Chapter 1. INTRODUCTION

1.1 Freshwater fishes: biodiversity and taxonomy

1.1.1 Ichthyology

Fishes constitute almost half of the total number of vertebrates in the world. They exhibit enormous diversity in their morphology, in the habitats they occupy and in their biology. The term "fishes" is not a taxonomic rank rather it is used conveniently to describe those vertebrates studied by ichthyologists and covered in ichthyological courses. Some studies have restricted the term "fish" to the jawed bony fishes, namely, Actinopteygii, Latimeridae, and Dipnoi. Some studies included shark, rays and other closely related members. Nelson used the term to designate an assemblage of a paraphyletic group, where the most recent common ancestor was included but some descendants from the common ancestors (like tetrapods) were not included. He concluded that despite their diversity fishes can be simply but artificially, defined as aquatic vertebrates that have gills throughout life and limbs, if any, in the form of fins (Nelson 1976, 1984). The field of ichthyology, the study of fish systematics, is enormously vast and exciting. Many controversies and problems exist and ichthyologists have numerous opportunities to make discoveries of new taxa, both extinct and extant, and to address phylogenetic and biogeographic problems.

Fishes live in almost all aquatic habitats. Primarily fishes are divided into two broad groups: freshwater and marine. The estimated 13,000 strictly freshwater fish species live in lakes and rivers that cover only 1% of the earth's surface, while the remaining 16,000 species live in salt water covering a full 70% (Lévêque et al. 2008). While freshwater species belong to some 170 families (or 207 if peripheral species are also considered), the bulk of species occur in a relatively few groups: the Characiformes, Cypriniformes, Siluriformes, and Gymnotiformes, the Perciformes (notably the family Cichlidae), and the Cyprinodontiformes.

1.1.2 Global geographical distribution of freshwater fishes

Freshwater fishes are one of the most important groups zoogeographically because they are more or less confined to drainage systems which can be thought of as dendritic islands of water surrounded by land, which in turn is bordered by salt water barrier. The freshwater fishes provide a relatively conservative system for examining patterns of distribution that may reflect continental changes. The family is the taxon that best reflects the evolution and dispersal of a group, and, infact, zoogeographic patterns form strong evidence of evolution.

The present distribution of freshwater fishes has been shaped by millions of years of changes in the global water cycle. In relation to climate change, the nature and dynamics of surface freshwater systems have evolved continuously, at various spatial and temporal scales. Many of the surface freshwater systems have, therefore, been transient; their fauna and flora usually disappeared when the systems disappeared, or were able to survive by developing adaptations to the changing circumstances. The dual processes of speciation and extinction have interacted with climatic and geological events that have both isolated fish populations and provided opportunities for migration and colonization of new habitats (Berra 2001).

Biogeographically the distribution of strictly freshwater species and genera are, respectively 4,035 species (705 genera) in the Neotropical region, 2,938 (390 genera) in the Afrotropical, 2,345 (440 genera) in the Oriental, 1,844 (380 genera) in the Palaearctic, 1,411 (298 genera) in the Nearctic, and 261 (94 genera) in the Australian (Lévêque et al. 2008).

Globally, introduced fishes comprise an important impact on freshwater ecosystems. Environmental degradation, including the alteration of habitat, is often an important factor in interactions between native and introduced species. As a result, the quantity and quality of habitats important to native freshwater organisms are altered, and thus native species may be more vulnerable to competition and predation by introduced species that may be more tolerant of degraded systems.

1.1.3 Taxonomy of fishes

Fishes are a paraphyletic group: that is, any clade containing all fish also contains the tetrapods, which are not fish. For this reason, groups such as the "Class Pisces" seen in older reference works are no longer used in formal classifications. Traditional classification divide fish into three extant classes, and with extinct forms sometimes classified within the tree, sometimes as their own classes (Benton 1998).

Ichthyologists used to distinguish three major groups of freshwater fish according to their tolerance to saltwater and their hypothesized ability to disperse across marine barriers (Myers 1949): the primary division fish being strictly intolerant of salt water; the secondary division able occasionally to cross narrow sea barriers; and the peripheral division including representatives of predominantly marine families that have colonized inland waters from the sea.

The popular internet site FishBase (http://www.fishbase.org) adheres to a slightly different classification with fresh and brackish water fish species falling into three categories: (1) exclusively freshwater, (2) occurring in fresh and brackish waters, (3) or in fresh, brackish and marine waters. The first category covers more or less the primary and secondary divisions of Myers, while categories 2 and 3 cover the peripheral division (Froese and Pauly 2013).

Taxonomy of fishes has been studied since ages. Pliny first recorded 144 species at about 77 A.D. Since then new species have been routinely discovered and there have been a significant appreciation in number of species. One of the most current and most accurate counts of valid species is the Catalog of Fishes database (Eschmeyer 2014). It is the authoritative reference for taxonomic fish names, featuring a searchable on-line database, that has been maintained for about 25 years, and which tracks all new taxa and many revisional studies on an almost daily basis. The Catalog of Fishes covers more than 53,000 species and subspecies, over 10,000 genera and subgenera, and includes more than 16,000 bibliographic references.

1.1.4 Impediments in freshwater fish taxonomy

Taxonomic impediments have remained the biggest barrier in freshwater fish diversity studies as in case of other taxonomic groups. In recent times, taxonomy at the species level has tended to be neglected within ecological researches. The identification of organisms within communities to species level is one of the greatest constraints in terms of time and costs in ecological studies.

Some studies have suggested that working at a taxonomic level higher than species does not result in an important loss of information (Taxonomic sufficiency) (Giangrande 2003). Taxonomy has always been considered a marginal science even during the pioneer descriptive period of ecology, and traditionally has received little financial support. The result was the production of many misidentifications and erroneous records. Before the widespread use of digital taxonomy, the developing experimental ecological approach has led to an improvement in scientific methods, but concurrently to a reduction in the number of expert taxonomists for many groups. Further cases of loss of voucher specimen, independent identification in different geographical location and lack of uniform database over a long history of taxonomy has resulted in many conflicting taxonomy.

Before the widespread use of digital taxonomy, the developing experimental ecological approach has led to an improvement in scientific methods, but concurrently to a reduction in the number of expert taxonomists for many groups. Biodiversity, particularly ‘species richness’ has long been thought to influence temporal variability. Efforts to clarify the temporal variability relationship and to demonstrate the lack of such a relationship should continue. Such information is essential in order to maintain the ecological function despite the loss of component species, an important topic not only to ecologists but also to policy makers. Many species appear to have overlapping niches, and as such it could be argued that it is not essential for all species to be present. In contrast the crucial role of keystone species has been embraced in conservation biology as a tool to help highlight species requiring priority for protection (Berra 2001).

1.2 Indian freshwater fish diversity and problem in taxonomy

1.2.1 Fish biodiversity in India

India is one of the 12-mega biodiversity countries having two biodiversity hotspots, namely the Western Ghats and the Eastern Himalayas that are included amongst the top eight most important hotspots in the world. It also has rich freshwater (rivers, irrigation canals, tanks, lakes, reservoirs) fish diversity. This diversity is not only the wealth of India and the world but it also has serious implications on fishery. This diversity is being eroded each day mainly because of unending anthropogenic stress.

The country is endowed with vast and varied resources possessing river ecological heritage and rich biodiversity. Freshwater fishery sites are varied like 45,000 Km of rivers, 1,26,334 Km. of canals, ponds and tanks 2.36 million hectares and 2.05 million hectares of reservoirs (Ayappan S and Birdar S R 2004). The assessment of freshwater fishes is done mainly on the basis of 6 drainage systems in the country (shown in Figure 1.1). These are Indus river system, Upland cold-water bodies, Gangetic river system, Bramhaputra river system, East flowing river system and West flowing river system. India is rich in fishery resources and comprises of around 2508 fish species (Eschmeyer et al. 2012) of which 856 are freshwater inhabitants (Froese and Pauly 2013, Menon 1999). Indian freshwater fishes represent about 8.9% of the known fish species of the world and occupy the ninth position in terms of freshwater fish diversity (Lévêque et al. 2008). Distribution of freshwater fishes in India is widely varied with the north eastern zone alone claiming 40% of the reported species.

The Indian fish fauna is divided into two classes, viz., Chondrichthyes (cartilage fishes) and Osteichthyes (bony fishes). The endemic fish families form 2.21 per cent of the total bony fish families of the Indian region. 223 endemic fish species are found in India, representing 8.75 per cent of the total fish species known from the Indian region. The Western Ghats is the richest region in India with respect to endemic freshwater fishes.

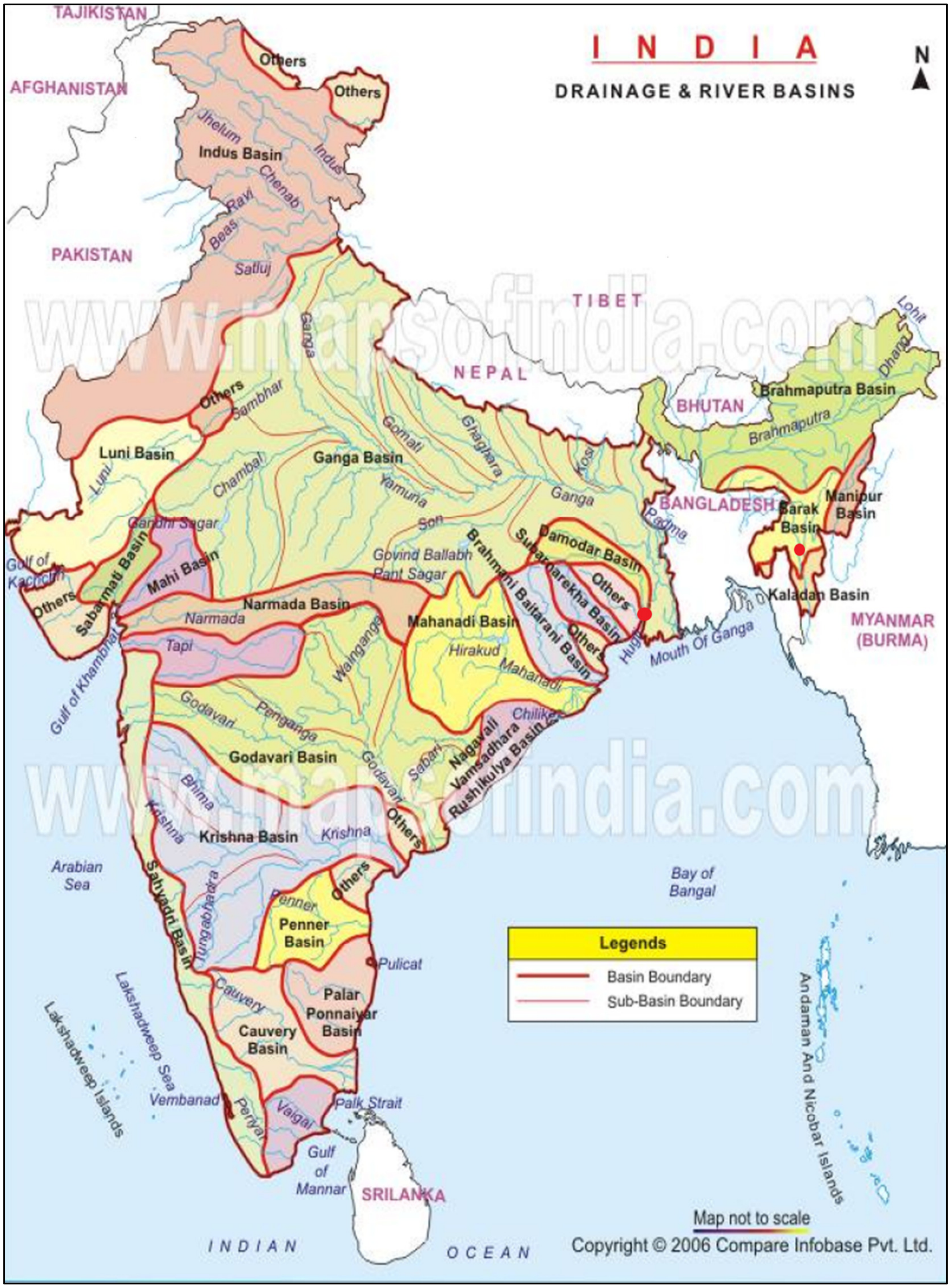

Figure 1.1 Map showing drainage and river system of India.

The red dots indicate the locations from where samples have been collected for this study. (The map have been downloaded from www.mapsofindia.com)

Northeastern India, which has a very high diversity among freshwater fish, does not have many endemic species within India because of its jagged political boundary. There are about 450 families of freshwater fishes globally. Roughly 40 are represented in India (warm freshwater species). About 25 of these families contain commercially important species. Number of endemic species in warm water is about 544.

1.2.2 Taxonomic impediments in Indian freshwater fishes

Freshwater fishes in India are a poorly studied group since information regarding distribution, population dynamics and threats is incomplete, and most of the information available is from a few well-studied locations only (CAMP). Threats to Indian freshwater fishes are physical in nature, such as habitat destruction, fragmentation, poisoning, pollution, pesticides, destructive fishing, and other kinds of human interference. Trade is an important contributing factor in threatening some freshwater fish taxa in India. This is mainly because of unsustainable harvest, poor scientific practices in fishing and an ever-growing demand. Moreover, the actual number of fish species found in India is still not accurately known because of taxonomic impediments (Hoagland 1996) arising due to lack of exploration, indiscernibility among some alike species, and ambiguity in taxonomic keys (Pethyagoda 1994). As a result, many species have been considered as cryptic some of which may also be latent (Darshan et al. 2010, Pethyagoda 1994). Furthermore, due to lack of proper morphological description with respect to sexual dimorphism, allopatric population, etc., the statuses of a few nominal species have been contentious (Darshan et al. 2010, Kottelat and Lim 1995, Ng and Hadiaty 2009, Roberts 1992, Vishwanath and Linthoingambi 2007). In India, in addition to the absence of an updated compiled checklist of freshwater fishes, the identification keys for many valid species have not been updated since, Talwar and Jhingran, 1991 (Jayaram 1999). During the last decade, many new species have been described in India, However due to lack of an updated compiled checklist or database, the newly discovered species are often overlooked in further assesment using taxonomic and molecular methods.

1.3 Molecular methods to resolve taxonomic impediments

1.3.1 Molecular vs. morphological methods

The development of molecular techniques has helped invigorate studies of fish systematics. The realm of methods developed for molecular systematics offer new suites of characters for analyzing relationship among fishes and have been applied effectively from population level to orders (Ferraris 1996). Morphological studies have been successful in defining species and organizing them into genera. Molecular approaches help to confirm these groupings and provide new insight into classification. Molecular studies also revealed some cryptic species and identified some incorrectly split groups (Rocha-Olivares et al. 1999, Stepien and Rosenblatt 1996, Stepien et al. 2001). Studies have proved good concordance between morphological and molecular approaches. Although morphological studies have been successful in defining genera, it is rare to find studies which present a hypothesis of relationship above the level of species comprising a genus, primarily due to lack of congruent characters. However, this is one of the strength of molecular data and inter and intrageneric relationships are now routinely tested and elucidated. Molecular data are also the primary means for generating phylogenetic relationship among species, populations and higher taxonomic orders. Studies at lower systematic levels are exploiting the mechanisms underlying the diversity of fishes (Kocher and Stepien 1997). Both morphological and molecular methods have problems in discerning higher taxonomic level relationship. In both types of data the main problem is in identifying homologous characters and finding synapomorphies to identify lineages with statistical confidence.

1.3.2 Mitochondrial DNA in species identification

The mitochondrial genome has many properties that made it useful for reconstructing recent phylogenetic history (Avise 1994, Simon 1994). Fish mitochondrial genomes are haploid and apparently nonrecombining. The evolution of the molecule therefore corresponds exactly to the model of bifurcating evolutionary trees. Moreover, mtDNA evolves more quickly than most nuclear genes, allowing the identification of informative phylogenetic characters among closely related species and populations. Two feature of mtDNA were typically listed as advantageous for phylogenetic analysis. First, mtDNA is maternally inherited. Second, substitutions in mtDNA do not accumulate according to strictly neutral process. Patterns of sequence differentiation suggested that selective sweeps may be common (Ballard and Kreitman 1994) and lab rotary experiments suggested competitive differences among mitochondrial haplotypes (Hutter and Rand 1995). Early studies of mtDNA analyzed restriction fragment length polymorphisms (RFLPs). Whole mtDNA were digested with specific endonucleases and the products were separated by using gel electrophoresis. Restriction sites were then mapped and their presence or absence was scored (Dowling 1996). Previously RFLP studies had been a popular approach in quantifying degree of divergence within and among populations. However, in applying this approach to species and higher systematic questions, the homology of restriction site characters became less certain. A better approach for these comparison involved direct analysis of DNA sequences. In coding regions, the variation in DNA sequences were evaluated among first, second and third codon positions and at the amino acid level in order to increase potential phylogenetic utility at higher systematic levels. The relative strength of the phylogenetic signal with the codon position and between the nucleotide and amino acid level were critically evaluated by (Naylor and Brown 1998).

1.3.3 The advent of DNA barcoding

In last decade, among the molecular methods used for taxon diagnosis, DNA barcoding has been extensively used for species identification as well as species discovery in various groups of organism (Hajibabaei 2012, Mendonca et al. 2012). The idea behind DNA barcodes was to find a single segment of DNA which is useful for identification of all living taxa. Hebert et al. (2003a) proposed the use of a 650 bp fragment of mitochondrial *COI* gene as sufficient for use in species identification of majority of animal taxa. The idea has gained wide popularity with over 1000 publications in less than a decade’s time.

As mentioned earlier, molecular-based identification is not a new concept. Many molecular identification systems have been developed. However, DNA barcoding has several advantages over previous methods. One advantage is its availability. The standard DNA barcode region, a fragment of *COI*, is very efficient for species identification. This region has good discrimination power for most animal groups. The universal primer, originally designed for marine invertebrates, can be applied to almost all animal phyla (Folmer et al. 1994, Hebert et al. 2003a, Hebert et al. 2004a, Hebert et al. 2004b, Hebert et al. 2003b). A 648-bp fragment has enough information and can be directly sequenced with a sequencer. The alignment process is not difficult because this is a protein-coding region. Errors can be detected by checking whether the obtained sequence is translatable. These useful features are the reason why the *COI* region was selected as the standard DNA barcode. Thus, DNA barcoding proved to be a simple but powerful method for non-experts, especially those who routinely identify a large number of samples.

Verifiability of identification of voucher specimens through relationships with taxonomy is another advantage of DNA barcoding. DNA barcoding is authorized by taxonomic experts who identify the voucher specimens from which DNA barcodes were obtained. A barcode record requires a species name, voucher specimen data (locality, date, depository of specimen, photographs etc.), a sequence, polymerase chain reaction (PCR) primers and trace files (sequencer's original outputs). CBOL and the National Center for Biotechnology Information (NCBI) have already proposed a standard format (keyword "BARCODE") for barcode sequences in GenBank (Consortium for the Barcode of Life 2005). Information on voucher specimens and trace files help to confirm whether the previous identification and sequence data are correct.

Effectiveness of DNA barcoding has now been validated for many groups of animals (Waugh 2007), both invertebrates and vertebrates (Hajibabaei et al. 2006a, Hajibabaei et al. 2006b, Hebert et al. 2004a, Hernandez-Davila et al. 2012), with fishes being one of the most extensively studied groups among them. (Becker et al. 2011, Ward 2012). Many successful nationwide studies on ichthyofaunal diversity have been undertaken using this method for both marine and freshwater fishes (de Oliveira Ribeiro et al. 2012, Mabragana et al. 2011, Wang et al. 2012). These studies have resolved several cases of crypticism and diagnosed many latent species (Carvalho et al. 2011, Pereira et al. 2011). Furthermore, these studies have also generated a large scale of barcode data that are available in BOLD (Ratnasingham S 2007) and NCBI.

The global success of DNA barcoding led the enthusiast to gather large scale DNA barcode data's from all over the world for every species that persist. This common campaign has been given the name "The Barcode of Life Project". In 2004, this project was formally initiated by the establishment of the Consortium for the Barcode of Life (CBOL). Ten years after DNA barcoding was initially suggested as a tool to identify species, millions of barcode sequences from more than 1100 species are available in public databases. While several studies have reviewed the methods and potential applications of DNA barcoding, most have focused on species identification and discovery, and relatively few have addressed applications of DNA barcoding data to ecology.

1.3.4 Fish DNA barcoding

The economic importance and identification challenges associated with fishes prompted the launch of an international Fish Barcoding of Life initiative (http://www.fishbol.org/) with the aim of barcoding all fishes. The Fish Barcode of Life Initiative (FISH-BOL) is a global effort to co-ordinate an assembly of a standardized reference sequence library for all fish species, one that is derived from voucher specimens with authoritative taxonomic identifications. The benefits of barcoding fishes include facilitating species identification for all potential users, including taxonomists; highlighting specimens that represent a range expansion of known species; flagging previously unrecognized species; and perhaps most importantly, enabling identifications where traditional methods are not applicable. The FISH-BOL is a wing of CBOL and at present stores DNA barcode records of nearly 30,000+ species (as of 04-03-2013) of marine, freshwater and estuarine fish of the world.

One important issue that needs to be more fully addressed in FISH-BOL concerns the initial misidentification of a small number of barcode reference specimens. This is unsurprising considering the large number of fish species, some of which are morphologically very similar and others yet unrecognized, but constant vigilance and ongoing attention by the FISH-BOL community is required to eliminate such errors. Once the reference library has been established, barcoding enables the identification of unknown fishes at any life history stage or from their fragmentary remains. The many uses of the FISH-BOL barcode library include detecting consumer fraud, aiding fisheries management, improving ecological analyses including food web syntheses, and assisting with taxonomic revisions (Ward 2012). The technique further bears application in monitoring fish products for health safety (Lowenstein et al. 2010) and in regulating the exploitation of fish species under aquarium trade (Wong et al. 2011).

Many sequences of Indian freshwater fishes have been submitted to the database and few studies have addressed problems specific to certain groups. Phylogenetic relationships among the commercially important Indian sciaenids was evaluated by Lakra et al. (2009). The taxonomic ambiguity, and systematic position of the Malabar snakehead, *Channa diplogramma*, was resolved by Benziger et al. (2011) using morphological and molecular genetic information, in addition to making an attempt to understand its phylogenetic relationships and evolutionary biogeography. A detailed study of Indian catfishes was undertaken by Bhattacharjee et al. 2012. Taxonomic discrepancy of Mahasheer fish was studied and resolved by Laskar et al. 2013.

However, a comprehensive assessment of DNA barcodes of Indian freshwater fishes has not yet been done through a similar study has been done for the Indian marine fishes (Lakra et al. 2011). Furthermore, in order to ensure that the reference barcode library is accurate, it is important to detect the ambiguities in the base collection of barcode at an early stage.

1.4 DNA barcode for higher taxon assignment

DNA barcode has proven to be an effective tool in species identification and resolving various taxonomic impediments. Hebert et al., while proposing barcoding of all animal life using *COI* gene (2003), advocated that diversity in the nucleotides and coded amino acid of the 5' section of this gene, was sufficient to reliably place species into higher taxonomic categories along with discriminating the closely allied species (Hebert et al. 2003a). They suggested that amino acid divergence can assign species to phylum and order level while nucleotide divergence can effectively discriminate among closely related species. Over the years, the efficacy of nucleotide divergence of *COI* gene in accurate species identification has been tested for various groups of animals and various degrees of success rates have been achieved (Dinca et al. 2011, Sass et al. 2007, Saunders 2005, Ward 2012). *COI* has been thus, gradually accepted as a standard barcode gene for discriminating animal life at the species level. This is evident from the explosive rise of *COI* gene sequence data in both global database (GenBank) and Barcode specific database (BOLD). As of 2013, the Nucleotide query 'barcode [keyword]' retrieves over 500 000 barcode sequences in GenBank, over 300 000 of which were added in the previous year (Benson et al. 2013).

However, efficacy of *COI* barcode, in assigning species to their higher taxa level viz. genus, family, order, is still dubious. Since the first proposal of DNA barcode, proponents of the barcode concept had positive view in the ability of *COI* gene in enabling higher taxon assignment, but the fact is yet to be standardized (Hebert and Gregory 2005). Some groups have proposed the use of nucleotides in higher taxon assignment while others have suggested the use of amino acid as a better alternative. Many methods of higher taxon assignment using DNA barcodes have been proposed over the years (Wilson et al. 2011). The conventional method of using Kimura's-two-parameter (K2P) divergence has been often used to assign higher taxa to the species. However limited rate of success have been achieved conforming the unsuitability of this method for higher taxon assignment (Ball et al. 2005, Ward et al. 2005). Alternatively, character based method have been employed to achieve the task but again achieving limited success rate (Bergmann et al. 2013, Rach et al. 2008).

Diagnostic attributes of DNA barcode is an outcome of two phenomena. First, flexibility acquired from various combinations of four alternate nucleotides at each position over an approximate of 648 sites. Secondly, constriction, governed by several constraints like codon bias, degenerate nature of genetic code and transition – transversion bias (DeSalle et al. 2005, Hebert et al. 2003a, Hebert et al. 2003b). These two phenomena cumulatively define unique characteristic of each group of closely related members. Species are basic unit of biological classification and success of barcode in species identification reveals presence of unique character at species level. Closely related species, sharing certain morphological features, have been classified into higher taxonomic groups as families and orders. Thus the short 5' segment of DNA barcode is expected to have the inherent signal to classify species to their higher taxon level. The underlying signal can be simply unraveled by accessing sequence variability and sequence information content across the taxa under study.

Schneider and Stephens devised a method of analyzing the degree of sequence conservation per site where the consensus sequence , the relative frequency of bases and the information content (measured in bits) at every position in a site or sequence were retrieved (Schneider and Stephens 1990). Both significant residues and subtle sequence patterns of nucleotide and amino acid sequence are derived using this method. Ward et.al (Ward and Holmes 2007) used this method to estimate sequence conservation in *COI* gene of fishes and unraveled the role of third codon variation in species discrimination power of *COI* barcode.

.

1.5 Factors effecting efficacy of DNA barcodes

1.5.1 Reference barcode library

The most important factor affecting the accuracy of species identification is the coverage and reliability of available barcode libraries (Ekrem et al. 2007). Barcode based identification will fail if the DNA barcode data of the species in question has not been registered to a library. In fact, most identification errors are caused by a lack of reference data (Virgilio et al. 2010). In addition, intraspecific variation might be underestimated when the samples included in the library do not reflect the overall genetic diversity and/or do not include all clades of non-monophyletic species groups, and interspecific variation might be overestimated if data on closely related species are unavailable. Wiemers and Fiedler (2007) reported that the barcoding gap in Lycaenidae (Lepidoptera) is an artifact caused by insufficient sampling across taxa. It should be emphasized that the misidentification of reference barcode data is another serious problem. Many records from misidentified samples have been submitted to GenBank (Taylor and Harris 2012, Virgilio et al. 2012).

Meier et al. (2008) reported that misidentified barcode data are submitted to the BOLD database, which does not have a mechanism for verifying records. The DNA barcodes obtained from misidentified specimens are detected by comparison with multiple barcodes of the species. Then, misidentifications can be corrected by re-identification of voucher specimens by taxonomic experts. Thus, quality control in collaboration with taxonomists is required for the proper construction of reference DNA barcode libraries. Costa et al. (2012) proposed a ranking system to attribute a confidence level to species identifications associated with DNA barcode records from reference libraries of DNA barcodes. They tested the ranking system on a newly generated reference library of DNA barcodes for marine fish of Portugal.

1.5.2 Sample condition

An important limitation of DNA barcoding is sample condition. Various traditional methods of preserving samples are often not suitable for extraction of DNA at a later point of time. The DNA of dried, pinned specimens, the most popular method of insect preservation, is degraded by heat, oxidation and fumigation gas (Zimmermann et al. 2008). Similarly DNA of samples preserved in chloroform also undergoes degradation. Thus, DNA barcoding has mainly been used only on fresh samples or specimens preserved in an ideal manner for molecular work (refrigerated or stored in ethanol or acetone).

In general the amplification of DNA fragments becomes extremely difficult for specimens that have been preserved for more than 50 years. Strange et al. (2009) showed that molecular markers work well in Bombus specimens up to 101 years old, although the amplification rate is significantly lower in materials that are more than 60 years old. Surprisingly, Thomsen et al. (2010) obtained DNA from fossilized Coleoptera preserved in permafrost for more than 10 000 years, even though only a short fragment of DNA was amplified by PCR. Other technical advances such as efficient DNA extraction methods, the discovery of high-efficiency DNA polymerase, reagents that decrease the effect of impurities that inhibit PCR and DNA-repairing enzyme have helped to improvise extraction of DNA from old and fossilized specimens.

However, recent methodological and technical advances allow the extraction of archival or ancestral DNA from historical museum specimens or fossilized samples. The extraction and amplification of this DNA has become one of the hottest trends in molecular ecology, evolutionary biology, paleobiology and anthropology, and many different methods have been used for animals, plants and fungi (Huijsmans et al. 2010, Mason et al. 2011, Perry et al. 2014, Petralia et al. 2013, Samarakoon et al. 2013, Shokralla et al. 2014).

(Paijmans et al. 2013, Särkinen et al. 2012, Schweitzer and Marshall 2012, Seguin-Orlando et al. 2013) have made it more feasible to amplify DNA from historical and fossilized specimens. DNA amplification from ancient specimens may also depend on the length of the amplified fragments. Fragments that are shorter than 200 bp are relatively well amplified even from old specimens, whereas longer ones are not. Indeed, most attempts to amplify such DNA have adopted primer sets for 20–200-bp fragments (Hajibabaei et al. 2006c). This low amplification success rate for longer fragments may be caused by fragmentation of DNA within the specimens. Two strategies have been proposed for addressing this problem. The first is to identify species based only on short fragments that are easily amplified. Several authors (Fan et al. 2009, Hajibabaei et al. 2006c, Meusnier et al. 2008) tested this process and showed that short barcodes are effective for species identification when the taxonomic group of the sample is preliminarily confined. This strategy is especially effective when DNA barcoding is used to identify historical samples by comparing them to a reference barcode library. The second strategy is to obtain a full length DNA barcode by connecting the short fragments. Van Houdt et al. (2010) demonstrated such a method by amplifying 269–363-bp fragments within the barcode region using newly developed universal primers and then connecting these fragments using a complete barcode guide sequence obtained from a fresh sample of the same species (or congeneric species) using the Bayesian algorithm. Although much time and effort is required, this strategy makes it possible to obtain full length barcodes from archival specimens such as type specimens. This progress in the barcoding of old specimen increases the value of museum collections as a source of genetic diversity information that is relevant to ecology evolutionary biology, population genetics and conservation biology (Wandeler et al. 2007).

Most primers used for DNA barcoding are universal and it is possible to amplify DNA from a wide range of organisms. This raises the risk of contaminating archival DNA with contemporary DNA. Thus, archival or ancestral DNA barcoding have been suggested to be conducted under very specific conditions including at least two repetitions of PCR amplification and the elimination of contemporary DNA from the laboratory.

1.5.3 Algorithms for DNA barcode based identification

The development of algorithms for DNA barcode based identification is a challenge in the field of bioinformatics. In the identification engine of the BOLD system, sequences similar to a query are collected from the reference library by a linear search (Ratnasingham S 2007). The result is also available as a cladogram based on the neighbor-joining (NJ) method. In the tree-based approach, a query sequence is assigned to a species when the query is included in a cluster consisting entirely or even partially of conspecifics (Hebert et al. 2003a).

There is controversy about the accuracy of tree-based approaches, such as the NJ method, for DNA barcoding-based identification. Meier et al. (2006) introduced distance-based criteria, in which a query sequence is assigned to a species of the best-matched barcode regardless of similarity (best match method) or when the degree of difference between the query and the best-matched barcode is less than 95% for all intraspecific distances.Virgilio et al. (2010) compared the performance of DNA barcoding-based identification among insect orders and these two criteria, and concluded that the distance-based criterion showed higher and more robust performance than the tree-based one.

The distance based method; uses a comparative analysis based on distance based clustering of the studied species and explore the relative affinity of a specimen to the available barcodes. Second, the character based method in which, the presence of a set of nucleotide characters at a particular set of positions, assigns a sequence to a species (Bergmann et al. 2009). Thus yielding a binary result, the sequence will show either a perfect match or none. In contrast, the more commonly used distance based method of DNA barcoding uses a more subjective approach. The decisive approach of the character-based barcoding can be used to develop short tags, which will have practical application in the design of the most lucrative proposition of the barcoding project: the handheld barcoder- a machine that matches the barcode of a specimen and gives an exact result (Bergmann et al. 2013).

This criterion may provide more accurate results than distance based approaches in which all variation is reduced to a single vector, even for subspecies and populations that show very little variation (Lowenstein et al. 2010, Rach et al. 2008). The accuracy of the character based approach tends to be low without a comprehensive library of species or species complexes (Sass et al. 2007). Many algorithms based on different approaches have been proposed and their performances have been compared (Fišer Pečnikar and Buzan 2014, Frezal and Leblois 2008).

The Character based system uses diagnostic nucleotide traits, which comprises of two associated parameters to define the barcode, (i) the character of the nucleotide and (ii) its relative position in the full-length barcode. The use of a two-parameter array of data amplifies the memory requirement for storing the data in large datasets. Moreover, in customary character-based system, as attributes are located in various positions in the full-length barcode, the entire length of the *COI* barcode has to be sequenced. Further, the diagnostic positions depend on the study group. Thus, the set of diagnostic positions may vary between study groups having different combination of species. This results in a lack of uniform character keys that can be used as a standard for species diagnosis. To solve the problem, a short barcode composed of continuous nucleotide positions in a standard known fragment can be used. The short continuous character array will require a lesser storage space and will be advantageous for cases having problem in recovering long DNA fragment.

One of the key problems in using longer barcode segment has been the difficulty in recovering PCR fragments, longer than 200bp in case of archival specimen and processed biological material (Goldstein and Desalle 2003, Wandeler et al. 2007). Some studies have proposed the concept of ‘minibarcode’ where a smaller fragment has been found to be competent with the full length DNA barcode in delimiting closely related species (Hajibabaei and McKenna 2012, JK et al. 2010, Lijtmaer et al. 2011, Meusnier et al. 2008). Moreover, a shorter fragment has the advantage of being cost effective and can potentially pave way for the design of a DNA barcode probe (Xu et al. 2009). Meusnier et al. have found that *COI* sequences of about 100 and 250 base pairs are sufficient to solve the purpose of species delimitation (Meusnier et al. 2008). A major problem in determining a standard mini barcode region is selecting a subset of relevant features within the large 650 bp of *COI* barcode. Thus, the key is to identify the most informative region within the existing barcode.

The 650bp of conventional barcode includes many uninformative and less informative sites. The vast majority of nucleotide substitutions within the *COI* fragment occur at the third codon position, which might lead to rapid saturation (Lin and Danforth 2004). Transition and transversion substitutions cumulatively define the nucleotide variations in protein coding gene (Kumar 1996, Zhao et al. 2012). Transition bias appears more pronounced in animal mitochondrial gene than the nuclear gene (Brown et al. 1982, Hasegawa and Kishino 1989). However, with time, most of the transitions are reverted to the original nucleotides while only 50% of the transversions have chances of reversal (Salemi et al. 2009).

Further, among single-step substitutions in the universal genetic code at third codon positions, only about 3% of transitions cause amino acid replacements, as compared to 41% in case of transversions (Wakeley 1996). This indicates that, transversion rich sites are more informative, carrying sufficient information to delimit a species from its close relatives. In a recent work, a 119 bp segment of *COI* gene, rich in transversion, was shown to be capable of delimiting species from 33 families of catfishes (Bhattacharjee and Ghosh 2014).

1.6 Prospect of the study

Indian Freshwater fish taxonomy is far from being complete. Taxonomic status of many species remains to be confirmed and discovery of new species still continues. There are various contentions regarding assignment of individuals to various taxonomic ranks. Lack of exploration, proper identification keys and lack of compiled up-to-date checklist of Indian freshwater fishes further enhance the problem. Thus there is an urgent need for proper inventorisation and documentation of this fish diversity in order to develop a fresh water fish diversity information system having both bioinformatics and georeferenced databases of fish and fish habitat. Therefore, for legible characterization of Indian freshwater fishes, there is an urgent need of species scrutiny using advanced molecular methods. Though, DNA barcoding project have been initiated in India there are some major issues that needs immediate attention as the reference library of barcodes remain incomplete. Moreover there is a lack of comprehensive assessment of DNA barcodes of Indian freshwater fishes

The conventional method of DNA barcoding used popularly, promises to provide resolution only at species level. Some studies have tried to look further and implement DNA barcoding for higher taxa but seldom encountered positive results with the conventional methods. Moreover, since fishes have paraphyletic origin any tree based method cannot ensure monophyly at higher taxa level thus making identification difficult. Character based method have the potential to collectively identify members of any taxa level as it is based on the fundamental concept that members of a given taxonomic group share attributes that are absent from comparable groups. Thus, it is more suitable for field based application of DNA barcoding. However the character profiles developed from random sites within full length barcodes have the inherent drawbacks like sequencing of full length barcode, use of sparse data matrix and lack of a uniform diagnostic position for each studied group. To resolve the problem, a short continuous stretch of fragment can be used for character based species identification by creating barcode motifs.

Through this study, first an attempt is made to cover a comprehensive assessment of both recorded and barcoded data of Indian freshwater fishes. The study will help to highlight the fish groups which need attention. The assessment of entire freshwater fish barcode data available so far in the database will aid in quality assurance of the barcode reference database. This is essential because in near future this reference database will act as a guide in species identification. The study focuses on evaluation of genetic diversity and shared nucleotide traits of DNA barcodes of Indian freshwater fishes which will further aid in resolving taxonomic discrepancies underlying in this group. A detailed compositional analysis is undertaken to understand the underlying potential of *COI* barcode which can aid in delimiting higher taxa level. Both character based and distance based methods are assessed to achieve higher taxonomic rank assignment. The efficiency of both the methods is compared and attempts are made to improvise the existing methods of DNA barcoding.

1.7 Objective of the study

1) Assessment of the status of DNA barcoding of freshwater fishes in India and enrichment of the global database with DNA barcodes of Indian freshwater fishes.
2) Compositional analysis of *COI* DNA Barcodes of Indian freshwater fishes.
3) DNA barcode sequence based species level discrimination and other hierarchical level taxon assignment of Indian freshwater fishes using conventional distance based method.
4) Development of character profiles for different hierarchical levels of a taxon.