vitamins and minerals; including social religious and cultural value so are consumed largely in many forms. They are at the point of extinction directly or indirectly by human interference. 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. Morphology based 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 Scope of the study

This type of research carries tremendous scopes in global bio-identification of Nepali flora and fauna and its biodiversity from low level land Terai to highest altitude of the world Sagarmatha. Nepal being rich is Natural biodiversity this small work help to study genetic variation of fishes not within country but also outside countries. 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. 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. It also helps to prevent illegal poaching of Nepali endemic fishes outside the country.

CHAPTER 2: LITERATURE REVIEW

2.1 Molecular Taxonomy or DNA Taxonomy

2.1.1 Mitochondrial DNA

Mitochondria are cellular structures that convert the energy from food through the oxidative phosphorylation and regulate self-destruction of cell and production of cholesterol and heme (component of hemoglobin). 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. Molecular tolls had been widely used for species separation and identification throughout the past two decades (Scherer and Sontag 1986). Among mitochondrial genes, the only 2 protein coding genes that occur in all eukaryotes are cytochrome c oxidase subunit I and cytochrome b. 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). Mitochondrial DNA barcoding is a novel system designed to provide rapid, accurate, and automatable species identifications by using short, standardized gene regions as internal species tags. As a consequence, it will make the Linnaean taxonomic system more accessible, with benefits to ecologists, conservationists, and the diversity of agencies charged with the control of pests, invasive species, and food safety. Protein coding genes and regulatory genes (control region) of mtDNA are used as markers for investigating intra species and inter species 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). 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). mtDNA a relatively fast mutation rate, resulting in the generation of diversity within and between populations over relatively short evolutionary timescales (thousands of generations) 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 phylogenic, 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 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 P.D.N. *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 J.W.O. *et al.,* 2004). Furthermore, the peculiar nature of the mitochondrial organelle arises 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 L.C. *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 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 A.R. *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 pseudo-genes 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 D.C. *et al.*, 2011).

2.1.2 COI as DNA Barcode Region

2.1.2.1 Cytochrome *c* Oxidase

Cytochrome oxidases is a 13 protein subunits intrinsic membrane metallo-protein complexes that contain 2 heme iron, 2 copper centers, 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 (Complexes I, II, III, and IV) activates di-oxygen, 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 ferro-

cytochrome c and the four-electron/ four-proton reduction of di-oxygen to water (Brunori M. *et al.,* 1987).

Cytochrome *c* oxidase receives electrons from cytochrome c, a small, soluble, mobile protein that moves from complex III to complex IV. The electrons pass through a number of metal centers and the final reduction of oxygen to water occurs at the a3/Cu B binuclear center. CCO donates electrons from cytochrome c to reduce oxygen and generate water. The enzyme pumps two protons and generates one water molecule per two cytochrome c molecules.

For cytochrome c oxidase, the overall reaction is:

4Fe+2-cytochrome c + 4H+in + O2 --> 4Fe+3-cytochrome c + 4H+out + 2 $H2O._P$

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 P.C. *et al.,* 2003).

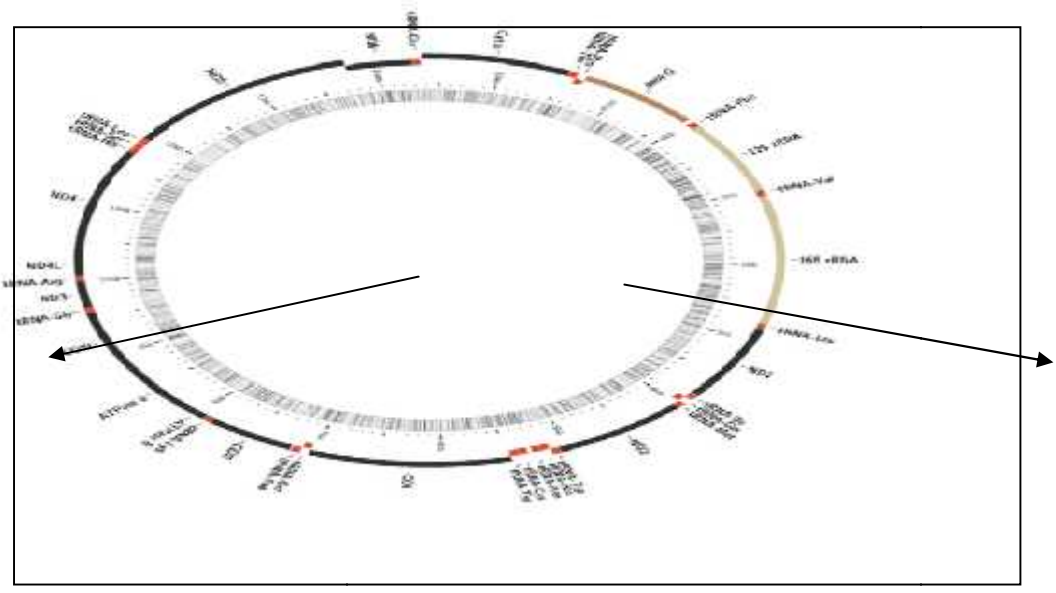


Figure 2.1: Mitochondrial DNA showing location of genes and other key regions. Below shows COI of *Cirrhinus mrigala* GenBank (Accession no. JQ838173.1). The 1550 bp COI sequence (5486-7036) extracted is shown above along with the indication of primer binding sites. In degenerate primer Y refer to either C or T and R refer to A or G. Primers sites are highlighted yellow and the M13 tail in each primer is underlined and highlighted in blue. (Diagram source: Mitochondrial Genome Database of Fish)

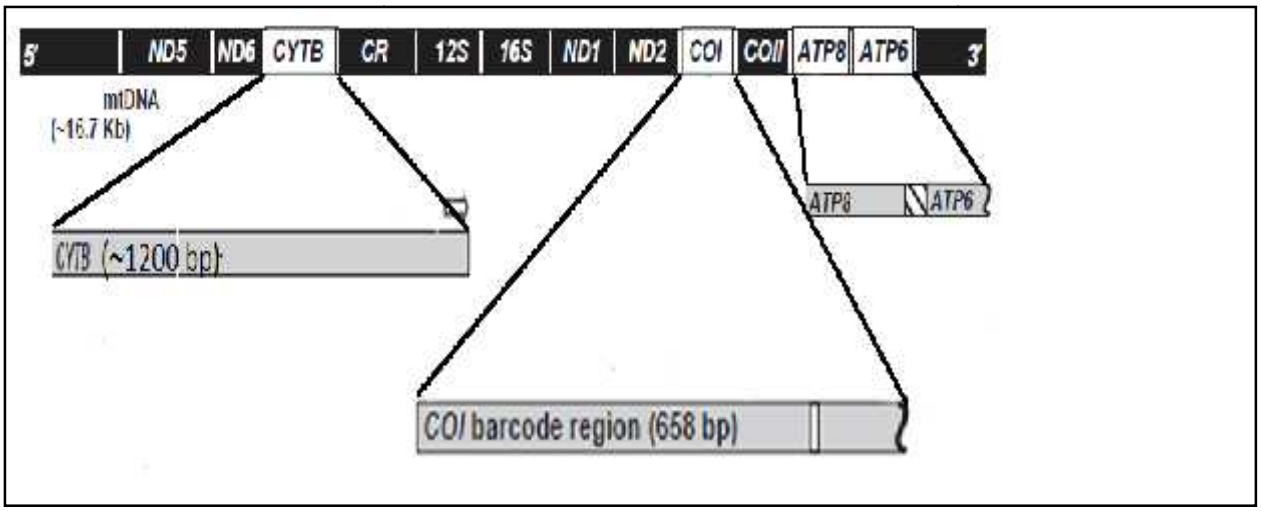


Figure 2.2: Schematic view of a linearized mitochondrial DNA showing the relative positions of most coding and noncoding regions (Chaves P.B. *et al.,* 2012).

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 D.C. *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 P.D.N. *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 L.C. *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 phylo-geographic 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 P.D.N. *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 region for fungi (Seifert *et al*. 2007; Seifert 2009) and supplementary locus (CBOL Plant Working Group, 2009) because incomplete concerted evolution, fungal contamination and difficulties of amplification and sequencing (Hollingsworth *et al.,* 2011). It shows powerful marker for interspecific divergence (A´ lvarez &Wendel, 2003). 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 C.L. *et al.,* 2012). ITS spacers from nuclear ribosomal DNA (nrITS) also represent the fundamental barcode in some parasitic plants with highly reduced genome (Hollingsworth P.M. *et al.,* 2011).

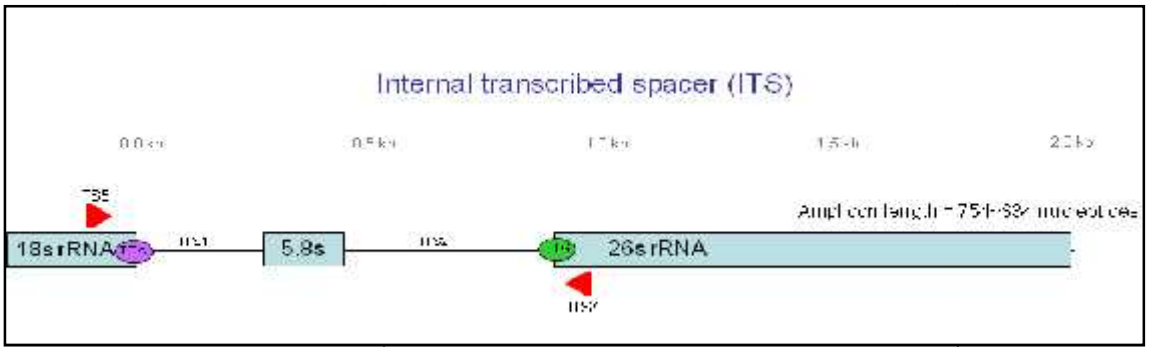


Figure 2.3: Showing internal transcribed spacer (ITS)

b) 18S Region:

The 18S nuclear ribosomal subunit rRNA gene (SSU) is used in phylo-genetics but has less hyper variable domains (Schoch C.L. *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 P.D.N. *et al.,* 2002).

d) rbcL+matK:

The combination of rbcL (ribulose 1, 5-biphosphate carboxylase/oxygenase) and matK forms a perfect plant barcode region of plastid genome which are portions of two plastid coding regions (Hollingsworth P.M. *et al.,* 2011). 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* has a high evolutionary rate, suitable length and obvious interspecific divergence as well as a low transition transversion rate (Min & Hickey, 2007; Selvaraj, Sarma & Sathishkumar, 2008).90% success in Angiosperm ,83% in gymnosperm and 10% in cryptogams PCR with

multiple primer shows unsatisfied single Universal region for plant barcode. 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 in land plants. But, it has comparatively low evolution rate and lowest divergence in flowering Plants (Kress *et al.*, 2005). The combination of *rbcL+matK* cannot avoid the low PCR efficiency of *matK* and secondly, the success of *rbcL+matK* in discriminating plants is typically lower than that of CO1 in animals and the amplification of one gene effect other gene and create problem in analytical Part.

2.1.3 Numts (Nuclear mitochondrial pseudogenes)

Nuclear mitochondrial pseudogenes (Numts) have been found in the genome of many eukaryote DNA sequences homologous to mitochondrial DNA (mtDNA) these pseudogenes should be very useful in the study of ancient mtDNA and nuclear genome evolution. The existence of mitochondrial DNA-like sequences in the nuclear genome of eukaryote cells was first suggested in 1967 by the results of hybridization experiments between mouse mtDNA and nuclear DNA. Lopez *et al.* discovered a recent insertion of mtDNA in the nuclear genome of the domestic cat and named it “Numt” (for nuclear mtDNA segment). A “pseudo gene,” as defined in the Human Genome issue of *Science*, is a nonfunctional copy of a normal gene that has been slightly altered so that it is no longer expressed. 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 pseudo genes, homologs or para logs. They vary widely among eukaryotes, with human and plant genomes harboring the largest repertoires (Antunes A and Ramos MJ, 2005). 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 N.V., 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 pseudo gene (Vallinoto M. *et al.,* 2000). PCR ghost bands, extra bands in restriction profiles, sequence ambiguities, frame shift mutations, stop codons and unexpected phylogenetic placements are indications for mitochondrial pseudo genes. Sequence ambiguities result if the pseudo genes 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 pseudo genes 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).

2.1.5 rpoB gene

The chloroplast gene (rpoB), a series of transcription and translational fusion of the gene for the beta subunit of RNA polymerase is used in DNA barcoding in land Plants (Hollingsworth P.M. *et al.,*)

2.1.6 *trnH-psbA*

The *trnH-psbA* spacer, although short (~450-bp), is the most variable plastid region in angiosperms and is easily amplified across a broad range of land plants is the best plastid option for a DNA barcode sequence that differentiate all flowering plant species have greater potential for species-level discrimination than any other locus and needs further investigation. (Kress W.J. *et al.,* 2007)

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 is a major element in barcoding. Collected specimen should be identified by taxonomist and review many journals and articles with labeled specimen identity like Genus species, site of collection, photographs and GPS navigation. After the collection, specimens should be preserved in cyanide or ethanol or freeze. 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 treasure troves of identified species. 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. 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 and Standard operating protocol (SOP) 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. 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 base callers, automatic alignment, coting 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 and maintained 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), 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 Nucleotide Sequence Database Collaborative** – It is the collaborative organization between Gen Bank 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. Specimens are identified by 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 e*t. al.,* 2005).

The Figure 2.4 in the next page illustrates each and every step 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 and then tissues are stored followed by lysis of cells and DNA extraction. 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 cleanup 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 Gene Bank. 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.

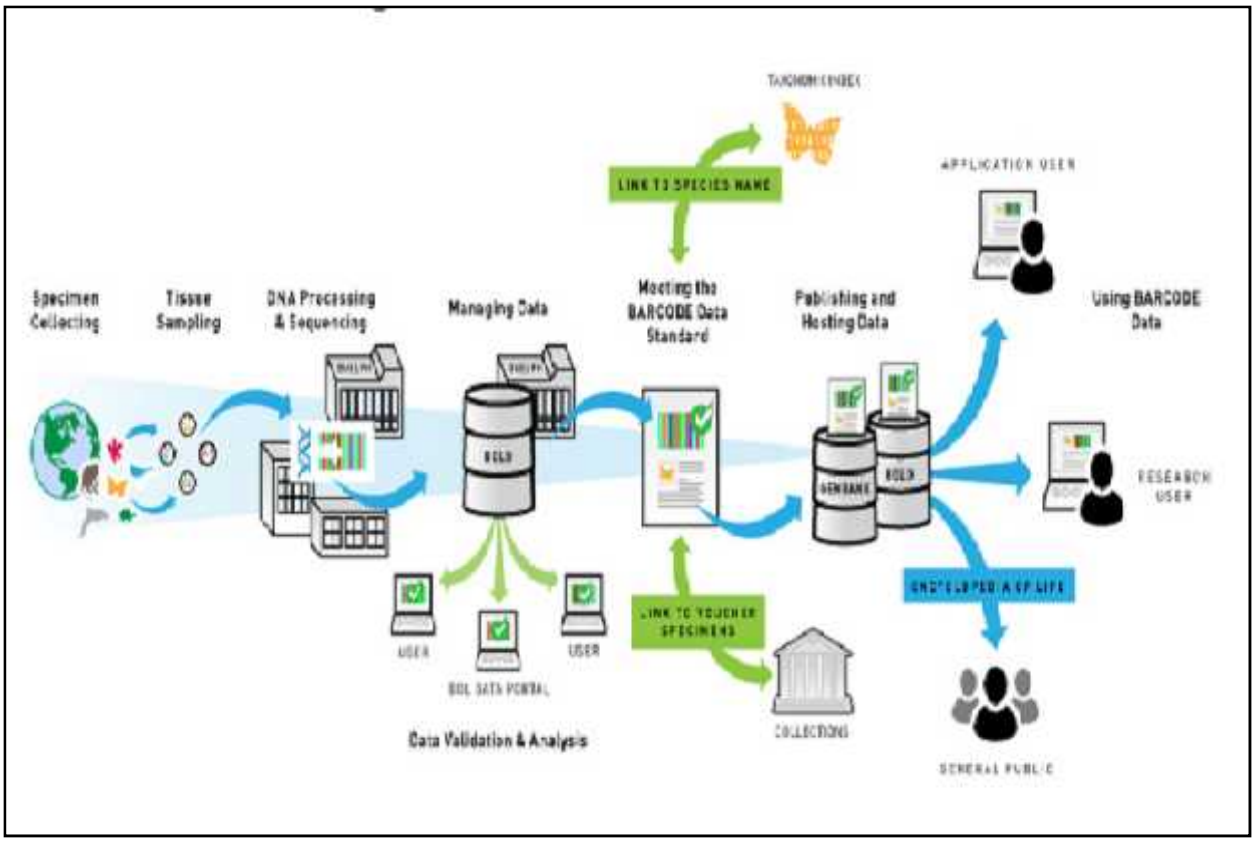


Figure 2.4: DNA barcoding workflow (Hebert P.D.N. et al., 2002), source: BOLD

2.2.2 International Barcoding Efforts

The Consortium for the Barcode of Life, CBOL (http://www.barcoding.si.edu/) started in May 2004 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 and at present includes more than 120 organizations from 45 countries. The Canadian center for DNA barcoding (CCDB) is the world's first and largest high throughput DNA facility established in 2006 within the Biodiversity Institute of Ontario (BIO) at the University of Guelph. 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 an 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 center and the multiple taxonomic communities to establish and strengthen data standards (Ratnasingham S. *et al.,* 2007).

2.2.3 Fish BOL

The Fish barcode of life (FISH-BOL) launched in June 2005 (*Ward et al.,* 2009) initiative seeks to establish a reference sequence library of short, standardized mitochondrial gene sequences derived from the 50 end of the cytochrome c oxidase subunit I gene (DNA barcodes) to facilitate the rapid, accurate, and cost-effective DNA-based identification of all fishes, regardless of life-stage, sex, or specimen condition.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 E.R. *et al.,* 2008). The FISH-BOL campaign has adopted Fish Base (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 Fish Base 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 to aid taxonomical activities and to facilitate collaboration. 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. 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 2015 February, this campaign had barcoded about 107099 fish species with 2300 unnamed barcode recognized for the cytochrome *c* oxidase subunit I (COI) gene (Kress W.J. *et al.,* 2012). 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 E.R. *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 E.R. *et al.,* 2008).

2.2.4 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 E.R. *et al.,* 2008). Fish Base 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. Fish Base provides also a range of country, regional, and Ecosystem specific information (http://www.worldfishcenter.org/fishbase).

2.2.5 BOLD (Barcode of life Systems)

The BOLD (Barcode of Life Data Systems) data system is a DNA public Portal Database provides an integrated bioinformatics platform that support analytical pathway from specimen collection to tightly validated barcode library which is created and maintained by University of University of Guelph in Ontario, Canada. 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 collection, management, distribution, species identification tools and analytical tools to support their validation (Hanner R. *et al.,* 2011). First, it is a repository for the specimen and sequence records that form the basic data unit of all barcode studies. Second, it is a workbench that aids the management, quality assurance and analysis of barcode data. Third, it provides a vehicle for collaboration across geographically dispersed research communities by coupling flexible security and data entry features with web based delivery.

BOLD consist 150 genetic markers including: COI-5P, ITS, rbcl+ matK. 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 (catalogue number and institution storing)
c. Collection record (collector, collection date, location with GPS coordinates)
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 7 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 5 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 Gene Bank (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

2.2.6 Barcode Index Number (BIN)

The Barcode Index Number System (BIN) is an online framework that clusters barcode sequences algorithmically, generating a web page for each cluster (www.boldsystems. Org/BOLD handbook2013). This system consists of three parts. Operational taxonomic unit

(OTU) is generated for the barcode sequence on BOLD using Refined Single Linkage (RESL) 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 be 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 Barcode and Molecular Phylogenetics

Molecular Phylogenetics is the branch of phylogeny that analyses the hereditary molecular evolutionary relationships among groups of organisms (e.g. species, populations), which are discovered through molecular sequencing data (DNA, RNA and proteins) matrices which is expressed in phylogenetic tree. 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 H.A. *et al.,* 2003). mtDNA barcoding provide phylogenetic and phylogeographic analysis thus helping the characterization diversity both within and among local species assemblage, in case of morphologically cryptic species (Emerson *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. 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). Rearchers 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 has been performed earlier, e.g. genes encoding: heat shock proteins, phytochrome, actin, transcription factors encoding gene MADS box genes (in plants) (Soltis D.E. *et al.,* 2003). 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).COI is not an adequate tool to build a fish phylogeny ,insights on relationship will

grow as taxon coverage expands(Valdez-Moreno *et al.,*2009).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 D.E. *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, mtDNA have been used to find the population genetics relationships (Mu X *et al.,* 2012). Mitochondrial DNA markers are haploid and uni-parentally inherited, so they are frequent targets for analysis and have made a particularly strong contribution to population-level studies (Avise J.C. 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. 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-neighbor 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 coalescent 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 facilitates biodiversity study surveys when large number of specimens from diverse taxa are studied and enables the identification of unknown new species and its lineage. Besides, even non-specialists are able to use identifying tools fast, cheaply and

reliably with more practical and fundamental applications (Radulovici A.E. *et al.,* 2010). It can replace the requirement of expert taxonomist and chances of misidentification and reduce cost of producing and sustaining taxonomist in all organismal groups. 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 K.K. *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 H.A.I. *et al.,* 2012; Stoeckle M. *et al.,* 2004).

2.5.2 Scope of DNA Barcoding

DNA barcoding has quite expanded utilities in various fields. Barcoding facilitates numerous applications including detection of putative cryptic species, identification of ambiguous life history stages, estimates shifts in species ranges, issues relating to tracking valuable/endangered species, analysis of food webs and trophic dynamics. Barcoding is flourishing as a useful tool in diagnosing cryptic species which had previously been misidentified as single morphologically based species (Dasmahapatra K.K. *et al.,* 2006). Barcoding can be used to explore life cycles of any organism, facilitate basic biodiversity inventories and its lineages and phylogenetics. 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 H.A.I. *et al.,* 2012). 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 F.O. *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 Woucheria) 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 K.K. *et al.,* 2006). Pseudo genes, 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 number 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 A.E. *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. In Institutional level, Central Department of Biotechonology, Tribhuvan University is the only institute that has barcoded some fishes from Begnas Lake and Koshi River under the supervision of Prof. Dr Tilak Ram Shrestha and all the credit goes to him. From governmental organization, National Academy of Science and Technology (NAST) located at Khumaltar, which has been working in the related field in various plant species have been identified using barcoding method in NAST. 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. From Private sector Centre for Molecular Dynamics-Nepal (CMDN) situated at Kathmandu has also begun working in the field of DNA barcoding. There are uncountable valuable diversities of plants and animals which need to be correctly distinguished. But due to several technical, political and economic problems, establishment of well-facilitated DNA barcoding center 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 Begnas Lake of Pokhara valley. So, sampling of fishes has been done from various water bodies of Begnas Lake and small rivulets. Samples were collected during the month of November 1 to November 7, 2014. 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. Thus fishes were collected randomly as per availability of fishes by fisherman in the Begnas lake.

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: A

Figure 3.1: B

Figure: 3.1: Fishes used for barcoding**.** Figure 3.1 A and 3.1 B represent *Catla catla* lateral view and *Clarias batrachus* dorsal view respectively.

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 store 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 and "Ichthyology of Nepal" by Dr. Tej Kumar Shrestha. Also, various reliable electronic databases like Wikipedia, Fishbase, etc. were also used for identification purpose.

3.6 DNA EXTRACTION USING PROMEGA KIT

The alcohol dipped samples were air dried on a tissue paper for 2-3 minutes and transferred into the microfuge tubes each. 60µl 0.5 M EDTA and 250µl Nuclei Lysis Buffer was added to each tube and crushed with sterilized scissors and forceps. After crushing, 3 µl of Proteinase K was added to each tube and 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, 100µl Protein precipitation solution was added than after vortexed vigorously for 20 seconds followed by centrifugation at 16,000 rpm for 10 min. Then the supernatants were taken carefully in a fresh microfuge tube discarding the pellet and after adding chilled 100µl isopropanol clouded solution was formed indicate the presence of DNA in the

Supernatant. The samples were then centrifuged at 16,000 rpm for 10 min. The supernatants was discarded and pellet was taken. 300 µl of 70% chilled ethanol was added and centrifuged at 16,000 rpm for 10 min for washing. Again, the supernatants were decanted and 300 µl of absolute alcohol was added to each tube which was followed by centrifugation at 16,000 rpm for 10 min. The supernatants were decanted and the pellets were dried at RT for overnight. Lastly, the dried pellets were dissolved in 25 µl Nuclease Free Water (NFW).

3.7 Quantification of DNA

The quantification of DNA was done on Nano drop ND1000 Spectrophotometer by using ND 1000 V3.7.1 software. Firstly, initialization of spectrophotometer was done by placing 1.5 µl of NFW water for washing. Then, after swabbing it with tissue paper, 1.5 µl NFW was kept on same pore to set blank. It was wiped again and 1.5 µl of sample DNA dissolved in NFW was kept. Absorbance was taken at 260 nm and 280nm for calculating the DNA concentration. DNA Concentration (ng/µl) = OD_{260} * 50. The ratio value of OD at 260 nm and 280 nm was used to find out the purity of DNA and contaminating factor. For Good quality DNA, the ratio (260nm/280nm) should lie ~1.8 (accepted as pure DNA). DNA contaminated with RNA shows

260nm/280nm ratio value >1.8 and protein, phenol or other contaminants shows ratio value <1.8.The ratio of OD at 260 and 230 nm was used to judge/check the contamination of Phenolic compounds. DNA which shows ratio value around 1.8 were taken and others deviating from this value were discarded. 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 was prepared in 1XTBE buffer to check the DNA samples by electrophoresis. For this, 1X TBE buffer stained with Ethidium bromide (500µg/µl) 3µl per 50 ml of gel was used. 5 µl of the DNA sample was mixed with 2µl of gel loading dye and was loaded into the wells in gel. The samples were run for 15 min with constant current of 100V and were then visualized under UV trans-illuminator system. Gel images were taken using 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 COI region

Mitochondrial region Cytochrome Oxidase subunit-1 gene (COI gene) was amplified using the following M13 tailed (underlined portion in Primer) as universal cocktail primer for amplification of COI gene for fishes. The sets of primers used are mentioned below.

Fish Cocktail Primer

VF2_t1: FishF2_t1: FishR2_t1:FR1d_t1

Forward Primer

FishF2_t1: TCGACTAATCATAAAGATATCGGCAC

VF2_t1: TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC

Reverse Primer

FishR2_t1: ACTTCAGGGTGACCGAAGAATCAGAA

FR1d_t1: CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAYCARAA

The reaction mixture for the Polymerase Chain Reaction (PCR) was composed as shown in the table. Following program was set up in the Thermal Cycler (ABI Verity, USA):

Table: 3.1: PCR reagents composition and reaction volume (ABI Verity, USA)

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

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	40 sec	35 cycle
	Annealing	51˚C	40 sec	
	Extension	72˚C	1 min30 sec	
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 component together.

3.10 PCR Amplification Check up

1% agarose gel was prepared in 50 ml of 1X TAE 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 100 mA current along with a size standard or marker of 100bp ladder, 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.

Make a Mastermix of Exonuclease I and Shrimp Alkaline Phosphatase for 10 µl of PCR product as per the table below.

Table 3.3: Exo-SAP reagent and volume

Components	Units per Rn	VOL(µl) Per Run	VOL(µl) for 100 Runs
Exo I (20U/ µl)	0.5	0.025	2.5
SAP (1U/ µl)	0.5	0.5	50
PCR Buffer 10X	1X	0.1	10
MilliQ		0.375	37.5
Total Volume		1	100

Add 1 µl of the Mastermix to 10 µl of PCR product and set up the following incubation protocol in a thermal cycler. Then the mixtures were incubated at 2 different conditions 37° C for 120 min and 85° 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 3.4 below (Hajibabei *et.al*, 2005).

Table 3.4: 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 3.5 below.

Table 3.5: Cycle sequencing PCR Condition

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 3.6: 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	Total:	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 DNA Sequencing

In each tube, 15 µl of Hi-Dye 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 capillaries of 50 cm length was used. The readied samples were loaded into the wells finally for obtaining sequences bi-directionally.

3.15 DNA Sequence Alignment

The sequence trace files were assembled using MEGA 6.06 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(Barcode of life Data System) and NCBI(National Center for Biotechnology Information) along with all the requirements needed as per standard format provided by them such as specimen data sheet including species name, voucher data, collection record, identifier name, and PCR primers used to generate the amplicon and trace files. All the COI sequence are required to submit to BOLD for acquiring validation of the fish Barcode sequence and BOLD accession number which can be used by scientist of different countries.

3.17 Data Analysis

Sequence divergences were calculated using Tamura Nei distance model. To provide a graphic representation of it, the mid-point rooted Neighbor Joining (NJ) tree was created. Bootstrap values for Neighbor Joining (NJ) tree was estimated using searches with 1000 pseudo replicates. To infer phylogenetic relationships between the sample species from matrix of sequences, Maximum likelihood(ML) and Maximum Parsimony(MP) analysis were conducted using MEGA 6 (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 validation of 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, 16 species were proceeded for sequencing. They were found to be belonging to 3 orders comprising 6 families, 13 genera and 16 individual species. Out of them six species were taken for analysis. Among Six species three individual of Cypriniformes family Cyprinidae (3) and two individuals from Siluriformes family Heteropneustidae(1) and Clariidae (1) and one individuals from channidae is taken for analysis for further analysis. All the fish specimens were of the class Actinopterygii.

Table 4.1: Family wise number of individuals studied.

Class	Order	Family	*Genus Species*	Individuals Studied
Actinopterygii	Siluriformes	Clariidae Heteropneustidae	*Clarias batrachus* *Heteropneustus fossilis*	1 1
Actinopterygii	Cypriniformes	Cyprinidae Nemacheilidae	*Cyprinus carpio, Catla catla* *Schistura corica*	2 1
Actinopterygii	Perciformes	Channidae	*Channa orientalis*	1

4.2 Morphological Character based identification of the fishes

All the fishes were morphologically identified with their morphological character common Name and local Name as well as two books “ICHTHYOLOGY OF NEPAL" by Dr. Tej Kumar Shrestha and "FISHES OF NEPAL" by Jeewan Shrestha helped very much for identification of the fishes.

Table 4.2: Morphological character based identification of individual’s species

Code	Class	Order	Family	Genus	Species
BNLF01	Actinopterygii	Siluriformes	Clariidae	*Clarias*	*batrachus*
BNLF02	Actinopterygii	Cypriniformes	Heteropneustidae	*Heteropneustus*	*fossilis*
BNLF03	Actinopterygii	Perciformes	Channidae	*Channa*	*orientalis*

BNLF04	Actinopterygii	Cypriniformes	Cyprinidae	*Catla*	*catla*
BNLF05	Actinopterygii	Cypriniformes	Cyprinidae	*Cyprinus*	*carpio*
BNLF06	Actinopterygii	Cypriniformes	Nemacheilidae	*Schistura*	*corica*

4.3: DNA Processing Results

4.3.1 Genomic DNA Processing

After the morphological identification of fishes, they were digitally photographed for digital records. The pectoral fin tissue samples were then used for DNA extraction and forwarded for PCR amplification of DNA. The isolated DNA was examined on 1% agarose gel electrophoresis for quality assurance as shown in figure below **4.1**. Those sample having A260/280 ratio between 1.7 and 2.1 (appendix 3) and un-fragmented genomic band was forwarded for amplification.

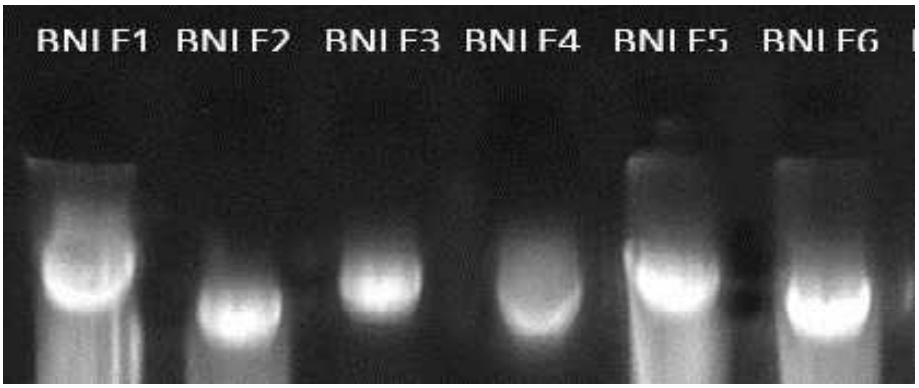


Figure 4.1: Genomic DNA in 1% agarose of six specis (BNLF01 *Clarias batrachus,* BNLF02 *Heteropneustus fossilis,* BNLF03 *Channa orrientali,s* BNLF04 *Catla catla,* BNLF05 *Cyprinus carpio,* BNLF6 *Nemacheilus (Schistura)* corica)

4.3.2 PCR amplification of COI gene

The successful amplification of mitochondrial COI using the cocktail primers in PCR, was confirmed by agarose gel electrophoresis (1% agarose) as shown in Figure **4.2.** Fragment size of approximately 650bp was obtained and verified by 100bp DNA ladder (NEB) run along with it.

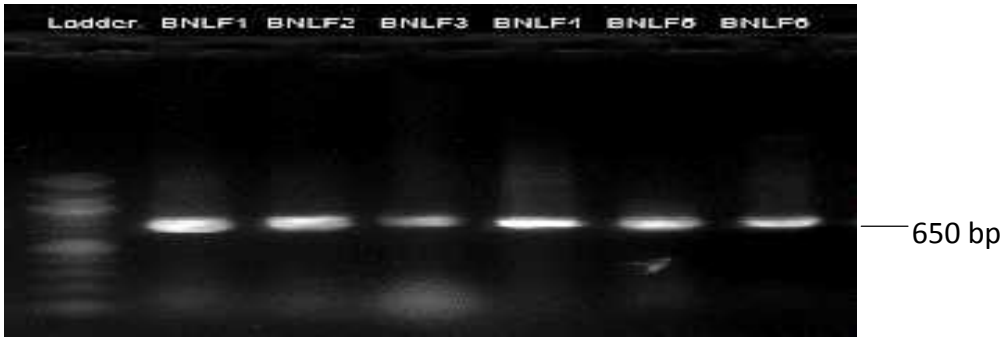


Figure 4.2: Agarose gel Electrophoresis (1%) of PCR product showing 100bp (NEB) Ladder and 650 bp PCR amplified products of samples ID (BNLF01: *Clarias batrachus,* BNLF02: *Heteropneustus fossilis,* BNL3: *Channa orrientali,s* BNLF04: *Catla catla,* BNLF05: *Cyprinus carpio and* BNLF6: *Nemacheilus (Schistura)* corica.

4.4 NCBI and BOLD Verification of Samples with Percentage similarity from BLAST, BOLD and BIN ID (Barcode Index Number) respectively.

Mitochondrial Barcodes for all six species fishes belonging to 3 orders comprising 4 families, 4 genera and 6 individual species was under the study. The specimen sequence based identification was done using NCBI BLAST tool and BOLD identification search engine to verify the morphological character based identification. In case of sample BNLF05 due to low quality trace files it fails to generate BIN ID. The following NCBI and BOLD % similarity and BIN ID was obtained for the sequences in the table 4.3.

Table 4.3: NCBI and BOLD % similarity with accession no. Obtained from gene bank and BOLD submission respectively. Similarity description used in the study: 97%-100% = Significant, 92%-96%= Moderate, ≤91%= Insignificant

Sample	Name of Species	NCBI% Similarity	BOLD% Similarity	NCBI submission number	BIN IDs
BNLF01	*Clarias batrachus*	99%	99.83%	KX249809	BOLD: AAM1926
BNLF02	*Heteropneustus fossilis*	99%	99.8%	KX240819	BOLD: ACR4875
BNLF03	*Channa orrientalis*	98%	99.76%	KX249815	BOLD: ACH0185
BNLF04	*Catla catla*	100%	100%	KX249820	BOLD: AAK2267
BNLF05	*Cyprinus carpio*	100%	100%	KX249814	Not Assign

BNLF06	*Nemacheilus (Schistura)* corica	97%		KX249824	NBOLD: ADB2979

In accordance with the Fish BOL campaign, all sequences and collateral specimen information were deposited within the BOLD and BIN ID was generated automatically. In case of BNLF06 is BOLD recognize it as *Nemacheilus* corica and NCBI as *Schistura* corica both are same with different generic name.

4.5 COI- Based study

Mitochondrial gene sequences have been widely used to infer phylogenetic relationships across different taxonomic levels (Simon et al., 1994). Of the several mitochondrial genes employed for this purpose, the mitochondrial cytochrome oxidase subunit I(COI) and subunit II (COII) have perhaps been the most frequently used. Bioinformatics analysis based on the fragment of the 5' fragment of COI, mitochondrial gene , sequence was done The nucleotide composition transition/transversion rate at each codon position of each barcode sequence, species divergence and barcode gap based on K2P distance between the sequences, nucleotide similarity among the sequences as well as any possible variability in the amino acid residue of important functional site and evolutionary study was carried out which are described in detail in the following sections.

4.5.1 Nucleotide composition

Nucleotide Compositional changes are a major feature of gene or codon evolution. Homogenous or heterogeneous changes in Nucleotide composition leads changes in amino acid sequences which on transmission to progeny gives rise new species. Overlooking nucleotide composition differences among sequences can seriously mislead phylogenetic reconstructions of the nucleotide substitution process used to reconstruct the phylogeny of the important group of species.

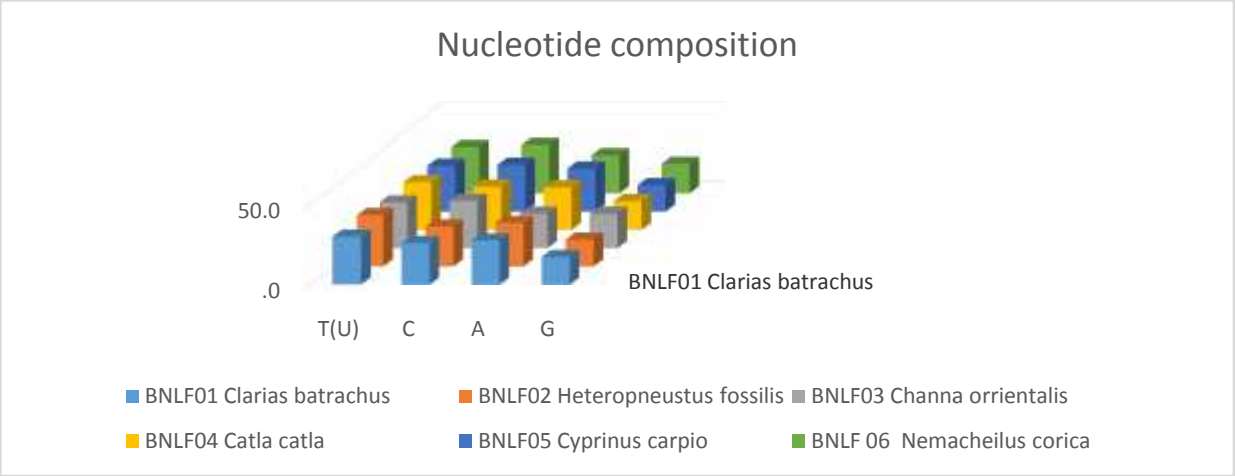


Figure 4.4: Nucleotide composition of six species

The average nucleotide frequencies in COI nucleotide of all the taxa are 25.9% Adenine(A), 29.6% Thymine(T), 26.9% Cytosine(C), and 17.6% Guanine(G). The nucleotide sequence analysis based on the mt DNA COI sequence showed 26.2% (A), 27.8% (T), 23.8% (C), and 22.00% (G) the domination of A: C. Average nucleotide frequencies in COI nucleotide of all the taxa are which shows the high percentage Thyamine. The average nucleotide of composition of Siluriformes family (BNLF01&BNLF03) are T (30.8%), C (25.2%), A (27.1%), G (16.9%). Similarly, the average nucleotide of composition of Cyprinus family (BNLF04, BNLF05, BNLF06) are T (27.0%), C (24.4%), A (27.1%), G (21.5%).

4.5.2 Pairwise Nucleotide composition

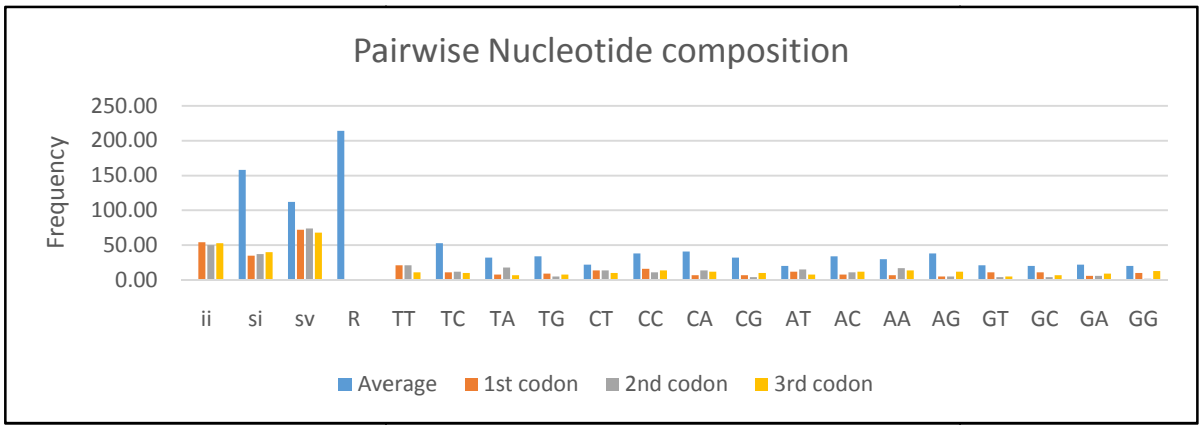


Figure 4.4: Pairwise Nucleotide composition of six species

The first codon shows dominant of AA (1.0) and AT (10.0), the second codon shows dominant of TT (11.0) and AT (10.0), third codon shows dominancy of TT (12.0) TA (10.0)

and GT (10.0) and Transversional Pair (sv) is in higher number (65.0) as shown in the above figure.

4.5.3 Nucleotide frequency at various positions

The nucleotide frequencies within all taxa were counted as displayed in the table below which frequencies are averages (rounded) over all taxa. The frequencies varied to greater extent. In 1st codon position, TT content was 17.00 followed by CC (16.00) TC (12.00) GC AT and GT (11.00) GG (11.00). At 2nd codon position the concentration of TT is highest 22.00, TA (18.00) CT (13.00) &CA (12.00), TG& AG are 5.00, GG (2.00) is less than 5.00. On the contrary, at the 3rd codon position, all pairs of nucleotide were present with CC and AA (14.00) TT (11.0) GG (13.00) being highest followed dinucleotide codon. On average, TT (16.67) was present more commonly followed by CC (13.67) AA (12.00), CT (12.67) and AA is 12.00 and GG (10.). The least occurring were TG, CG, AG and GT overall as illustrated in the figure below.

Table 4.4: Nucleotide frequency at various positions

	TT	TC	TA	TG	CT	CC	CA	CG	AT	AC	AA	AG	GT	GC	GA	GG
1stPos	17	12	9	10	13	16	8	9	11	8	8	6	11	10	6	11
2ndPos	22	15	18	5	15	11	12	4	14	12	14	6	6	4	5	3
3rdPos	11	12	8	7	10	14	13	9	9	12	14	11	6	8	9	10
Avg.	16.6	13	11.6	7.3	12.6	13.6	11	7.3	11.3	10.6	12.0	7.6	7.6	7.3	6.6	8.0

4.5.4 Codon composition of the six sequences

A codon is a sequence of three DNA or RNA nucleotides that corresponds with a specific amino acid or stop signal during protein synthesis. DNA and RNA molecules are written in a language of four nucleotides; meanwhile, the language of proteins includes 20 amino acids. Codons provide the key that allows these two languages to be translated into each other. Each codon corresponds to a single amino acid (or stop signal), and the full set of codons is called the genetic code. The genetic code includes 64 possible permutations, or

combinations, of three-letter nucleotide sequences that can be made from the four nucleotides. Of the 64 codons, 61 represent amino acids, and three are stop signals.

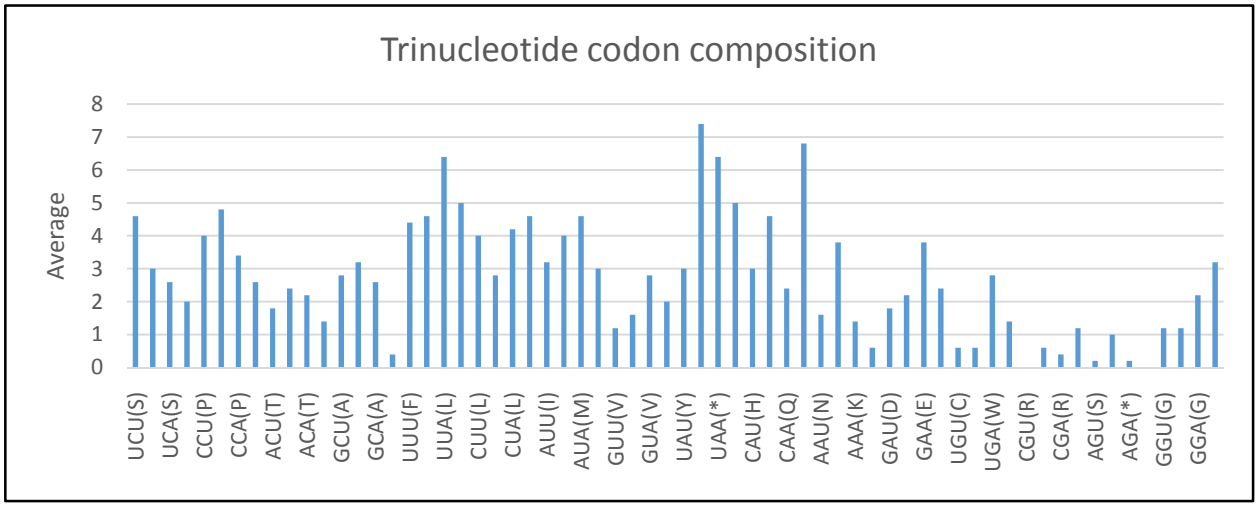


Figure 4.5 Codon composition of the six sequences

For example, the codon CAG represents the amino acid glutamine, and TAA is a stop codon. The genetic code is described as degenerate, or redundant, because a single amino acid may be coded for by more than one codon. When codons are read from the nucleotide sequence, they are read in succession and do not overlap with one another. Just as the rate of change can differ for different nucleotide substitutions. So the rate vary at different sites within a gene. For example, when analyzing protein-coding DNA sequences, researchers immediately noticed that the three different positions within each codon (1, 2, and 3) tend to evolve at different rates, with the third position evolving much more rapidly than the others. This is primarily because the third position can often change without changing the amino acid at that position because of the degeneracy of the genetic code. Because of this, more weight might be given to changes at the first and second positions than to changes at the third positions. Although this is similar in principle to giving more weight to transvers ions than transitions (see above), here, instead of using a substitution matrix or the like, the different positions in the gene are differentially weighted in the phylogenetic reconstruction. All frequencies are averages over all taxa.

4.5.5 Maximum Composite Likelihood (MCL) Estimate of the Pattern of Nucleotide Substitution

Table 4.5: MCL Pattern of Nucleotide Substitution

From /to	A	T	C	G
A	-	*7.34*	*6.45*	**8.9**
T	*6.21*	-	**14.03**	*4..77*
C	*6.21*	**15.96**	-	*5.77*
G	**11.59**	*7.34*	*6.45*	-

Each entry shows the probability of substitution (r) from one base (row) to another base (column). For simplicity, the sum of r values is made equal to 100. Rates of different transitional substitutions are shown in **bold** and those of transversionsal substitutions are shown in *italics*. The nucleotide frequencies are 25.26% (A), 29.57% (T/U), 25.89% (C), and 19.28% (G). The analysis involved 6 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All ambiguous positions were removed for each sequence pair. The transition/transversion rate ratios are k_1 = 1.245 (purines) and k_2 = 2.266 (pyrimidines). The overall transition/transversion bias is R = 0.924, where R = [A*G*k_1 + T*C*k_2]/[(A+G) *(T+C)]. The analysis involved 6 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All ambiguous positions were removed for each sequence pair.

4.5.6 Maximum Likelihood Estimate of Transition/Transversion Bias

In comparing a pair of nucleotide sequences, we distinguish two types of differences; if homologous sites are occupied by different nucleotide bases but both are purines or both pyrimidines, the difference is called “transition”, while, if one of the two is a purine and the other is a pyrimidine, the difference is called “transversion”. The estimated Transition/ Transversion bias (R) is 0.90. Substitution pattern and rates were estimated under the Kimura (1980) 2-parameter model. The nucleotide frequencies are A = 25.00%, T/U = 25.00%, C = 25.00%, and G = 25.00%. For estimating ML values, a tree topology was automatically computed. The maximum Log likelihood for this computation was -2581.860. The analysis involved 6 nucleotide sequences. Codon positions included were 1st+2nd+3rd. All positions containing gaps and missing data were eliminated. There was a total of 418

positions in the final dataset. For estimating ML values, a tree topology was automatically computed. The maximum Log likelihood for this computation was -2377.201. The estimated Transition/ Transversion bias (*R*) is 1.55 for first codon, 1.69 for second Codon and 1.56 for third codon.

4.5.7 Maximum Likelihood Estimate of Substitution Matrix

The number of base substitutions per site from averaging over all sequence pairs are shown in the figure below. Analyses were conducted using the Maximum Composite Likelihood model. The analysis involved five nucleotide sequences. Codon positions included were 1st + 2^{nd} + 3^{rd} + Noncoding. All positions containing gaps and missing data were eliminated. There was a total of 347 positions in the final dataset.

Table 4.6: Maximum Likelihood Estimate of Substitution Matrix

From /to	A	T/U	C	G
A	-	*6.40*	*6.40*	**12.40**
T/U	*6.40*	-	**12.01**	*6.40*
C	*6.40*	**12.20**	-	*6.40*
G	**12.20**	*6.40*	*6.40*	-

Each entry is the probability of substitution (*r*) from one base (row) to another base (column). Substitution pattern and rates were estimated under the Tamura-Nei (1993) model. Rates of different transitional substitutions are shown in **bold** and those of transversionsal substitutions are shown in *italics*. Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 100, the nucleotide frequencies are A = 25.36%, T/U = 28.55%, C = 24.52%, and G = 21.57%. For estimating ML values, a tree topology was automatically computed. The maximum Log likelihood for this computation was -3150.910. The analysis involved 6 nucleotide sequences. Codon positions included were 1^{st} + 2^{nd} + 3^{rd} + Noncoding. All positions containing gaps and missing data were eliminated. There were a total
of 418 positions in the final dataset.

4.5.8 Amino acid composition

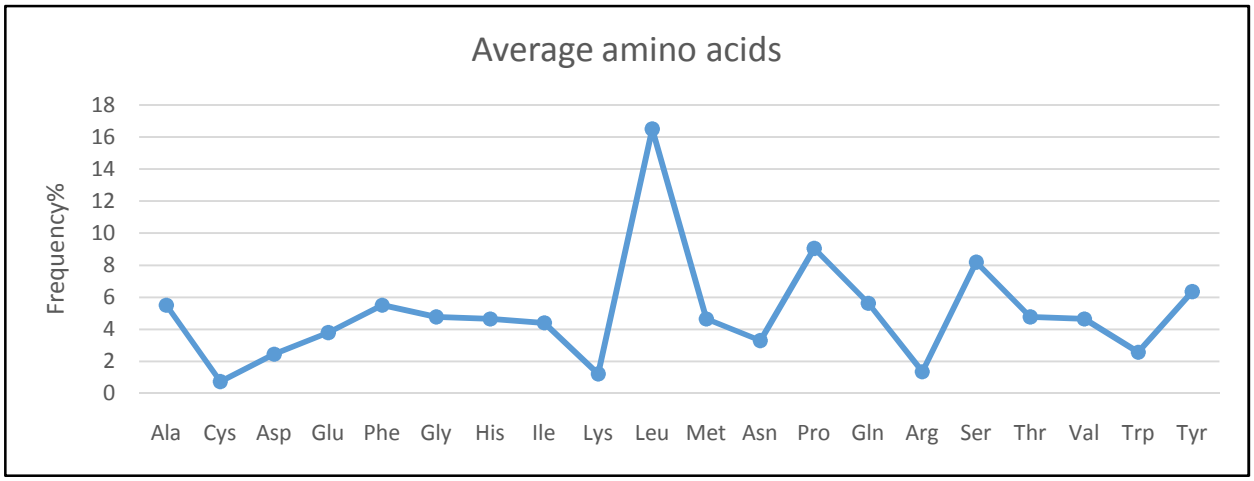


Figure 4.6: Plot illustrating the average amino acid composition in the sequences obtained

All frequencies are given in percent. The overall amino acid composition in all barcode sequences obtained showed high prevalence of hydrophobic amino acid leucine (16.26%) followed by Proline (9.04%), Isoleucine (4.62%), Glycine (4.56%), Alanine (5.50%), Valine (4.64%) Phenylalanine (5.50%) respectively. Among the hydrophilic amino acid sequence serine (8.1%) is the most abundant followed by Asparagine (2.44%), Tryptophan (1.28%), and Histidine (4.32%) which are the key amino acid in enzymatic action of COI. The lowest amino acid consists of tryptophan (2.56%), Cysteine (0.73%) and methylene (4.6%) respectively. All the amino acid composition shows normal appearance of amino acid percentage of hydrophobic group of COI. The amino acid composition of order Cypriniformes(BNLF04, BNLF05) Siluriformes (BNLF01&BNLF02) shows high dominance of Alanine, Cysteine, Glycine, Threonine and Low percent of Asparagine, Glutamine, Phenylalanine, Histidine, Isoleucine, Lysine, Leucine, Methylene, Proline, Glutamine, Arginine, Serine.

4.5.9 Amino acid variability in six individual species

We investigated the variability in amino acid sequences between mitochondrial cytochrome c oxidase subunit I (COI) domains, as well as that of gene sequences encoding the corresponding codons. These results indicate that the heterogeneous substitution rates between COII domains, as well as genes encoding the domains, might be closely related to the inner membrane environment where each region of the amino acid sequence is laid.

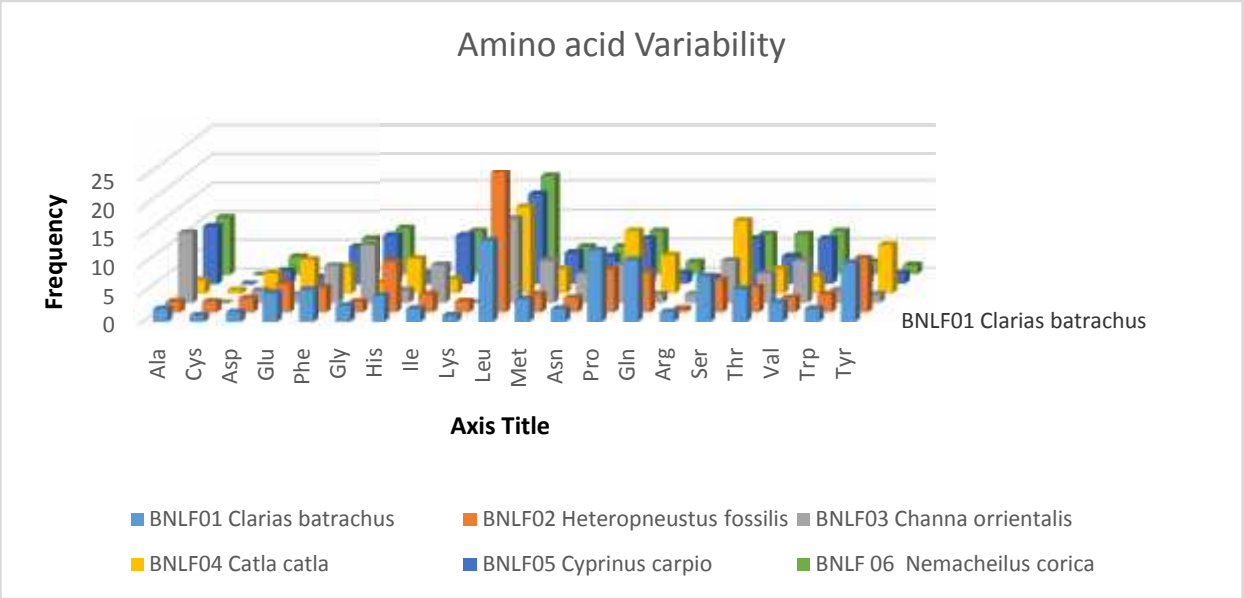


Figure 4.7: Amino acid variability in the individual species

The above figure shows the wide range of Amino acid Variability in the individual species with divergence in its composition too. The average amino acid variability Siluriformes family (*C. batrachus* and *H. fossilils*) shows leucine (18.76%), cystine(1.46%), Histidine(6.45%), Proline (9.9%), Gln(8.79%), Tyrosine (9.6%) and Cyprinoformes shows Alanine (8.0%), cysteine (0.39%) Glycine (6.1%) leucine (11.41%), Proline (9.05%), Gln (3.19%) Tyrosine (3.0%) in average. The amino acid variability of Perciformes family (*C. orientalis*) shows Alanine (12.23%) Glycine (10.07%) Methonine, serenine and valine are 7.19%, and Theronine (28.2%) but cystine and lystine is 0.0%.

4.5.10 Pair wise distance of all species

Pair wise distance-matrix methods of phylogenetic analysis explicitly rely on a measure of "genetic distance" between the sequences being classified, and therefore they require an MSA

(Multiple sequence alignment) as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches. Neighbor-joining methods apply general data clustering techniques to sequence analysis using genetic distance as a clustering metric. The simple neighbor-joining method produces un-rooted trees, but it does not assume a constant rate of evolution across lineages. Pairwise distances of COI gene are shown in Table 4.7. The overall pairwise distance was found to be the pairwise distance of COI sequences among the 6 fish species revealed the shortest genetic distance between species. The lowest distances among the six species is between gb|KX249816|Cyprinus carpio(BNLF05) and gb|KX 249820|BOLD: AAk2267 Catla catla(BNLF04) which is 0.1395 which is expected in Cyprinoformes family. In siluriformes family gb|KX249809|BOLD: AAM1926 Clarias batrachus (BNLF01) and gb|KX249819|BOLD: ACR4875 Heteropneustus fossils (BNLF02) shows 0.1485 barcode gap. The above all barcode gap analysis shows normal barcode gap distances between all the respective species.

Table 4.7: Pair wise distance of all six fish species (BNLF01 *Clarias batrachus,* BNLF02 *Heteropneustus fossilis,* BNLF03 *Channa orrientali,s* BNLF04 *Catla catla,* BNLF05 *Cyprinus carpio,* BNLF6 *Nemacheilus (Schistura)* corica

Species 1	Species 2	Distance
gb|KX249809|BOLD: AAM1926Clarias batrachus(BNLF01)	gb|KX249815|BOLD: ACH0185 Channa orrientalis (BNLF03)	0.2321
gb|KX249809|BOLD: AAM1926 Clarias batrachus(BNLF01)	gb|KX249816|Cyprinus carpio(BNLF05)	0.1965
gb|KX249815|BOLD: ACH0185 Channa orrientalis (BNLF03)	gb|KX249816|Cyprinus carpio(BNLF05)	0.2274
gb|KX249809|BOLD: AAM1926 Clarias batrachus(BNLF01)	gb|KX249819|BOLD: ACR4875 Heteropneustus fossilis(BNLF02)	0.1485
gb|KX249815|BOLD: ACH0185 Channa orrientalis (BNLF03)	gb|KX249819|BOLD: ACR4875 Heteropneustus fossilis(BNLF02)	0.2451
gb|KX249816|Cyprinus carpio(BNLF05)	gb|KX249819|BOLD: ACR4875 Heteropneustus fossilis(BNLF02)	0.2067
gb|KX249809|BOLD: AAM1926 Clarias batrachus(BNLF01)	gb|KX 249820|BOLD: AAk2267 Catla catla(BNLF04)	0.1880
gb|KX249815|BOLD: ACH0185 Channa orrientalis (BNLF03)	gb|KX 249820|BOLD: AAk2267 Catla catla(BNLF04)	0.2297
gb|KX249816|Cyprinus carpio(BNLF05)	gb|KX 249820|BOLD: AAk2267 Catla catla(BNLF04)	0.1395
gb|KX249819|BOLD: ACR4875 Heteropneustus fossilis(BNLF02)	gb|KX 249820|BOLD: AAk2267 Catla catla(BNLF04)	0.1849
gb|KX249809|BOLD: AAM1926 Clarias batrachus(BNLF01)	gb|KX249824|BOLD: ADB2979 Nemacheilus corica (BNLF06)	0.2149

gb|KX249815|BOLD: ACH0185 Channa orrientalis (BNLF03)	gb|KX249824|BOLD: ADB2979 Nemacheilus corica (BNLF06)	0.2208
gb|KX249816|Cyprinus carpio(BNLF05)	gb|KX249824|BOLD: ADB2979 Nemacheilus corica (BNLF06)	0.1869
gb|KX249819|BOLD: ACR4875 Heteropneustus fossilis(BNLF02)	gb|KX249824|BOLD: ADB2979 Nemacheilus corica (BNLF06)	0.2068
gb|KX 249820|BOLD: AAk2267 Catla catla(BNLF04)	gb|KX249824|BOLD: ADB2979 Nemacheilus corica (BNLF06)	0.1771

4.6 Barcode Gap Analysis

The Successfully amplified sequences were analyzed of evolutionary divergence using the K2P model, and the 3% cut-off criteria suggested for species level divergence (Hebert *et al*. 2003), identified clades that were in concordance with recognized taxonomic units based on morphological characters. Our data were all in BOLD and calculated using BOLD analytical tools. The intraspecific divergence over all sequence pair was found to be 2.946 whereas within group of Siluriformes is 3.24 and with Cypriniformes is 3.93. The evolutionary divergence between groups of sequences is 0.226. In case of Perciformes family, the lowest barcode gap of gb|KX249815|BOLD: ACH0185 *C. orrientalis* (BNLF03) is found with KF742420.1 C. orientalis isolate KRF3 and KF742438.1 *C. orientalis* isolate PVF9 with 0.01114 ie 1.14% difference which verify of same species. Similarly, in Cyprinoformes family gb|KX249816|Cyprinus carpio (BNLF05) shows barcode gap of 0.02221 with KF558283.1 Cyprinus carpio isolate Cca A12 shows the nearest resembling species. The intraspecific divergence and interspecific divergence between groups is given in the table below **4.8** A and **4.8** B respectively.

Table 4.8: The intraspecific divergence and interspecific divergence between groups is given in the table below **4.8** A and **4.8** B respectively.

Species 1	Species 2	Distance
Siluriformes	Perciformes	0.239
Siluriformes	Cyprinoformes	0.201
Perciformes	Cyprinoformes	0.227

4.8 A

Species 1	Species 2	Distance
Siluriformes	Perciformes	0.208
Siluriformes	Cyprinoformes	0.047
Perciformes	Cyprinoformes	0.187

4.8 B

4.7 Phylogenetic Analysis

The purpose of this study was to investigate whether the COI barcode provided sufficient resolution to identify genus and species of different fishes. The NJ analysis showed that the COI barcode is an effective tool for identification purposes. All species were resolved as reciprocally despite low COI divergences between some individuals. The phylogenetic tree was generated based on K2P/NJ model in MEGA6 software. The phylogenetic analysis shows Cyprinoformes cluster in one clade (BNLF04, BNLF05, BNLF06), siluriformes(BNLF01, BNLF02) in separate another clade and Perciformes(BNLF03) in other separate clade from Cyprinoformes and siluriformes. The phylogenetic tree shows correct analysis of phylogenetic tree.

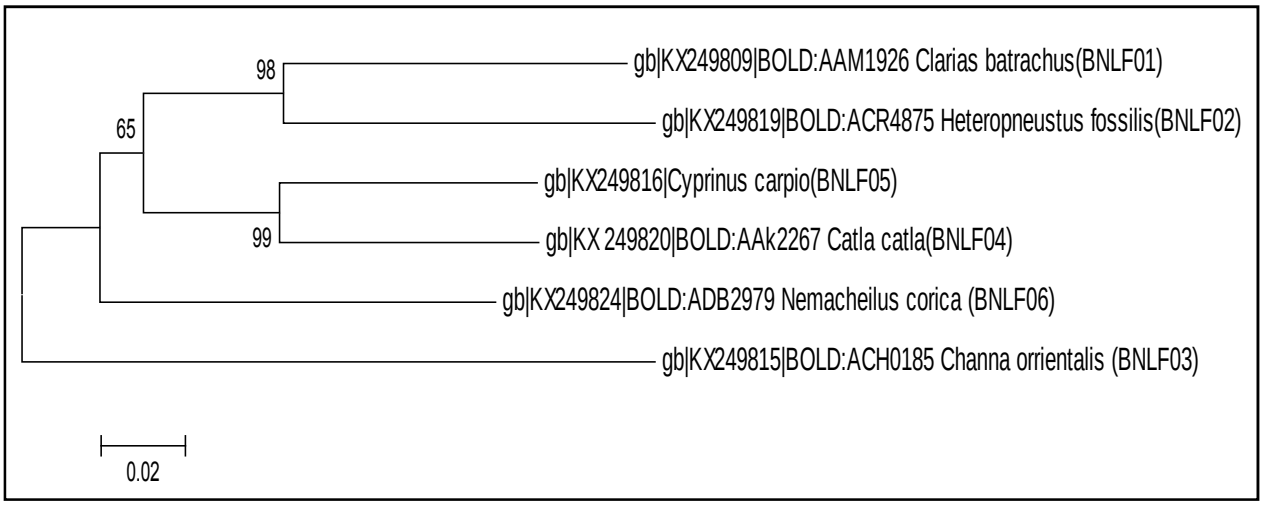


Figure 4.8: Phylogenetic tree inferred using Neighbor joining method based on K2P distance using MEGA6 software. Trees were constructed with the barcode fragment of the COI gene sequences. Numbers given at the main branches refer to bootstrap proportions among 1,000 bootstrap replicates.

4.7.1 Cypriniformes clusters

Cypriniformes is an order of ray-finned fish, including the carps, minnows, loaches and relatives. This order contains 11-12 families, over 400 genera, and more than 4,250 species, with new species being described every few months or so, and new genera being recognized frequently. They are most diverse in southeastern Asia, but are entirely absent from Australia and South America. In this study, the three species Catla *catla(BNLF04), Cyprinus carpio(BNLF05),* and *Nemachelius corica(BNLF06) COI sequence were studied.*

These all common carp belong to the Cyprinidae, the largest freshwater teleost family (Nelson, 1994), and is probably the oldest and most extensively cultured fish species in the world which has been acclimatized to a wide range of habitats and environmental conditions and therefore enjoys a world-wide distribution. The evolutionary history was inferred using the Neighbor-Joining method computed using the Kimura 2-parameter method in the units of the number of base substitutions per site in MEGA6. In the below evolutionary tree *Cyprinus carpio* (BNLF05) shows maximum similarity with *C. carpio* (HQ600723.1) in the same clade along with other *carpio*. *Catla catla* (BNLF04) shows the maximum similarity with *Catla catla* (KF4299.14) isolte CC2608 in the same clade with other *catla*. *Nemacheilus corica* (BNLF06) shows similarity with *S. corica* voucher NBFGR: NC02.

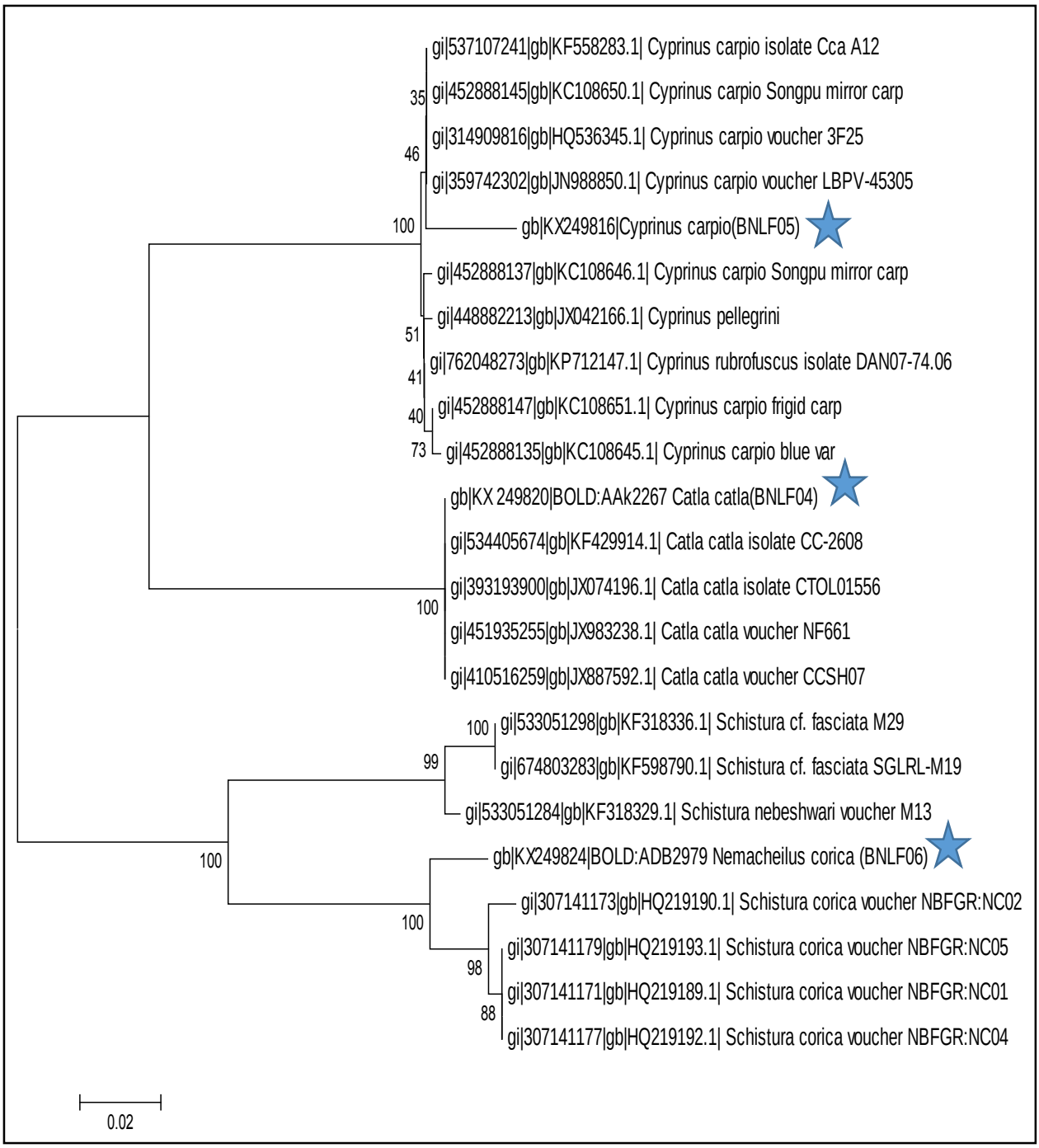


Figure 4.9: K2P distance NJ tree of COI sequences from the species of the Order Cypriniformes analyzed with the sequences obtain from GenBank**.** The arrow sign in the figure shows the position of our sequences in the tree and Numbers given at the main branches refer to bootstrap proportions among 1,000 bootstrap replicates.

4.7.2 Siluriformes clusters

Siluriformes are the diverse group of fish belonging to the superorder Malacopterygii including catfishes having prominent barbels. In siluriformes clustering olny two species were clustered C. batrachus and H. fossilis where second one is exotic in context to Nepal. In NJ based clustering system C. batrachus shows maximum similarities with similar species of karnali and Begnas lake river C. batrachus (KF742432.1) in the same clade. Similarly, H. fossilis shows maximum similariy with H. fossilis (KR021972.1). The evolutionary history was inferred using the Neighbor-Joining method.

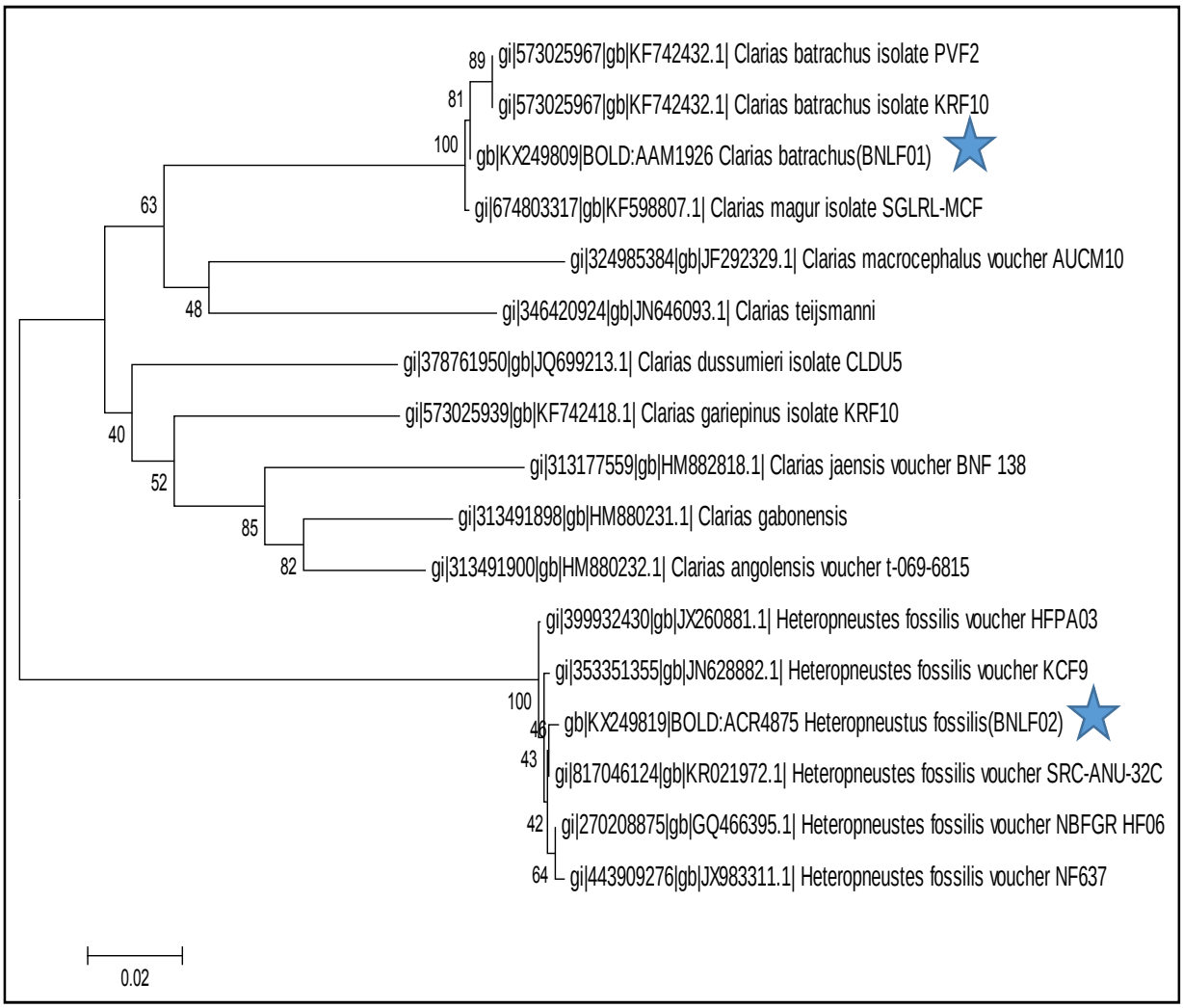


Figure 4.10: K2P distance NJ tree of COI sequences from the species of the Order Siluriformes analyzed in the present work and of GenBank**.** The star sign in the figure shows

the position of our sequences in the tree and Numbers given at the main branches refer to bootstrap proportions among 1,000 bootstrap replicates.

4.7.3 Perciformes cluster

Perciformes are the largest group of all bony fishes including channadae which in indigenous fish of Nepal. In Perciformes clustering only one species Channa orientalis was under study where evolutionary history was inferred using the Neighbor-Joining method in which taxa was clustered together in the bootstrap test (1000 replicates) is shown down below the figures. The evolutionary distances were computed using the Kimura 2-parameter method and are in the units of the number of base substitutions per site in MEGA6 software. *C.* orientalis shows maximum similarity with *C.* orientalis (KF742420.1) and *C.* orientalis (KF42438.1) of koshi river of Nepal.

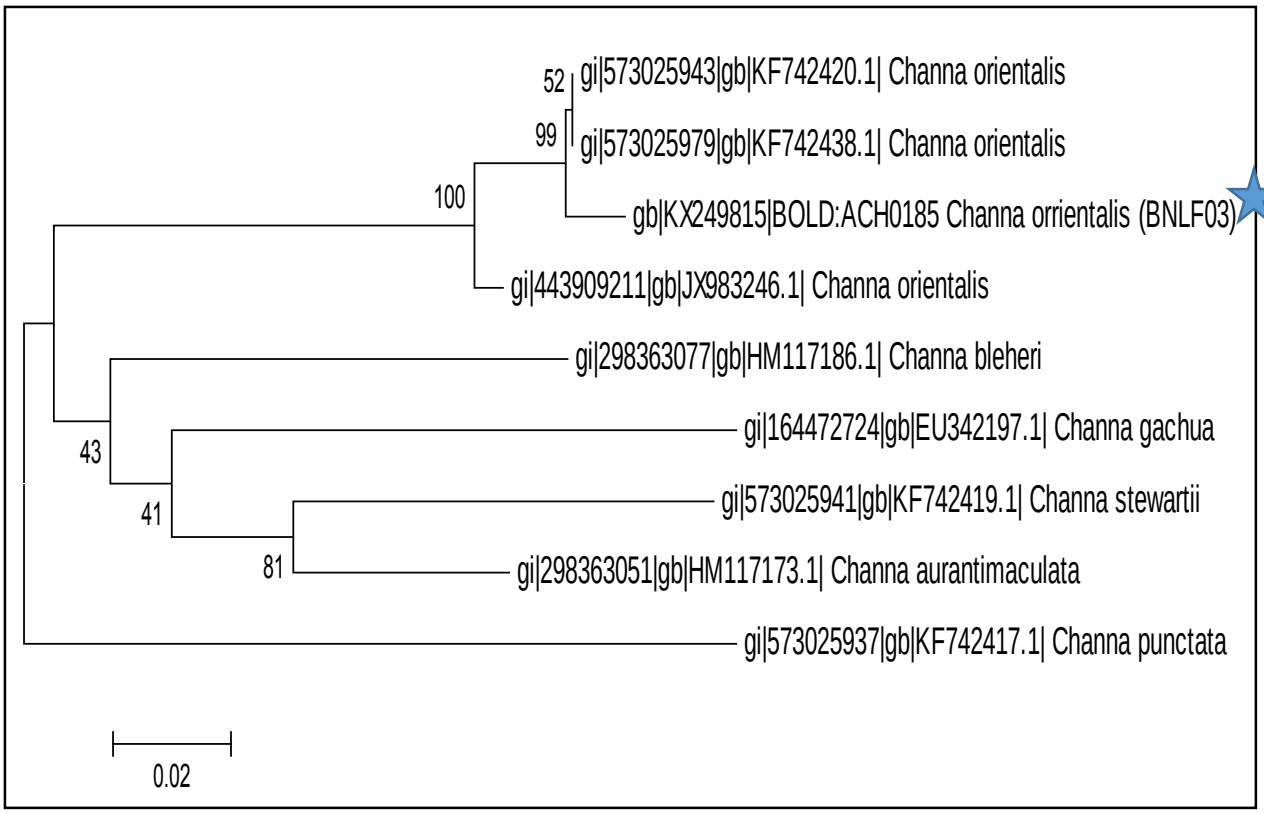


Figure 4.11: K2P distance NJ tree of COI sequences from the species of the Order Perciformes analyzed in the present work and of GenBank. The arrow sign in the figure

shows the position of our sequences in the tree and Numbers given at the main branches refer to bootstrap proportions among 1,000 bootstrap replicates.

4.7.4 Evolutionary Relationship of all six individuals with other maximum similarity sequences

Mitocondrial evolutionary relationship shows common ancestry for populations and species using molecular evidence. mtDNA is inherited from the mother only, no recombination occurs providing a much more direct genetic lineage and ideal for comparing organisms within a species or those who have diverged in a relatively short time.

The below figure shows the evolutionary relationship between Cyprinoformes, Siluriformes and Perciformes family in total having Six individuals. All Six of them show their respective phylogenetic relationship with maximum similarity sequences. *Clarias batrachus* (BNLF01) shows maximum similarity with gi|573025967|gb|KF742432.1| *Clarias batrachus* isolate PVF2 isolated from pokhara valley and gi|573025939|gb|KF742418.1| *Clarias gariepinus* isolate KRF10 from koshi river of Nepal. *Heteropneustus fossilis* (BNLF02) shows similarity with gi|399932430 gb|JX260881.1| *Heteropneustes fossilis* voucher HFPA03. Similarly, *Catla catla* (BNLF04) shows maximum similarity with gi|534405674|gb|KF429914.1| *Catla catla* isolate CC-2608. *Cyprinus carpio* (BNLF05) shows similarity with gi|314909816|gb|HQ536345.1| *Cyprinus carpio* voucher 3F25. *Channa orrientalis* (BNLF03) shows similarity with gi|573025943 |gb|KF742420.1| *Channa orientalis* and gi|573025979|gb|KF742438.1| *Channa orientalis* in same clade which are similar species of Nepal from Koshi River. *Nemacheilus corica* (BNLF06) shows similarity with gi|307141173|gb|HQ219190.1| *Schistura corica* voucher NBFGR: NC02.

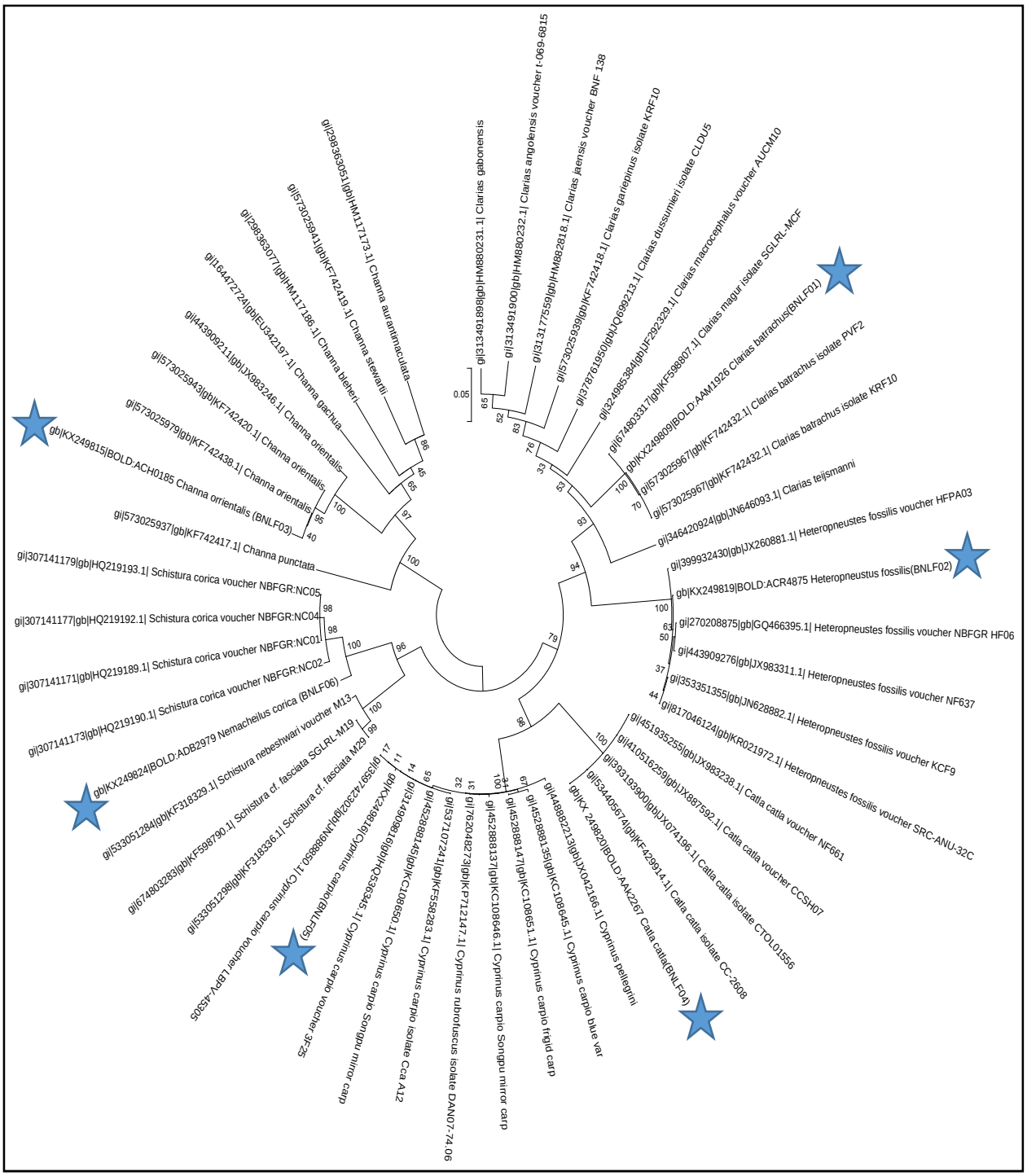


Figure 4.12 Showing evolutionary relationship with respective maximum similarity sequences obtained from Genbank along with six sequences. The star sign in the figure shows the position of our sequences in the evolutionary tree and Numbers given at the main branches refer to bootstrap proportions among 1,000 bootstrap replicates.

CHAPTER 5: DISCUSSION

The present study represents the molecular phylogenetic analysis of Pokhara's freshwater ichthyofauna. Amplification of the COI 5'region (652 bp) was successful for all assayed individuals. A total of five freshwater fishes from Lake Begnas were barcoded.

5.1 Mitochondrial COI as Barcode

This study has strongly supported the efficacy of COI barcodes for diagnosing the freshwater fishes since all 6 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 L.C. *et al.,* 2009).

5.2 Species Identification Based on BLAST and BOLD

In additional to morphological character based identification Nucleotide based identification was performed by online library BLAST (www.ncbi.nlm.nih.gov/BLAST) and BOLD (www.boldsystems.org /identification).Species identification through DNA barcoding is based upon the principle that interspecific divergence sufficiently outscores intraspecific divergence and the biological species can be clearly demarcated by a threshold value, which corresponds to the divergence between the nearest neighbors within a group (Hebert *et al*., 2003). 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 E.H.K. *et al.,* 2008). The Percentage similarity search result of the two species C.

carpio(BNLF02) and Catla catla (BNLF04) shows 100% match with other similar species while other rest species shows more than 97% match BLASTN and BOLD excluding Nemacheilus corica(BNLF06) which donot show any result on BOLD. 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 are 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 neighbor 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. Because of some ambiguities in the K2P distance analysis the experimental fishes taken couldn't be determined for the ranking system.

5.4 Compositional Analysis of COI Sequences

The compositional analysis of COI sequence of six species shows overall mean A= 27.1%, G= 19.6%, C=24.7%, T= 28.5% and CT bias AT=53.3% and GC=45.5%. In cyprinoformes A= 27.0 %, G= 21.5%, C=24.4%, T= 27.0% similarly in Siluriformes A= 27.1%, G= 16.9%, C=25.2%, T= 30.8% and Perciformies shows A= 21.3%, G= 21.3% C=29.2%, T= 28.2% also indicate C: T dominanace.

5.5 Phylogenetic Analysis

Phylogenetic analysis based on Nucleotide sequences helps us study evolutionary history, development and relationship among group of organism. 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 (Hubert *et al.,* 2003). 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. 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). 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 (Hebert *et al.,* 2008) and Indian carangid fishes (Persis *et al.,* 2009). It is hypothesized that genetic divergences should increase with the increasing taxonomic levels. For any particular group of animals this tree could identify the ancestors and closest relatives of the group (Hebert P.D.N. *et al.,* 2003). 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. When genetic distances are low, the K2P model provides the best metric (Nei and Kumar 2000). 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.

5.6 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. 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). The fishes of Cyprinidae family are cosmopolitan but their phyletic classification is ambiguous. The great diversity in family Cyprinidae causes an obstacle for clearing the phylogeny the arbitrary rules and characters were creating a hurdle for the resolution of phylogeny of Cyprinidae. The transition/transversion bias R deviate from the neutral evolution (R = 0.5) Li *et al* also reported similar results. The reason behind this deviation may be due to the structure of nucleotide bases and the complementary base pairing as discussed by Topal and Fresco also this was established by deviation of bias towards transitional mutation.

5.7 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 S.A. *et al.,* 2011). It is needed to broaden the collaboration in order to allow the assembly of global database of fish COI sequence.

CHAPTER 6: SUMMARY AND CONCLUSION

6.1 SUMMARY

Under this study, we have studied Begnas Lake and small rivulets fishes from Pokhara sampling station. Among them 6 fishes 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. All good sequences were analyzed using bioinformatics tools and then deposited to BOLD. They were found to be belonging to out of them five species were taken for analysis. Among Six species three individual of Cypriniformes family Cyprinidae(3) and two individuals from Siluriformes family Heteropneustidae(1) and Clariidae (1) and one from Perciformes family Channdae(1) were proceed for further analysis. All the fish specimens were of the class Actinopterygii. Overall, this study demonstrated the ability of DNA barcoding to help calibrate current taxonomic resolution and shed new light on the biodiversity of fishes of Begnas Lake. Thus, all 6 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, closest relatives were observed to be packed in same groups. Finally, our results stress the need for more taxonomic research, because it appears that even for economically important fishes that have benefited from over a century of scientific inquiry, additional work is required to create a more accurate picture of species diversity. DNA-based methods are not demonstrably more objective, accurate, or useful than orphology or other sources of phenotypic data for species identification or other taxonomic purposes. DNA barcoding as presented by Barrett and Hebert (2005), and DNA taxonomy more generally, is just another technique for species identification that may be useful in particular situations but for which the general utility as a global identification system remains undemonstrated.It is difficult to envisage DNA aiding students to learn a flora or fauna, identify living or preserved specimens, or conduct fieldwork, without first training them to know the organisms, understand their features, and identify them visually.

6.2 CONCLUSION

This study has strongly validated the efficacy of COI barcodes for identifying fish species COI barcoding for species identification is far more powerful than other method. Barcoding discriminated all of the fish species we examined and would clearly be capable of unambiguously identifying individually isolated fish eggs, larvae, fillets and fins from these species degeneracy. This study establishes the feasibility of developing a COI-based identification system for animals-at-large. COI-based identification systems can also aid the initial delineation of species. The general ease of species diagnosis reveals one of the great values of a DNA-based approach to identification. The prospect of using a standard COI threshold to guide species diagnosis in situations where prior taxonomic work has been limited is appealing. It is, however, important to validate this approach by determining the thresholds that distinguish species in other geographical regions and taxonomic groups. Thresholds will particularly need to be established for groups with differences in traits, such as generation length or dispersal regime, that are likely to alter rates of molecular evolution or the extent of population subdivision. The likely applicability of a COI identification system to new animal groups and geographical settings suggests the feasibility of creating an identification system for animals-at-large. Certainly, existing primers enable recovery of this gene from most, if not all, animal species and its sequences are divergent enough to enable recognition of all but the youngest species. Where species bound arise are blurred by hybridization or introgression, supplemental analyses of one or more nuclear genes will be required. Similarly, when species have arisen through polyploidization determinations of genome size may be needed. While protocols will be required to deal with such complications, a COI-based identification system will undoubtedly provide taxonomic resolution that exceeds that which can be achieved through morphological studies. Moreover, the generation of COI profiles will provide a partial solution to the problem of the thinning ranks of morphological taxonomists by enabling a crystallization of their knowledge before they leave the field. This study also has the potential of finding maternal inheritance, recombination, inconsistent mutation rate, heteroplasmy, and compounding evolutionary processes of fish of Begnas Lake.

REFERENCES

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

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

Barrett, R.D.H. and Hebert, P.D.N. (2005) Identifying spiders through DNA barcodes. Can. J. Zool. 83, 481–491

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

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.

CBOL Plant Working Group (2009) A DNA barcode for land plants Proceedings of the National Academy of Sciences 106: 12794–12797.

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

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

Edds, D.R., 1985. New records of fish species for Nepal. J. Nat. Hist. Mus. 9(1-4): 41-

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

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

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*

Gunther, A., 1861. List of the cold-blooded Vertebrata collected by B.B. Hodgson Esq. in Nepal. Proc. Zool. Soc. London, pp. 213-227

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.

Hamilton, F., 1822. An Account of the Fishes found in the River Ganges and its Branches. Edinburg

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

Hebert, P.D.N. *et al*. (2004) Identification of birds through DNA barcodes. PLoS Biol. 2, e312

Hebert, P.D.N. *et al*. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. R. Soc. Lond. B. Biol. Sci. 270 (Suppl. 1), S96–S99

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

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

Hora, S.L., 1937. Notes on fishes in the Indian Museum: On a collection of fish from Nepal. Rec. Ind. Mus. 39: 43-46.

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

Hebert 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

Hollingsworth PM, Graham SW, Little DP. (2011) Choosing and Using a Plant DNA Barcode. Steinke D, ed. PLoS ONE.; 6(5): e19254. doi: 10.1371/journal.pone.0019254.

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

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.

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

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. (2005 Jun 7) Use of DNA barcodes to identify flowering plants Proc Natl Acad Sci U S A.;102(23):8369-74.

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

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

McClelland, J., 1839. Indian Cyprinidae: Second Part of the Nineteenth Volume of Asiatic Researcher; or

Menon, A.G.K., 1949. Notes on Fishes; XLIV - Fish from the Koshi Himalayas, Nepal. Rec. Ind. Mus. 47: 231-237.

Muchlisin Z.A., 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

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

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

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

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)

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

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

Rajbanshi, K. G., 1976. Looping of "Snow Trout" - Asla. J. Science, Kathmandu 6(1): pp.59 - 64.

Rajbanshi, K. G., 1982. A General Bibliography on Fish and Fisheries of Nepal, Royal Nepal Academy, Kamaladi, Kathmandu, Nepal.

Rajbanshi, K. G., 1996. Conservation Status of the Inland Fish Fauna of Nepal. Royal Nepal Academy of Science and Technology, Kathmandu, Nepal.

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

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 P.D.N. (2013) A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN)

System. PLoS ONE 8(8): e66213

Ratnasingham S and Hebert P.D.N. (2007) BOLD: The Barcode of Life data system (http://www.barcodinglife.org). Mol. Ecol. Notes 7: 355–364

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.

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, J., 1981. Fishes of Nepal. C.D.C. Tribhuvan University, Kathmandu, Nepal.

Shrestha, J., 1994. Fishes, Fishing Implements and Methods of Nepal. Smt. M.D Gupta. Lalitpur Colony, Lashkar (Gwalior), India. 150 p.

Shrestha, J., 1995. Enumeration of the Fishes of Nepal. Publication No.10. HMG/N & Govt. of Netherlands. 417/4308. 263p.

Shrestha, J., 1998. Aquatic Habitat and Natural Water Fish and Fisheries in Nepal. In: Environmental Assessment Background Training Paper. ADB TA 2613 Nep., NEAED, Kathmandu, Nepal, 28p.

Shrestha, J., 1999. Cold water fish and fisheries in Nepal. FAO Fisheries Technical Paper. No. 385: 13-40. FAO, Rome.

Shrestha, J., 2001. Taxonomic Revision of Fishes of Nepal. Environment and Agriculture. In: Biodiversity, Agriculture and Pollution in South Asia (P.K. Jha et. al., eds): 171-180. ECOS, Kathmandu.

Shrestha, T.K., 1990. Rare fishes of Himalayan Waters of Nepal. J. Fish Biol. 37 (Suppl.): 213-216.

Shrestha, T.K., 1990a. Resource Ecology of the Himalayan Waters. CDC Tribhuvan University 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.

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

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

T. Petr, Deep Bahadur Swar (2002) Cold water Fisheries in Trans-Himalayan Countries

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

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

Lopez, J. V., Yuhki, N., Masuda, R., Modi, W., and O'Brien, S. J. (1994). Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. 39: 174–190

Ward, R.D. *et al*. (2005) DNA barcoding Australia's fish species. Phil. Trans. R. Soc. Lond. B Biol. Sci. 360, 1847–1857

Websites

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

www.fishbase.org

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

http://www.thekingdomofnepal.com/

http://www.worldfishcenter.org/fishbase

APPENDICES

Appendix 1: Composition of various reagents used.

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

Appendix2: Classification table of species under study including its IUCN red list of threatened species.

Code	Order	Family	Genus Species	IUCN status
BNLF01	Siluriformes	Clariidae	*Clarias batrachus*	Least concern
BNLF02	Siluriformes	Heteropneustidae	*Heteropneustus fossilus*	Least concern
BNLF03	Perciformes	Channidae	*Channa orientalis*	Not accessed in IUCN list yet
BNLF04	Cypriniformes	Cyprinidae	*Catla catla*	Endangered
BNLF05	Cypriniformes	Cyprinidae	*Cyprinus carpio*	Least concern
BNLF06	Cypriniformes	Nemacheilidae	*Schistura Corica*	Not accessed in IUCN list yet

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

Sample Id	A_{260}	A_{280}	$A_{260/280}$	ng/µl	FinalConc. (ng/µl)	Total volume (µl)	DNA conc. taken(µl)	NFW added(µl)
BNLF01 *C. batrachus*	17.060	8.746	1.95	853.01	100	25	2.93	22.07
BNLF02 *H. fossilis*	42.288	22.47	1.93	2100.8	100	25	1.18	23.82
BNLF03 C. *orientalis*	24.6	12.60	1.94	598.87	100	25	4.17	20.83
BNLF04 *Catla catla*	21.69	11.12	1.95	1084.8	100	25	2.30	22.7
BLF05 *Cyprinus carpio*	7.406	3.765	1.97	370.28	100	25	6.75	18.25
BLF06*Schistura corica*	3.903	2.011	1.94	195.15	100	25	12.81	12.19

Appendix 4: List of all the specimens belonging to 3 families Cypriniformes, Siluriformes and Perciformes extracted from GenBank for analysis with the accession numbers.

Cypriniformes:

Name of species	NCBI Accession no.	Name of species	NCBI Accession no.
C. carpio voucher 3F25	HQ536345.1	*C. rubrofuscus* DAN07-74.06	KP712147.1
C. carpio voucher BW-1763	EF609339.1	*Cyprinus carpio 'blue var'*	KC108645.1
C. carpio isolate Cca_A12	KF558283.1	*C. carpio* 'Songpu mirror carp'	KC108646.1
C. carpio 'frigid carp'	KC108651.1	*Cyprinus pellegrini*	JX042166.1
C. carpio voucher LBPV-45304	JN988851.1	*Catla catla* voucher CCSH07	JX887592.1
Catla catla voucher NF661	JX983238.1	*Catla catla* isolate CTOL01556	JX074196.1
Catla catla isolate CC-2608	KF429914.1	*Catla catla* voucher NF696	JX983237.1
Schistura cf. fasciata SGLRL-M19	KF598790.1	*S. corica* voucher BFGR:NC04	HQ219192.1
S. corica voucher NBFGR:NC04	HQ219192.1	*Schistura cf. fasciata* M29	KF318336.1
S. corica voucher NBFGR:NC02	HQ219190.1	*Schistura nebeshwari* voucher M13	KF318329.1

Siluriformes:

Name of species	NCBI Accession no.	Name of species	NCBI Accession no.
C. gariepinus isolate KRF10	KF742418.1	*C. macrocephalus*	JF292329.1
C. batrachus *isolate PVF2*	KF742432.1	*Clarias teijsmanni*	JN646093.1
C. magur SGLRL-MCF	KF598807.1	*C. gabonensis*	HM880231.1
Clarias fuscus	KF011505.1	*Clarias dussumieri*	JQ699213.1

Clarias jaensis	HM882818.1	*Clarias angolensis*	HM880232.1
Clarias gariepinus	KC500413.1	*Heteropneustes fossilis*	JN628882.1
Heteropneustes fossilis	JX260881.1	*Heteropneustes fossilis*	KR021972.1
Heteropneustes fossilis	GQ466395.1	*Heteropneustes fossilis*	JX983311.1

Perciformes:

Name of species	NCBI Accession no.	Name of species	NCBI Accession no.
Channa orientalis	KF742420.1	*Channa bleheri*	HM117186.1
Channa punctata	KF742417.1	*Channa gachua*	EU342197.1
Channa stewartii	KF742419.1	*Channa orientalis*	JX983246.1
Channa orientalis	KF742438.1	*Channa aurantimaculata*	HM117173.1

Appendix 5: Collection Information and Data Sheet

Sample ID	**Collectors**	**Collection Date**	**Continent/ Ocean**	**Country**	**Zone**	**District**	**Sector**	**Exact Site**
BNLF01	Pradip Paudel	1/11/2014	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas lake
BNLF02	Pradip Paudel	3/11/2014	Asia	Nepal	Gandaki	Kaski	Pokhara	Small rivulets
BNLF03	Pradip Paudel	2/11/2014	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas lake
BNLF04	Pradip Paudel	3/11/2014	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas lake
BNLF05	Pradip Paudel	2/11/2014	Asia	Nepal	Gandaki	Kaski	Pokhara	Begnas lake
BLF106	Pradip Paudel	3/11/2014	Asia	Nepal	Gandaki	Kaski	Pokhara	Small rivulets

Appendix 6: Trace sequence of all species

>gb|KX249809|BOLD: AAM1926 *Clarias batrachus* (BNLF01)

[1]TTTATCTAGTATTTGGTGCCTGGGCCGGTATAGTCGGCACAGCCCTAAGCTTACTAATCC[60]
[61]GGGCAGAACTGGCACAACCCGGGGCTCTTTTAGGAGATGACCAGATTTATAATGTTATTG[120]
[121]TTACTGCCCACGCCTTCGTAATAATTTTCTTTATAGTAATACCAATTATGATTGGAGGTT[180]
[181]TCGGAAACTGACTTGTACCTCTAATAATCGGTGCCCCAGATATAGCATTCCCACGAATAA[240]
[241]ATAATATAAGCTTCTGATTACTACCCCCCTCCTTCCTACTGCTACTTGCCTCATCAGGCG[300]
[300]TTGAAGCGGGGGCAGGAACAGGGTGAACAGTATACCCACCCCTTGCAGGAAACCTGGCAC[360]
[361]ATGCAGGAGCTTCTGTAGACTTAACCATTTTTTCTCTACATCTAGCAGGTGTATCATCAA[420]
[421]TTCTTGCCTCCATTAACTTTATCACAACCATTATTAACATGAAACCGCCAGCCATCTCCC[480]
[481]AATATCAAACACCCCTATTTGTTTGATCCGTAATAATCACAGCAGTACTACTACTTCTGT[540]
[541]CCCTTCCAGTATTAGCTGCGGGAATCACTATATTATTAACAGACCGTAATTTAAACACAA[600]
[601]CCTTCTT[607]

>gb|KX249819|BOLD: ACR4875 *Heteropneustus fossilis* (BNLF02)

[1]CAGCCCTTAGCTTACTTATCCGGGCAGAATTAGCACAACCTGGTGCTCTACTGGGTGATG[60]
[61]ACCAAATTTATAACGTTATTGTTACTGCTCACGCATTCGTAATAATTTTCTTTATAGTAA[120]
[121]TACCAATTATGATTGGAGGCTTTGGAAACTGACTAGTACCCCTAATGATTGGAGCCCCTG[180]
[181]ATATAGCATTTCCACGTATGAATAACATAAGCTTCTGACTACTTCCACCATCTTTCCTAC[240]
[241]TACTGCTTGCATCTTCTGGAGTTGAAGCGGGGGCAGGAACAGGATGAACAGTGTATCCAC[300]
[301]CTCTTGCTGGGAATCTTGCACATGCTGGAGCCTCAGTAGATTTAACCATTTTCTCCCTAC[360]
[361]ACTTAGCAGGTGTCTCATCTATTCTAGCATCTATTAATTTTATTACTACTATTATTAACA[420]
[421]TGAAACCCCCAGCCATCTCACAATATCAAACACCACTATTTGTTTGATCAGTGTTAATTA[480]
[481]CAGCCGTACTACTACTACTCTCCCTACCTGTACTAGCCGCTGGAATTACCAT[532]

>gb|KX249815|BOLD: ACH0185 *Channa orrientalis* (BNLF03)

[1]ATAGTCGGCACCGCACTGAGCCTACTGATCCGGGCTGAACTTAGCCAGCCCGGTGCTCTT[60]
[61]CTAGGCAACGACCAAATTTATAATGTAATTGTTACGGCCCACGCCTTCGTCATGATCTTC[120]
[121]TTCATGGTAATGCCAATAATAATCGGGGGCTTTGGAAACTGACTGGTCCCGCTTATGATC[180]
[181]GGCGCCCCTGACATAGCCTTCCCTCGAATAAACAATATGAGTTTTTGACTTCTCCCCCCT[240]
[241]TCTTTCCTCCTTCTTCTGGCCTCTTCTGCAGTAGAAGCCGGAGCTGGGACAGGCTGGACA[300]
[301]GTTTACCCACCTTTAGCTGGCAATCTGGCTCACGCGGGGACATCCGTAGACCTAGCCATC[360]
[361]TTCTCTTTACACCTTGCAGGTGTCTCTTCAATTTTAGGGGCAATTAACTTCACCACGA[418]

>gb|KX 249820|BOLD: AAk2267 *Catla catla* (BNLF04)

[1]TTTATCTCGTATTTGGTGCCTGAGCCGGAATAGTAGGAACCGCCTTAAGCCTTCTCATCC[60]
[61]GGGCTGAACTAAGTCAACCCGGATCGCTTCTAGGTGATGACCAAATTTATAATGTTATTG[120]
[121]TAACTGCTCACGCCTTCGTAATAATTTTCTTTATAGTAATACCTATCCTCATTGGAGGAT[180]
[181]TTGGAAACTGACTCGTGCCATTAATGATCGGAGCCCCAGATATGGCATTCCCCCGTATAA[240]

241ATAATATAAGCTTCTGACTCCTACCCCCATCATTCCTATTACTACTAGCCTCCTCTGGTG300
301TAGAAGCTGGGGCTGGAACAGGATGAACAGTATATCCACCTCTTGCAGGCAACTTAGCCC360
361ACGCAGGAGCATCAGTAGACCTAACAATTTTCTCACTCCACTTGGCAGGAGTTTCATCAA420
421CTTGGCAGGAGTTTCATCAATCCTAGGGGCTATTAATTTCATCACCACAACTATTAATAT480
481AATATCAAACACCTTTATTTGTCTGATCCGTACTTGTAACCGCCGTACTACTTCTCCTAT540
541CGCTACCAGTACTGGCCGCTGGCATTACA 569

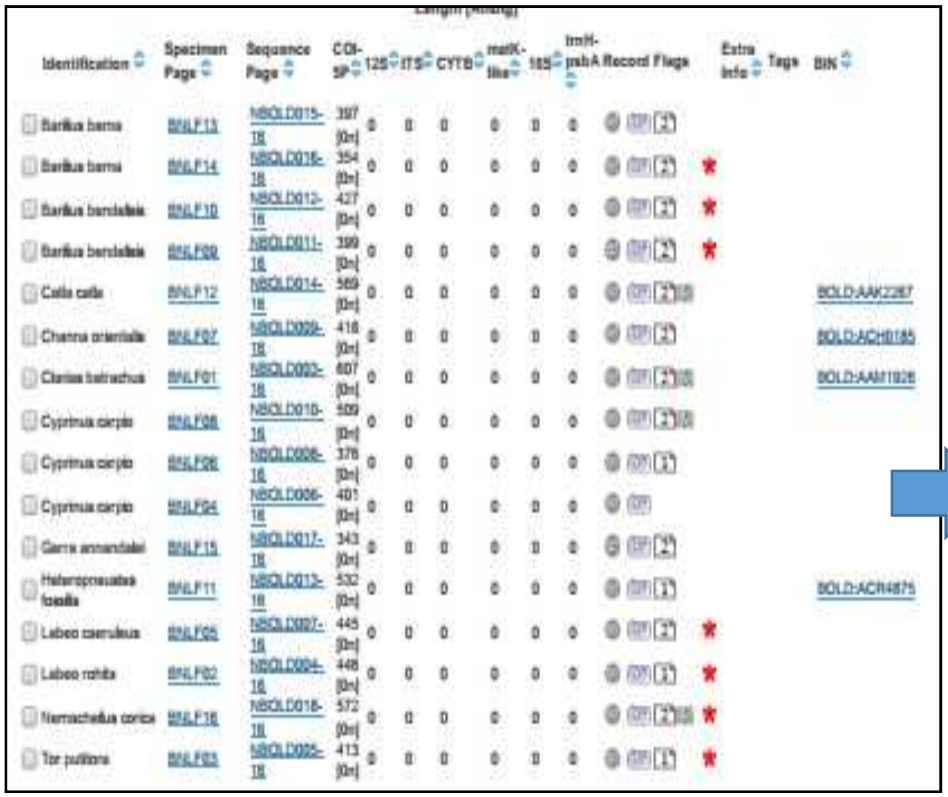
>gb|KX249816|*Cyprinus carpio* (BNLF05)

1ACCGCCTTAAGCCTCCTCATTCGGGCCGAACTTAGCCAACCCGGGTCGCTTCTAGGTGAT60
61GACCAAATTTATAACGTTATCGTCACTGCCCACGCCTTTGTAATAATTTTCTTTATAGTA120
121ATGCCTATCCTTATTGGAGGATTTGGAAACTGACTTGTACCACTAATAATCGGAGCCCCA180
181GACATAGCATTCCCACGAATAAATAACATAAGCTTCTGACTACTACCCCCATCATTCCTT240
241CTACTCCTAGCTTCTTCTGGTGTTGAAGCTGGAGCCGGAACAGGATGAACCGTATACCCA300
301CCTCTTGCAGGGAACTTAGCCCACGCAGGAGCATCAGTAGACCTAACAATTTTCTCACTT360
361CACCTAGCAGGTGTTTCATCAATTCTAGGGGCAATCAACTTTATTATTACAAACATCAAC420
421ATGAAACCCCCAGCCATCTCTCAATACAAAACACCCCTGTTCGTCTGATCCGTGCTTGTA480
481ACCGCTATGGTCTTCCTTCTACCTTTACC 509

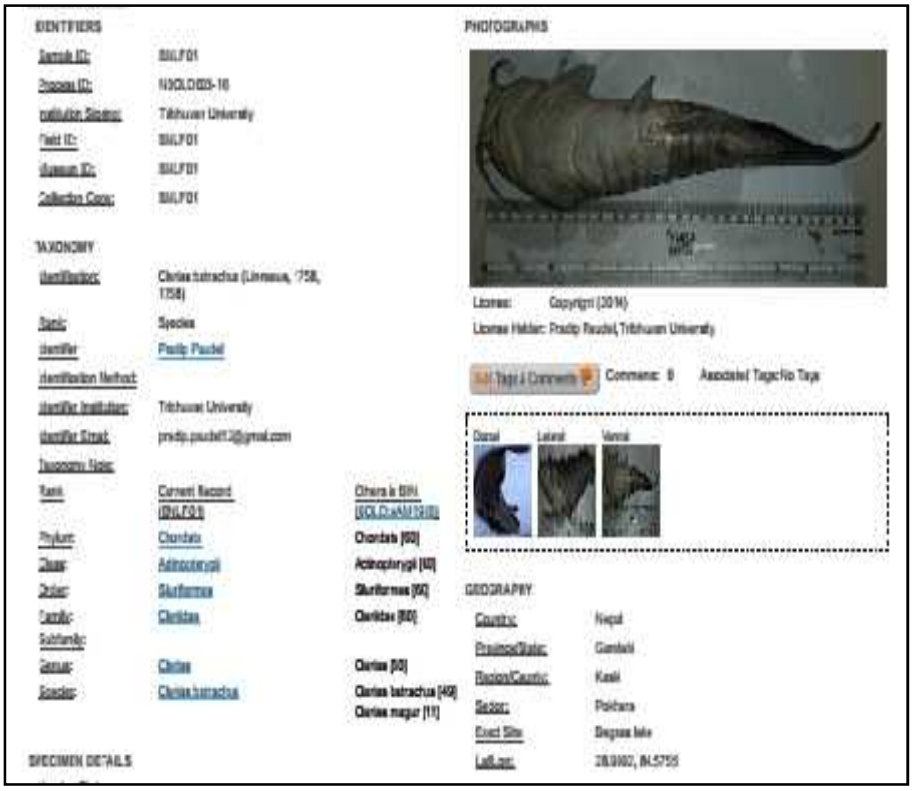
>gb|KX249824|BOLD: ADB2979 *Nemacheilus corica* (BNLF06)

1CTTAGCCTTCTAATCCGAGCTGAACTAAGCCAACCGGGATCCCTTCTGGGTGATGACCAA60
61ATTTATAATGTTATTGTTACTGCCCACGCTTTTGTTATAATTTTCTTTATAGTAATGCCT120
121ATCCTTATTGGGGGGTTTGGAAACTGACTCGTACCACTAATGATTGGAGCCCCCGATATG180
181GCATTCCCACGGATAAATAATATAAGCTTCTGACTTCTACCCCCCTCCTTTCTCCTGCTA240
241TTGGCCTCATCCGGCGTAGAAGCCGGGGCCGGGACAGGATGAACGGTCTACCCCCCACTA300
301GCTGGGAACCTAGCTCACGCAGGTGCCTCAGTAGATTTAACCATTTTCTCCTTACACCTT360
361GCCGGTGTCTCATCCATCCTCGGGGCAATTAATTTTATTACAACAACAATTAATATAAAA420
421CCCCCAGCCATCTCCCAGTACCAAACCCCCCTGTTCGTGTGGGCAGTCCTCGTAACCGCC480
481GTCCTCCTTCTTCTATCACTACCAGTTCTGGCCGCCGGAATTACTATGCTCTTGACAGAC540
541CGAAACTTAAATACCACATTCTTTGACCCCGC572

Appendix 7: Shows the glimpse of BOLD system



Identification	Specimen Page	Sequence Page	COI-5P	12S	ITS	CYTB	matK-like	18S	trnH-psbA	Record Flags	Extra Info	Tags	BIN
Barilius barna	BNLF13	NBOLD015-18	397 [0m]	0	0	0	0	0	0				
Barilius barna	BNLF14	NBOLD016-18	354 [0m]	0	0	0	0	0	0				
Barilius bendelisis	BNLF10	NBOLD012-18	427 [0m]	0	0	0	0	0	0				
Barilius bendelisis	BNLF09	NBOLD011-18	399 [0m]	0	0	0	0	0	0				
Catla catla	BNLF12	NBOLD014-18	569 [0m]	0	0	0	0	0	0				BOLD:AAK2287
Channa orientalis	BNLF07	NBOLD009-18	418 [0m]	0	0	0	0	0	0				BOLD:ACH0185
Clarias batrachus	BNLF01	NBOLD002-18	607 [0m]	0	0	0	0	0	0				BOLD:AAM1926
Cyprinus carpio	BNLF08	NBOLD010-18	509 [0m]	0	0	0	0	0	0				
Cyprinus carpio	BNLF06	NBOLD008-18	378 [0m]	0	0	0	0	0	0				
Cyprinus carpio	BNLF04	NBOLD006-18	401 [0m]	0	0	0	0	0	0				
Garra annandalei	BNLF15	NBOLD017-18	343 [0m]	0	0	0	0	0	0				
Heteropneustes fossilis	BNLF11	NBOLD013-18	532 [0m]	0	0	0	0	0	0				BOLD:ACN4875
Labeo caeruleus	BNLF05	NBOLD007-18	445 [0m]	0	0	0	0	0	0				
Labeo rohita	BNLF02	NBOLD004-18	448 [0m]	0	0	0	0	0	0				
Nemacheilus corica	BNLF16	NBOLD018-18	572 [0m]	0	0	0	0	0	0				
Tor putitora	BNLF03	NBOLD005-18	413 [0m]	0	0	0	0	0	0				



IDENTIFIERS

Sample ID: BNLF01
Process ID: NBOLD002-18
Institution Storing: Tribhuvan University
Field ID: BNLF01
Museum ID: BNLF01
Collection Code: BNLF01

TAXONOMY

Identification: Clarias batrachus (Linnaeus, 1758, 1758)
Rank: Species
Identifier: Pradip Paudel
Identification Method:
Identifier Institution: Tribhuvan University
Identifier Email: pradip.paudel12@gmail.com
Taxonomy Note:

Rank	Current Record (BNLF01)	Others in BIN (BOLD:AAM1926)
Phylum:	Chordata	Chordata [60]
Class:	Actinopterygii	Actinopterygii [60]
Order:	Siluriformes	Siluriformes [60]
Family:	Clariidae	Clariidae [60]
Subfamily:		
Genus:	Clarias	Clarias [60]
Species:	Clarias batrachus	Clarias batrachus [49] Clarias magur [11]

SPECIMEN DETAILS

PHOTOGRAPHS

License: Copyright (2016)
License Holder: Pradip Paudel, Tribhuvan University

Tags | Comments Comments: 0 Associated Tags: No Tags

GEOGRAPHY

Country: Nepal
Province/State: Gandaki
Region/County: Kaski
Sector: Pokhara
Exact Site: Begnas lake
Lat/Lon: 28.0902, 84.5755

NBOLD Project Page

Specimen Identification Page

Sequence Barcode and trace file Page

IDENTIFIERS

Sample ID:	BNLF01
Process ID:	NBOLD003-16
Identification:	Clarias batrachus
BIN:	BOLD:AAM1926

COI-SP

SEQUENCE DATA

Genbank Accession:

Translation Matrix: Vertebrate Mitochondrial

Last Updated: 2016-02-19

Clear Sequence | Edit Sequence

NUCLEOTIDE SEQUENCE

Sequence: 607 bp

TTTATCTAGTATTTGGTGCCTGGGCCGGTATAGTCGGCACAGCCCTAAGCTTACTAATCC
GGGCAGAACTGGCACAACCCGGGGCTCTTTTAGGAGATGACCAGATTTATAATGTTATTG
TTACTGCCCACGCCTTCGTAATAATTTTCTTTATAGTAATACCAATTATGATTGGAGGTT
TCGGAAACTGACTTGTACCTCTAATAATCGGTGCCCCAGATATAGCATTCCCACGAATAA
ATAATATAAGCTTCTGATTACTACCCCCCTCCTTCCTACTGCTACTTGCCTCATCAGGCG
TTGAAGCGGGGGCAGGAACAGGGTGAACAGTATACCCACCCCTTGCAGGAAACCTGGCAC
ATGCAGGAGCTTCTGTAGACTTAACCATTTTTTCTCTACATCTAGCAGGTGTATCATCAA
TTCTTGCCTCCATTAACTTTATCACAACCATTATTAACATGAAACCGCCAGCCATCTCCC
AATATCAAACACCCCTATTTGTTTGATCCGTAATAATCACAGCAGTACTACTACTTCTGT
CCCTTCCAGTATTAGCTGCGGGAATCACTATATTATTAACAGACCGTAATTTAAACACAA
CCTTCTT

Composition: A (166), G (104), C (156), T (181)

Ambiguous Characters: 0

Identify Sequence Using:

Full DB | Species DB | Published DB | Full Length DB

AMINO ACID SEQUENCE

Sequence: 213 residues

YLVFGAWAGMVGTALSLLIRAELAQPGALLGDDQIYNVIVTAHAFVMIFFMVMPIMIGGF
GNWLVPLMIGAPDMAFPRMNNMSFWLLPPSFLLLLASSGVEAGAGTGWTVYPPLAGNLAH
AGASVDLTIFSLHLAGVSSILASINFITTIINMKPPAISQYQTPLFVWSVMITAVLLLLS
LPVLAAGITMLLTDRNLNTTFX

ILLUSTRATIVE BARCODE

SEQUENCING RUNS: Tribhuvan University

Run Date	Direction	Trace File	Seq Primer	Quality
PCR Primers: VF2_t1/FR1d_t1				
2015-02-16	Reverse	BNLF01__R_A02.ab1	M13R	med
2015-02-16	Forward	BNLF01__F_A01.ab1	M13F	low

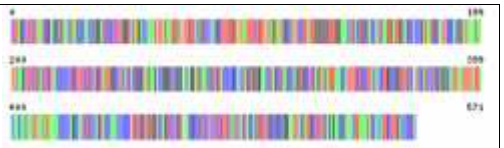
Sequence Editor | View Trace Files | Download

ANNOTATION

Add Tags & Comments Comments: 0 Associated Tags: No Tags

Appendix 8: Sequence and illustrative barcode for each six species

Species identification number	NCBI submission number	BOLD accession number	Species image	Sequence	Barcode
BNLF01	KX249809	BOLD:AAM1926		TTTATCTAGTATTTGGTGCCTGGGCCGGTATAGTCGGCACAGCCCTAAG CTTACTAATCCGGGAGATGACCAGGCAGAACTGGCACAACCCGGGGCT CTTTTAGATTTATAATGTTATTGTTACTGCCCACGCCTTCGTAATAATTTT CTTTATAGTAATACCAATTATGATTGGAGGTTTCGGAAACTGACTTGTAC CTCTAATAATCGGTGCCCCAGATATAGCATTCCCACGAATAAATAATATA AGCTTCTGATTACTACCCCCCTCCTTCCTACTGCTACTTGCCTCATCAGGC GTTGAAGCGGGGGGCAGGAACAGGGTGAACAGTATACCCACCCCTTGCA GGAAACCTGGCACATGCAGGAGCTTCTGTAGACTTAACCATTTTTTCTCT ACATCTAGCAGGTGTATCATCAATTCTTGCCTCCATTAACTTTATCACAA CCATTATTAACATGAAACCGCCAGCCATCTCCCAATATCAAACACCCCTA TTTGTTTGATCCGTAATAATCACAGCAGTACTACTACTTCTGTCCCTTCCA GTATTAGCTGCGGGAATCACTATATTATTAACAGACCGTAATTTAAACA CAACCTTCTT	
BNLF02	KX240819	BOLD: ACR4875		CAGCCCTTAGCTTACTTATCCGGGCAGAATTAGCACAACCTGGTGCTCT ACTGGGTGATGACCAAATTTATAACGTTATTGTTACTGCTCACGCATTCG TAATAATTTTCTTTATAGTAATACCAATTATGATTGGAGGCTTTGGAAAC TGACTAGTACCCCTAATGATTGGAGCCCCTGATATAGCATTTCCACGTAT GAATAACATAAGCTTCTGACTACTTCCACCATCTTTCCTACTACTGCTTGC ATCTTCTGGAGTTGAAGCGGGGGCAGGAACAGGATGAACAGTGTATCC ACCTCTTGCTGGGAATCTTGCACATGCTGGAGCCTCAGTAGATTTAACC ATTTTCTCCCTACACTTAGCAGGTGTCTCATCTATTCTAGCATCTATTAAT TTTATTACTACTATTATTAACATGAAACCCCCAGCCATCTCACAATATCAA ACACCACTATTTGTTTGATCAGTGTTAATTACAGCCGTACTACTACTACT CTCCCTACCTGTACTAGCCGCTGGAATTACCAT	
BNLF03	KX249815	BOLD: ACH0185		ATAGTCGGCACCGCACTGAGCCTACTGATCCGGGCTGAACTTAGCCAGC CCGGTGCTCTTCTAGGCAACGACCAAATTTATAATGTAATTGTTACGGCC CACGCCTTCGTCATGATCTTCTTCATGGTAATGCCAATAATAATCGGGG GCTTTGGAAACTGACTGGTCCCGCTTATGATCGGCGCCCCTGACATAGC CTTCCCTCGAATAAACAATATGAGTTTTTGACTTCTCCCCCCTTCTTTCCT CCTTCTTCTGGCCTCTTCTGCAGTAGAAGCCGGAGCTGGGACAGGCTGG ACAGTTTACCCACCTTTAGCTGGCAATCTGGCTCACGCGGGGACATCCG TAGACCTAGCCATCTTCTCTTTACACCTTGCAGGTGTCTCTTCAATTTTAG GGGCAATTAACTTCACCACGA	

BNLF04	KX249820	BOLD: AAK2267		TTTATCTCGTATTTGGTGCCTGAGCCGGAATAGTAGGAACCGCCTTAAG CCTTCTCATCCGGCTGAACTAAGTCAACCCGGATCGCTTCTAGGTGATG ACCAAATTTATAATGTTATTGTAACTGCTCACGCCTTCGTAATAATTTTCT TTATAGTAATACCTATCCTCATTGGAGGATTTGGAAACTGACTCGTGCCA TTAATGATCGGAGCCCCAGATATGGCATTCCCCCGTATAAATAATATAA GCTTCTGACTCCTACCCCCATCATTCCTATTACTACTAGCCTCCTCTGGTG TAGAAGCTGGGGCTGGAACAGGATGAACAGTATATCCACCTCTTGCAG GCAACTTAGCCCACGCAGGAGCATCAGTAGACCTAACAATTTTCTCACT CCACTTGGCAGGAGTTTCATCAATCCTAGGGGCTATTAATTTCATCACCA CAACTATTAATATGAAACCTCCGGCCATCTCACAATATCAAACACCTTTA TTTGTCTGATCCGTACTTGTAACCGCCGTACTACTTCTCCTATCGCTACCA GTACTGGCCGCTGGCATTACA	
BNLF06	KX249824	NBOLD: ADB2979		ACCGCCTTAAGCCTCCTCATTCGGGCCGAACTTAGCCAACCCGGGTCGC TTCTAGGTGATGACCAAATTTATAACGTTATCGTCACTGCCCACGCCTTT GTAATAATTTTCTTTATAGTAATGCCTATCCTTATTGGAGGATTTGGAAA CTGACTTGTACCACTAATAATCGGAGCCCCAGACATAGCATTCCCACGA ATAAATAACATAAGCTTCTGACTACTACCCCCATCATTCCTTCTACTCCTA GCTTCTTCTGGTGTTGAAGCTGGAGCCGGAACAGGATGAACCGTATAC CCACCTCTTGCAGGGAACTTAGCCCACGCAGGAGCATCAGTAGACCTAA CAATTTTCTCACTTCACCTAGCAGGTGTTTCATCAATTCTAGGGGCAATC AACTTTATTATTACAAACATCAACATGAAACCCCCAGCCATCTCTCAATA CAAAACACCCCTGTTCGTCTGATCCGTGCTTGTAACCGCTATGGTCTTCC TTCTACCTTTACC	
BNLF05	KX249814	Not Assign		CTTAGCCTTCTAATCCGAGCTGAACTAAGCCAACCGGGATCCCTTCTGG GTGATGACCAAATTTATAATGTTATTGTTACTGCCCACGCTTTTGTTATA ATTTTCTTTATAGTAATGCCTATCCTTATTGGGGGGTTTGGAAACTGACT CGTACCACTAATGATTGGAGCCCCCGATATGCATTCCCACGGATAAATA ATATAAGCTTCTGACTTCTACCCCCCTCCTTTCTCCTGCTATTGGCCTCAT CCGGCGTAGAAGCCGGGGCCGGGACAGGATGAACGGTCTACCCCCCAC TAGCTGGGAACCTAGCTCACGCAGGTGCCTCAGTAGATTTAACCATTTT CTCCTTACACCTTGCCGGTGTCTCATCCATCCTCGGGGCAATTAATTTTA TTACAACAACAATTAATATAAAACCCCCAGCCATCTCCCAGTACCAAACC CCCCTGTTCGTGTGGGCAGTCCTCGTAACCGCCGTCCTCCTTCTTCTATC ACTACCAGTTCTGGCCGCCGGAATTACTATGCTCTTGACAGACCGAAAC TTAAATACCACATTCTTTGACCCCGC