Chapter 2. REVIEW OF LITERATURE

2.1 Cataloging and classification of global fish species

Studies related to classification and cataloging of fishes have initiated long back in history. Various people in history had an appreciation of species numbers, from about 144 species as estimated by Pliny about 77 A.D. Günther (1864) created a catalogue of the Physostomi, containing the Families Siluridae, Characinidae, Haplochitonidae, Sternoptychidae, Scopelidae, Stomiatidae in the Collection of the British Museum. Bailey (1960) gave 15,000 to 17,000 species; Marshall (1965) mentioned an approximate of 20,000 species; Norman (1963) estimated an approximate of 25,000 species. However, for most of this study no mentions of method of estimation were made.

In 1970, Cohen presented an estimation of the total number of fish species in the world and in each of eight ecological groupings. He found that an "astonishingly high percentage of bony fishes live in freshwater habitats. According to Cohen's analysis, 41.2% (8,275 species) of all species live in freshwater, which included both primary and secondary freshwater fishes. He indicated that this high percentage might be a reflection of the degree of isolation possible in freshwater environments and referred to the great variety of habitats and ecological niches in freshwater and along tropical shores. Joseph S. Nelson created a standard reference for fish systematics in the four editions of his book, Fishes of the World. In various editions of his book, the number of species recognized as valid has increased quite dramatically Nelson (1976), 18,818 in 450 families; Nelson (1984), 21723 species in 445 families and Nelson (1994), 24618 species in 482 families. In the latest edition, he recognized 32,500 species belonging to 515 families (Nelson 2006). Berra (1997) found that the families Cichlidae, Cyprinidae, Characidae, Loricariidae and Cyprinodontidae constituted half of all described species between 1978-1993. The most new freshwater fishes came from South America (39%), Africa (32%), Asia (17%). Lundberg et al. (2000) estimated that over 10 000 fish species

live in fresh water; which constitutes, approximately 40% of global fish diversity and one quarter of global vertebrate diversity. Berra (2001) found that families receiving the most new species were the cyprinids, the gobiids, characins and cichlids. They also reported that of the newly named freshwater species, 35% were from south America, 30% from Africa, and 23% from Asia. These matched with Gilbert (1976) data.

Many groups of fishes are expanding with newly described species, whereas a few are decreasing because species are being synonymized faster than new ones are described. However, a net increase in species of fish is shown every year. In the Catalog of Fishes database, 2013, there were 428 new species added. So far in 2014, there have been 158 new species described which totals the number of valid species of fishes to about 33,167 (Eschmeyer 2014). In Fishbase 52 valid species were originally described in the year 2014. 32726, species were entered before 1st January 2014 and 121 species were entered from 1st January 2014 till 4th July 2014. Moreover, 30759, valid names were modified before 1st January 2014.

At the level of the biogeographic realms and taking into account only fully freshwater fish families (i.e. the primary and secondary divisions), the largest number of families by far (43) was found in the Neotropical region, with a high proportion of endemic families (33% or 77%) mainly belonging to the orders Characiformes and Siluriformes. This was followed by Oriental region (33 families, 15 endemic) and the Afrotropical region (32 families, 17 endemic). For strictly freshwater fishes, at the generic and species levels in the different biogeographic realms the overall pattern was quite similar to that at the family level with 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. When taking into account the fresh and brackish-water fishes, the figures are, respectively, 4,231 species (769 genera) in the Neotropical region, 3,272 (542 genera) in the Afro-tropical, 2,948 (609 genera) in the Oriental, 2,381 (551 genera) in the Palaearctic, 1,741 (402 genera) in the Nearctic, and 580 (1,232 genera) in the Australian (Lévêque et al. 2008).

2.2 Cataloging and classification of Indian freshwater fishes

Francis Hamilton (1822) presented the first comprehensive work on the fishes of the river Ganges and its tributaries which describes 269 species, with 97 figures of fishes arranged in 39 plates. In his work 15 new genera of fishes were described. Francis Day was a pioneer ichthyologist who first described 343 marine and freshwater fishes of British India. Day wrote a monograph, illustrated by George Henry Ford that was published between 1875–1878, with a supplement in 1888. Day published two volumes on "Fishes" in The Fauna of British India, Including Ceylon and Burma series in which he described over 1400 species (Day 1889a, b).

The work of Francis Day was reassessed by Whitehead and Talwar and published in (1976). Shaw and Shebbeare (1937) presented a detailed account on the fishes of North Bengal. Dey (1973) studied the distribution and taxonomy of the ichthyofauna of the hill streams of Kamrup-Khasi-Garo regions of Assam. (Menon 1974) created the first checklist of the fishes of the Himalayan and the Indo-Gangetic plains. Jayaram (1981) authored the first comprehensive guide for the freshwater fishes of India, Pakistan, Bangladesh, Burma, and Sri Lanka. Talwar and Jhingran (1991) provided a detailed account of freshwater fishes of Indian subcontinent, which is considered to be the standard catalogue of Indian freshwater fishes till date. Menon's Checklist of Fishes of the Himalayas and the Indo-Gangetic Plains is an example where the distribution of the torrential stream fishes along the Himalayas is explained in terms of the palaeogeography of the region (Menon 1974).

The importance of zoogeographical studies in the advancement of our knowledge of the phylogeny, especially of widely distributed groups of animals, has been brought out clearly in the ichthyological contributions, viz. 'Monograph of the Cyprinid fishes of the genus Garra hamilton' and 'A systematic monograph of the tongue soles of the genus Cynoglossus Hamilton Buchanan (Pisces: Cynoglossidae)' (Menon 1964, 1977). Menon was in fact the first Indian to contribute data to the FAO fish identification sheets. The checklist of Menon (1999) listed 446 primary freshwater species under 33 families and 11 orders from the Indian region alone.

A Conservation Assessment and Management Plan (C.A.M.P.) Workshop was conducted in 1997 for 329 taxa of freshwater fishes of India, to assess their status in the wild. Approximately half of all Indian freshwater fishes were assessed at the workshop, 329 species. According to the assessment of the workshop, a total of 227 Indian freshwater fishes are threatened (CAMP).

Jayaram (1999) listed 852 freshwater species of fishes under 272 genera, 71 families and 16 orders, including both primary and secondary freshwater fishes from India, Bangladesh, Myanmar, Nepal, Pakistan and Sri Lanka.

Specimens representing 160 nominal species of fishes that were named by Francis Day were among the nearly 2000 specimens sent to the Australian Museum by Day in 1884. The type status of each of these specimens was evaluated in light of new evidence obtained from the archival papers of Edward Ramsay, the curator responsible for the acquisition of the Day collection (Ferraris 2000). Rema Devi and Indra (2009) created the latest Checklist of species updated after a span of 10 years which included several new species mostly siluroids and a few cyprinoids and those resurrected from synonyms. This list excluded exotic and secondary freshwater species.

During the last decade, many new species have been described in India, particularly from the north-eastern region (Dishma and Vishwanath 2012, Nath and Dey 1989, Selim and Vishwanath 2002, Vishwanath and Linthoingambi 2005) and Western Ghats (Devi et al. 2010, Pethyagoda 1994).

Kottelat (2013) estimated 3108 valid and named native fish species in the inland waters of Southeast Asia between the Irrawaddy and Red River drainages, the small coastal drainages between the Red River and Hainan, the whole Indochinese Peninsula, Andaman and Nicobar Islands, Indonesia (excluding Papua Province, Waigeo, Aru), and the Philippines. The species belonged to 707 genera and 137 families. In addition, he confirmed the presence of 49 introduced and established species in this region.

2.3 Molecular approaches and use of DNA barcoding in fish systematics

2.3.1 Early molecular studies

Species identification by molecular analysis has been used for many years. Initially, allozyme differences were used (Avise 1975), followed by mtDNA examination (Avise 1994). Genomic approaches to taxon diagnosis exploit diversity among DNA sequences to identify organisms (Kurtzman 1994, Wilson 1995). A bacterial identification system using SrRNA sequences was proved to be effective (Busse et al. 1996). DNA being less sensitive to degradation has a huge advantage over their protein-based counterpart methods (Hanner 2005). Moreover DNA can be accessed in all stages from egg to adult. Furthermore, synonymous mutations can be recognized in sequencing approaches, and polymerase chain reaction (PCR) amplification protocols make it possible to analyses minute amounts of tissue. Perhaps most importantly, DNA sequence data are easier to replicate and interpret across laboratories.

An increasing sophisticated realm of molecular techniques has been developed since the mid-1970's to study the molecular similarities of organism. These methods were preceded by protein sequencing and immunology, the widespread use of molecular techniques in fish systematics began with the discovery of allozyme polymorphisms. Allozyme and isozyme studies have been one of the most popular approaches in examining population genetic and stock divergence questions in fishes. They have also been useful in identifying cryptic species and in testing bio-geographic hypothesis. However the main disadvantage of using allozyme approaches is that bands (alleles) that have same electric charge and migrate to the same point in the gel may not be homologous. The scoring of gel is often subjective and bands are difficult to interpret when weak or close together (Kocher and Stepien 1997). Variants have traditionally been assumed to be selectively neutral. However, several studies have shown that allozyme variants are not neutral markers and are under selection (Avise 1994, Pogson et al. 1995, Powers and Schulte 1996).

2.3.2 Mitochondrial DNA in taxonomy studies

Mitochondrial DNA regions have been well studied in fishes and knowledge of universal primer sequences (Kocher et al. 1989, Meyer et al. 1990, Palumbi 1996, Simon 1994) for amplification by PCR and sequencing has made them very accessible. This has been used to address many different levels of taxonomic queries, depending on the region sequenced and the use of various correction factors for types and positions of substitutions. Silent sites of mitochondrial protein coding genes and non-transcribed control region are shown to be particularly useful for analyzing relationships of recently diverged taxa such as among populations, species and genera. At higher taxonomic levels, more slowly evolving regions such as the 16S and 12S ribosomal regions were found to be more effective. The sequence evolution in mtDNA has been relatively well studied in fishes. Base substitution events occur relatively rapidly. MtDNA structure, gene order and secondary structure are largely conserved in fishes, as well as in other vertebrates. Until early 2000, cytochrome b was probable the best studied mitochondrial gene in fishes (Block et al. 1993, Carr and Marshall 1991, Kocher and Stepien 1997, Meyer et al. 1990, Zhu et al. 1994). Although it has been widely used, some have questioned the ability of this sequence to resolve phylogenies (Graybeal 1993, 1994, Meyer 1994).

Mitochondrial ribosomal genes were often used to study more distantly related taxa. Substitutions in the small subunit (12S) accumulate relatively slowly, approximating the average for the entire mitochondrial genome. However, those in the large subunit (16S) evolve even more slowly (Simon 1994). The 12s rDNA gene was used by Stepien and Kocher (1997) to examine relationships among species, genera, tribe, families and suborders of beloniform fishes, showing strong utility at these different levels and congruence with morphological based hypothesis.

The mitochondrial control region is involved in the control of mtDNA replication and RNA transcription. It is also called the displacement loop (D-loop) because one of the two strands of the helix is displaced by the synthesis of a new strand during replication. In fishes the control region is usually long, (888 to 1223bp in percids) and often

contains tandemly repeated segments. There is a set of conserved sequence blocks that are probably involved in controlling mtDNA replication and transcription, which may be useful for some systematic studies (Attardi 1985, Lee et al. 1995). The highly variable region has thus been a popular sequence for examining population structure and relationships among closely related species of fishes (Arnason and Rand 1992, Meyer et al. 1990). Although some areas of this rapidly evolving sequence were alignable even among distantly related fishes (Lee et al. 1995), the high rate of evolution of this sequence appeared to preclude analyses beyond the level of closely related species and genera.

2.3.3 Introduction of DNA barcoding

Hebert et al. (2003a) from the University of Guelph, Ontario, Canada, first established that the mitochondrial gene cytochrome c oxidase subunit 1 (*COI)* can serve as the core of a global bio identification system for animals by discriminating 200 closely allied species of Lepidoterians and proposed the compilation of public library of *COI* gene that would be linked to named specimens. In their following work, they suggested that the DNA-based identification system, founded on *COI* can aid the resolution of this diversity. While the previous work validated the ability of *COI* sequences to diagnose species in certain taxonomic groups, the next study extended these analyses across the animal kingdom. The results indicated that sequence divergences at *COI* regularly enabled the discrimination of closely allied species in all animal phyla except the Cnidaria. This success in species diagnosis reflected both the high rates of sequence change at *COI* in most animal groups and constraints on intraspecific mitochondrial DNA divergence (Hebert et al. 2003b). Hebert et al. (2004b) sequenced *COI* sequences of 260 of the 667 bird species that breed in North America. They found that every single one of the 260 species had a different *COI* sequence. The sequences were either identical or were most similar to sequences of the same species. Hebert and his collegues thus reffered to *COI* as potential barcode for species identification of animal kingdom.

The term “DNA barcode” was first coined by Arnot et al. (1993) in their paper describing the possibility of discriminating isolates of *Plasmodium falciparum* on the basis of a circumsporozoite gene.

The popularity of DNA barcoding was due to its inherent advantages over several previous methods. One advantage is its universality. The standard DNA barcode region, a fragment of *COI*, is very efficient for species identification and has good discrimination power for most animal groups. The universal primer, originally designed for marine invertebrates, proved effective for many animal phyla (Folmer et al. 1994, Hebert et al. 2004a, Hebert et al. 2003b). Thereafter, the Barcode of Life project was proposed to promote DNA barcoding as a global standard for sequence-based identification of eukaryotes. In 2004, this project was formally initiated by the establishment of the Consortium for the Barcode of Life (CBOL), which aims to develop a standard protocol for DNA barcoding and to construct a comprehensive DNA barcode library (Stoeckle 2004).

A major study evaluating the efficacy of DNA barcoding was focused on the Neotropical skipper butterfly, *Astraptes fulgerator* at the Area Conservacion de Guanacaste (ACG) in north-western Costa Rica. This species was already known as a cryptic species complex. The *COI* gene of 484 specimens was sequenced from the ACG. The studies finally concluded that *Astraptes fulgerator* consisted of 10 different species in north-western Costa Rica (Hebert et al. 2004a).

Lambert et al. (2005) examined the possibility of using DNA barcoding to assess the past diversity of the earth's biota. The *COI* gene of a group of extinct ratite birds the moa, were sequenced using 26 subfossil moa bones. As with Hebert's results, each species sequenced had a unique barcode and intraspecific *COI* sequence variance ranged from 0 to 1.24

In 2006, Smith *et al.* examined whether a *COI* DNA barcode could function as a tool for identification and discovery for the 20 morphospecies of *Belvosia* parasitoid flies (Tachinidae). Barcoding not only discriminated among all 17 highly host-specific

morphospecies of *Belvosia*, but it also suggested that the species count could be as high as 32 by indicating that each of the three generalist species might actually be arrays of highly host-specific cryptic species.

Min and Hickey (2007) showed that short sequences of *COI* DNA barcode can yield important, and surprisingly accurate, information about the composition of the entire genome. Thus, for unsequenced genomes, the DNA barcodes can provide a quick preview of the whole genome composition. Dawnay and Ogden (2007) introduced the application of DNA barcodes to forensics for wildlife crime investigation which previously routinely involved genetic species identification based on DNA sequence similarity. They assessed the *COI* gene for use in forensic analysis following published human validation guidelines. Borisenko et al. (2009) however, proposed efficient logistics of pre-laboratory specimen processing and seamless interfacing with molecular protocols for building a global library of DNA barcodes. The first description of a new species using a DNA barcode from the holotype was by Brown et al. (2003), who used this method to describe a new species of Xenothictis (Lepidoptera: Tortricidae). Since then, many new species have been described with DNA barcodes from the holotype or paratypes, not only in arthropods, but also in other animals e.g. (Adamski et al. 2009, Badek et al. 2008, Burns et al. 2007, Dabert et al. 2008b, Dabert J 2008, Yassin et al. 2008, Yoshitake et al. 2008).

2.3.4 Launch of iBOL and next phase of DNA barcoding

By then, the Barcode of Life project entered a new phase with the launch of the International Barcode of Life project (iBOL) (Stockle and Hebert 2008). The iBOL is a huge international collaboration of 26 countries that aims to establish an automated identification system based on a DNA barcode library of all eukaryotes. Costa and Carvalho (2010) explained the prospects of barcode of life initiative (BOLI) where they mentioned- “while genome projects yield spectacular insights into molecular evolution, they have targeted only a few species. In contrast, the Barcode of Life Initiative (BOLI)

proposes a horizontal approach to genomics, examining short, standardized genome segments across the sweep of eukaryotic life, all 10 million species".

Meiklejohn et al. (2011) brought forward a wonderful practical application of DNA barcoding by sequencing of a 658-bp 'barcode' fragment of *COI* gene of forensically important Sarcophagidae (Diptera) from 85 specimens, representing 16 Australian species from varying populations. All species were resolved as reciprocally monophyletic, except Sarcophaga dux. The *COI* 'barcode' sequence was found to be suitable for the molecular identification of the studied Australian Sarcophagidae. McGowin et al. (2011) focused on the importance of *COI* DNA barcode in identification of two marine turtle leeches (Ozobranchus margoi and Ozobranchus branchiatus. Using morphological taxonomy combined with distance- and character-based genetic sequence analyses, this study has established a DNA barcode for both species of Ozobranchus spp. leech and has shown that it can be applied successfully to the identification of leeches at earlier stages of development when morphological taxonomy cannot be employed.

Bennett et al. (2011) carried out DNA barcoding of an invasive mammal species, the small Indian mongoose (*Herpestes javanicus*) in the Caribbean and Hawaiian Islands. The work demonstrates the utility of using DNA barcoding approaches with mtDNA cytochrome b to discriminate between the two species and other sympatric members of the genus Herpestes (*Herpestes naso, Herpestes urva, and Herpestes edwardsii*). Bitanyi et al. (2011) evaluated the effectiveness of *COI* barcode in species identification of Tanzanian antelope species. A 470 base-pair region of the *COI* gene was examined in 95 specimens representing 20 species of antelopes, buffalo and domestic Bovidae. All the Tanzanian species showed unique clades. Further they demonstrated that even short *COI* fragments can efficiently identify antelope species.

Bucklin et al. (2011) reviewed the role of DNA barcoding in the study of marine metazoa, with concern that more than 230,000 known species representing 31 metazoan phyla populate the world's oceans and perhaps another 1,000,000 or more species remain to be discovered. The campaign of barcoding increased with the

involvement of more and more organisms (Bell et al. 2011, Clare et al. 2011, Janzen et al. 2011) to reveal the actual level of diversity among them.
Vargas et al. (2012) developed a DNA-barcoding workflow capable of processing potentially large sponge collections and is routinely used for the Sponge Barcoding Project with success. Sponge specific problems such as the frequent coamplification of non-target organisms were detected and potential solutions are currently under development. The initial success of this innovative project demonstrated considerable refinement of sponge systematics.

Nagy (2012) presented the first comprehensive study targeting the entire reptile fauna of the fourth-largest island in the world, the biodiversity hotspot of Madagascar. Compared with available multi-gene phylogenies, DNA barcoding correctly assigned most samples to species, genus and family with high confidence and the analysis of fewer taxa resulted in an increased number of well supported lineages. Taylor and Harris (2012) reviewed the DNA barcoding enterprise, its continued resistance to improvement and the implications of this on the future of the discipline. They anticipated the consistent failure of DNA barcoding to recognize its limitations and evolve its methodologies, reducing the usefulness of the data produced by the movement and throwing into doubt its ability to embrace NGS.

Many large-scale barcoding projects have been initiated in 2013 with many of them exploring new avenues for application of DNA barcodes (Alcántar-Escalera et al. 2013, Brodin et al. 2013, Di Pinto et al. 2013, Galimberti et al. 2013, Hebert et al. 2013, Keskin and Atar 2013, Maas et al. 2013, Pino-Bodas et al. 2013). Shen et al. (2013) used DNA barcoding to detect erroneous sequences in GenBank by evaluating deep intraspecific and shallow interspecific divergences to discover possible taxonomic problems and other sources of error. Galimberti et al. (2013) explored the effectiveness of DNA barcoding in food traceability, and to delineate some best practices in the application of DNA barcoding throughout the industrial pipeline. Porco et al. (2013) surveyed the occurrence and genetic structure of two major groups of soil invertebrates in both their native and introduced ranges: Collembola and earthworms. Their study

established that invasive species surveys employing DNA barcoding gain additional resolution over those based on morphology as they allow evaluation of cryptic lineages exhibiting different invasion histories. Newmaster et al. (2013) investigated herbal product integrity and authenticity with the goal of protecting consumers from health risks associated with product substitution and contamination. Product substitution occurred in 30/44 of the products tested and only 2/12 companies had products without any substitution, contamination or fillers. Some of the contaminants were found to pose serious health risks to consumers. In 2013 a major initiative was announced for formation of Cold Code, the international initiative to DNA barcode all species of the 'cold-blooded' vertebrates (Murphy et al. 2013). Crawford et al. (2013) used a DNA barcoding approach to survey mtDNA variation in captive populations of 10 species of Neotropical amphibians maintained in an ex situ assurance programme at El Valle Amphibian Conservation Center (EVACC) in the Republic of Panama.

Even after Ten years of initiation of DNA barcoding and mass scale barcoding of earth's biota, many new groups are still left and or have been recently barcoded for the first time (Alonso et al. 2014, González-Varo et al. 2014, Leavitt et al. 2014, McFadden et al. 2014, Shokralla et al. 2014, Vargas et al. 2014, Vierna et al. 2014). Amalgamation of NGS technology and DNA barcoding will be the new exciting turn point of barcoding in near future. Shokralla *et al.* 2014 demonstrated the potential application of next-generation sequencing platforms for parallel acquisition of DNA barcode sequences from hundreds of specimens simultaneously. 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. These data, and the associated information on the evolutionary histories of taxa that they can provide, offer great opportunities for ecologists to investigate questions that were previously difficult or impossible to address. Joly et al. (2014) presented an overview of potential uses of DNA barcoding relevant in the age of ecoinformatics, including applications in community ecology, species invasion, macroevolution, trait evolution, food webs and trophic interactions.

2.4 Applications of fish DNA barcoding

Bartlett and Davidson (1991) were among the first to use mtDNA sequencing for fish identification, showing that cytochrome b sequences could discriminate four species of tuna (Thunnus spp.). They subsequently proposed forensically important nucleotide sequences as a means of identifying fishes (Bartlett and Davidson 1992). A large community of scientists joined forces in 2005 to launch the Fish Barcode of Life (FISH-BOL) campaign to meet the needs of an accurate inventory of species and a more scalable and cost-effective approach to their reliable identification at any life-history stage (Lundberg et al. 2000).

Ward et al. (2005) sequenced a 655 bp region of the mitochondrial *COI* of 207 species of fish. All species could be differentiated by their *COI* sequence, although single individuals of each of two species had haplotypes characteristic of a congener. Marine biologists also considered the value of the barcoding technique in identifying cryptic and polymorphic species and have suggested that the technique may be helpful when associations with voucher specimens are maintained (Saunders 2005).

In an important study in 2007, the status of *Eumicrotremus eggvinii*, was reassessed using 21 meristic and 32 morphometric characters analyzed for a total of 83 specimens of *E. spinosus and E. eggvinii*. Mitochondrial (*COI*, *COII and cytb*) and nuclear (Tmo-4C4) genes were also sequenced for both species, along with *E. derjugini*. The results indicated that although *E. spinosus and E. eggvinii* were clearly separated by a considerable number of morphological characters, they in fact constituted a single, sexually dimorphic species. Thirteen specimens of *E. eggvinii* (including the holotype) and 59 *E. spinosus* could be sexed; all individuals of *E. eggvinii* turned out to be males and all *E. spinosus* were females. Identical DNA sequences were found in all *E. eggvinii and E. spinosus* for *COI*, *COII* and *Tmo-4C4*, and a single shared synonymous substitution found in *cyt-b* (Byrkjedal et al. 2007).

Cohen et al. (2009) practically applied DNA barcode in identifying commercially available puffer fish. In 2007, two individuals developed symptoms consistent with

tetrodotoxin poisoning after ingesting home-cooked puffer fish purchased in Chicago. Both the Chicago retailer and the California supplier denied having sold or imported puffer fish but claimed the product was monkfish. However, genetic analysis and visual inspection determined that the ingested fish and others from the implicated lot retrieved from the supplier belonged to the family Tetraodontidae. Tetrodotoxin was detected at high levels in both remnants of the ingested meal and fish retrieved from the implicated lot. The investigation led to a voluntary recall of monkfish distributed by the supplier in three states and placement of the supplier on the U.S. Food and Drug Administration's Import Alert for species misbranding. This case of tetrodotoxin poisoning highlighted the need for continued stringent regulation of puffer fish importation by the U.S. Food and Drug Administration, education of the public regarding the dangers of puffer fish consumption, and raising awareness among medical providers of the diagnosis and management of foodborne toxin ingestions and the need for reporting to public health agencies.

Asgharian et al. (2011) with regard to practical applicability of DNA barcoding generated a large scale datasets of mitochondrial *COI* gene of fish of the Nayband National Park in the Persian Gulf. Amor et al. (2011) molecular characterized *Hysterothylacium aduncum* (Nematoda: Raphidascaridae) from different fish caught off the Tunisian coast based on nuclear ribosomal DNA sequences instead of *COI* DNA barcode. Carvalho et al. (2011) unveild a high rate of mislabeling in a commercial freshwater catfish from Brazil through DNA barcoding where they reported on the molecular identification results from processed fish products (i.e. fillets) and whole fishes sold in Brazilian markets under the common name surubim (Pseudoplatystoma spp.). They found DNA barcoding revealed the incorrect labeling of around 80% of all samples analyzed, with mislabeling being more pronounced within fillets rather than whole fish. Benziger et al. (2011) resolved the taxonomic ambiguity, and discussed the identity as well as systematic position of the Malabar snakehead, *C. diplogramma*, using morphological and molecular genetic (mitochondrial 16S rRNA and *COI* gene) information, in addition to making an attempt to understand its phylogenetic relationships and evolutionary biogeography.

Both morphological and genetic analyses support *C. diplogramma* as a distinct and valid species endemic to peninsular India and reveal its importance for conservation.

Wong et al. (2011) developed and evaluated DNA barcodes for use in differentiating United States domestic and imported catfish species. They also suggest that as the United States heightens inspection and regulation requirements for seafood products, DNA barcoding will serve as an important tool in efforts to ensure consumer safety and fair international commerce. April et al. (2011) established a barcode reference library for more than 80% of the named freshwater fish species of North America. This study demonstrates that 90% of known species can be delineated using barcodes. Moreover, it reveals numerous genetic discontinuities indicative of independently evolving lineages within described species, which points to the presence of morphologically cryptic diversity. From the 752 species analyzed, our survey flagged 138 named species that represent as many as 347 candidate species, which suggests a 28% increase in species diversity. In contrast, several species of parasitic and nonparasitic lampreys lack such discontinuity and may represent alternative life history strategies within single species. Bucklin et al. (2011) calculated an average retrieval of 2% new species in larger fish DNA barcoding studies, and they extrapolated this rate to about 600 overlooked or cryptic species to await discovery through similar studies. From the 31,000 species currently listed in the Catalog of Fishes, about 4000 have been described new during the past 10 years (2000–2009), with 500 added in 2008 and 300 in 2009 (Eschmeyer et al. 2012).

Becker et al. (2011) provided a 5-year progress report on the campaign and includes an updated "Collaborators' Protocol" (Steinke and Hanner 2011) to facilitate its continued growth and success. The implementation of standards (Hubert et al. 2008) is attributed to the overarching success of barcoding and to this end, the new protocol aims to refine and further advance FISH-BOL best practices for the benefit of the user community. Key to this objective is the widespread adoption of specimen imaging and reporting of identification "confidence levels" as discussed in the new protocol, which also reiterates the importance of a shared informatics workbench, the Barcode of Life

Data system (Ratnasingham S 2007). The utility of FISH-BOL derives from the contributions of many and varied researchers from around the world who are dedicated to expanding the barcode coverage for global fishes. The accumulating data already support applications of DNA barcoding which reveal market substitution (Jackson et al. 2001, Naiman and Magnuson 1995, Naiman and Turner 2000) and enhancing our understanding of fisheries exploitation (Holmes et al. 2009; Doukakis et al. 2011).

Weigt et al. (2012) represented a DNA barcode data release for 3,400 specimens representing 521 species of fishes from 6 areas across the Caribbean and western central Atlantic regions (FAO Region 31). Merged with their prior published data, the combined efforts resulted in 3,964 specimens representing 572 species of marine fishes and constitute one of the most comprehensive DNA barcoding ‘‘coverages’’ for a region reported to date. Kadarusman et al. (2012) assessed the diversity of the Papua rainbowfishes with DNA barcoding. Unexpected levels of cryptic diversity and endemism were detected since additional cryptic lineages were detected in several watersheds from the Vogelkop and the Lengguru massif.

The FISH-BOL campaign until 2012 has barcoded for the cytochrome c oxidase subunit I (*COI*) gene about 8,000 of the 31,000 fish species currently recognized. This includes the great majority of the world's most important commercial species. Results thus far show that about 98% and 93% of marine and freshwater species, respectively, are barcode distinguishable (Ward 2012). From Africa (Malimqvist and Rundle 2002, Rahel 2002) and Europe (Postel and Ricther 2003), Oceania (Revenga et al. 2005) and South America (Carvalho et al. 2011, Pereira et al. 2011) a large number of researchers have contributed to the FISH-BOL campaign.

Pereira et al. (2013) aimed to test the effectiveness of the barcoding methodology (*COI* gene) to identify the mega diverse freshwater fish fauna from the Neotropical region. Isabelle et al. (2013) barcoded unidentifiable fish items from the stomachs of 130 lionfish captured on Bahamian coral reefs. They identified 37 fish prey species, nearly half of which had not previously been recorded in this region. Puckridge et al. (2013) employed *COI* sequencing alongside traditional taxonomic identification methods and

uncovered instances of deep intraspecific genetic divergences among flathead species. Sixty-five operational taxonomic units (OTUs) were observed across the Indo-West Pacific from just 48 currently recognized species. Ko et al. (2013) pointed out that due to insufficient morphological diagnostic characters in larval fishes, it is easy to misidentify them and difficult to key to the genus or species level. They tried to find out, by applying DNA barcoding, how inconsistent the identifications can be among larval fish taxonomists.

Hellberg et al. (2014) compared modified versions of three DNA extraction kits (i.e., Qiagen DNeasy Blood and Tissue Kit, Sigma-Aldrich Extract-N-Amp Kit; and Life Technologies MagMax-96 DNA Multi-Sample Kit) and two polymerase chain reaction (PCR) setup methods (manual vs. automated) for use in DNA barcoding, with a focus on minimizing time, costs, and labor. Overall, the modified Extract-N-Amp Kit offered the greatest reduction in time and costs, while the DNeasy Blood and Tissue Kit produced sequences with the highest quality and highest initial success rates. Automation of the PCR setup process resulted in slightly greater success (100 %) compared to manual PCR setup. Khedkar et al. (2014) described the species diversity of fishes of the Narmada River in India. Keskin (2014) used eDNA approach to investigate non-native freshwater fish species from fifteen different locations of Upper Sakarya Basin. They detected four of the most common invasive freshwater fish species. Their results clearly indicated that eDNA surveys could be used as an important molecular tool to monitor invasive fish species in freshwater ecosystems.

Exploration of ichthyodiversity across the globe using DNA barcoding and application of barcoding for commercial purpose continues to grow with many new fish groups being barcoded and using the barcodes for novel applications (Baldwin and Johnson 2014, Bhattacharjee and Ghosh 2014, Cutarelli et al. 2014, Jo et al. 2014, Jones et al. 2013, Laskar et al. 2013, Maralit et al. 2013, Nunes et al. 2014, Takahara et al. 2013, Winterbottom et al. 2014, Young et al. 2013, Zhu et al. 2013).

2.5 Evolution of DNA barcoding as a technique

2.5.1 Distance and monophyly based methods

Hebert et al. (2003a) proposed the use of genetic distance matrix based on K2P divergence for estimating the interspecies and intraspecies distance in the *COI* gene. For the species-level analysis, nucleotide-sequence divergences were calculated using the K2P model, the best metric when distances are low (Nei 2000) as in this study. Neighbor-joining (NJ) analysis, implemented in MEGA2.1 (Kumar 2001), was employed to both examine relationships among taxa in the profiles and for the subsequent classification of 'test' taxa because of its strong track record in the analysis of large species assemblages (Kumar and Gadagkar 2000).

Since its advent DNA barcoding received wide response from the scientific community and was followed by barcoding intiatives from across the globe for various organisms. Most of the studies used the distance based approach (Besansky et al. , Blaxter et al. 2004, Hogg and Hebert 2004). Some researchers have envisioned "DNA taxonomy", a concept of adopting DNA sequencing as a central criterion for taxonomic decisions and descriptions, and have proposed using DNA barcodes as the standard method of analysis (Blaxter 2003, Tautz et al. 2003, Vogler and Monaghan 2007). However, there is concern over adopting one specific sequence region as the only criterion for taxonomic studies (DeSalle et al. 2005, Lipscomb et al. 2003, Rubinoff 2006). In addition, it is quite apparent that the DNA barcode itself is not a new species concept (i.e. a species cannot be defined based on the barcode only); neither does it provide enough information to describe unknown specimens as a new species. The results of barcoding can only suggest new species candidates (Brown et al. 2003, Hajibabaei et al. 2007, Hajibabaei et al. 2006b, Waugh 2007) as well as other valuable supporting information (e.g. distribution, life history, host plants) for taxonomic studies (e.g. integrative taxonomy: (Dayrat 2005, Schlick-Steiner et al. 2010, Yoshitake et al. 2008). Species descriptions using barcodes based on type specimens will become more common and important in the near future.

2.5.2 Advent of Character based methods

DeSalle et al. (2005) pointed out that a major shortcoming of using distances in DNA Barcoding is that that similarity scores often do not give the nearest neighbor as the closest relative (Koski and Golding 2001). Nevertheless, similarity scores will always give a nearest neighbor. They suggested that an alternative approach including character based phylogenetic analysis is more appropriate for establishing or 'printing' barcodes.

Sarkar et al. (2002a) presented a novel and simple method that exhaustively scanned microarray data for unambiguous gene expression patterns. Such patterns of data were used as the basis for classification into biological or clinical categories. The method, termed the Characteristic Attribute Organization System (CAOS), was derived from fundamental precepts in systematic biology. In CAOS two types of characteristic attributes ('pure' and 'private') were defined in gene expression microarray data. They also considered additional attributes ('compound') that are composed of expression states of more than one gene that are not characteristic on their own. CAOS was tested on three well-known cancer DNA microarray data sets for its ability to classify new microarray samples. CAOS was found to be a highly accurate and robust class prediction technique.

Sarkar et al. (2008) introduced the existing CAOS software with a set of software tools that implement the previously described Characteristic Attribute Organization System for both diagnostic identification and diagnostic-based classification. The software is publicly available from http://sarkarlab.mbl.edu/CAOS.

Rach et al. (2008) demonstrated the potential of character-based DNA barcodes by analysing 833 odonate specimens from 103 localities belonging to 64 species. A total of 54 species and 22 genera could be discriminated reliably through unique combinations of character states within only one mitochondrial gene region (NADH dehydrogenase 1).

Reid et al. (2011) assessed variability within the barcode region and the utility of both distance-based and character-based methods for species identification for *COI* barcode

sequences (650 bp) for 174 turtle species. They suggested that complementing distance-based barcoding with character-based methods for identifying diagnostic sets of nucleotides provided better resolution in several cases where distance-based methods failed to distinguish species.

Zou et al. (2012) reported a comprehensive barcoding analysis of 22 Nassarius species. They integrated the mitochondrial and nuclear sequences and the morphological characters to determine 13 Nassarius species and revealed four cryptic species and one pair synonyms. Distance, monophyly, and character–based barcoding methods were employed.

2.5.3 Comparison of distance based and character based methods

Rosso et al. (2012), barcoded the fish fauna of the Pampa Plain *COI* sequences were analysed by means of distance (K2P/NJ) and character-based (ML) models, as well as the Barcode Index Number (BIN). K2P/NJ analysis was able to discriminate among all previously identified species while also revealing the likely occurrence of two cryptic species that were further supported by BIN and ML analyses. On the other hand, both BIN and ML were not able to discriminate between two species of Rineloricaria. Despite the small genetic divergence between *A. cf. pampa* and *A. eigenmanniorum*, a tight array of haplotypes was observed for each species in both the distance and character-based methods.

Many studies evaluated the potential of distance-based thresholds and character-based DNA barcoding for the identification of problematic species-rich taxa. Bergmann et al. (2013) sequenced and compared gene fragments of CO1 and ND1 for 271 odonate individuals representing 51 species, 22 genera and eight families. Their data suggested that (i) the combination of the CO1 and ND1 fragment forms a better identifier than a single region alone; and (ii) the character-based approach provides higher resolution than the distance-based method in Odonata especially in closely related taxonomic entities. Abdullah and Rehbein (2014) demonstrated that the *COI* gene could be more reliably used as a tool for Indonesian commercial tuna products authentication, if the

sequencing results were combined with the character-based identification using differences at certain nucleotide positions. van Velzen et al. (2012) compared six methods to correctly identify recently diverged species with DNA barcodes: neighbor joining and parsimony (both tree-based), nearest neighbor and BLAST (similarity-based), and the diagnostic methods DNA-BAR, and BLOG. Their results showed, that success rates are significantly lower for recently diverged species (~75%) than for older species (~97%) ($P<0.00001$). Similarity-based and diagnostic methods significantly outperform tree-based methods, when applied to simulated DNA barcode data.

White et al. (2014) investigated the effect of delimitation methods on outcomes of bioassessments based on DNA barcodes. They used 2 tree-construction methods (NJ, maximum likelihood) and 4 classes of species-delimitation criteria (distance-based, bootstrap support, reciprocal monophyly, and coalescent-based) with a DNA barcode data set consisting of 3 genera and 2202 COI sequences. They assessed congruence among trees and compared species abundances and estimated species richness among methods. NJ followed by use of a standard barcoding distance cutoff (2%) yielded the greatest number of putative species. All other delimitation methods yielded similar, but lower, richness.

2.5.4 New methods of barcoding

Many recent studies explored new methods of reading barcodes. Brown et al. (2012) created Spider: SPecies IDentity and Evolution in R. It is a new R package implementing a number of useful analyses for DNA barcoding studies and associated research into species delimitation and speciation.

Weitschek et al. (2013) created BLOG (Barcoding with LOGic) a diagnostic and character-based DNA Barcode analysis method. Its aim is to classify specimens to species based on DNA Barcode sequences and on a supervised machine learning approach, using classification rules that compactly characterize species in terms of DNA Barcode locations of key diagnostic nucleotides. Tanabe and Toju (2013) proposed two new computational methods of DNA barcoding and show a benchmark for

bacterial/archeal 16S, animal COX1, fungal internal transcribed spacer, and three plant chloroplast (rbcL, matK, and trnH-psbA) barcode loci that can be used to compare the performance of existing and new methods.

Fan et al. (2014) developed a practical program for both accurate and scalable species identification for DNA barcoding. The VIP Barcoding is a user-friendly software in graphical user interface for rapid DNA barcoding. It adopts a hybrid, two-stage algorithm. First, an alignment-free composition vector (CV) method is utilized to reduce searching space by screening a reference database. The alignment-based K2P distance nearest-neighbor method is then employed to analyze the smaller data set generated in the first stage. In comparison with other software, VIP Barcoding has higher accuracy than Blastn and several alignment-free methods and higher scalability than alignment-based distance methods and character-based methods.

Porter et al. (2014) pointed out that current methods to identify unknown insect (class Insecta) cytochrome c oxidase (*COI* barcode) sequences often rely on thresholds of distances that can be difficult to define, sequence similarity cut-offs, or monophyly. They used a naïve Bayesian classifier to automate taxonomic assignments for large batches of insect *COI* sequences such as data obtained from high-throughput environmental sequencing.

20 years since the introduction of DNA barcoding, it has moved from a theoretical concept to practical applied field. Information gathered from DNA barcodes can be used beyond taxonomic studies and will have far-reaching implications across many fields of biology, including ecology (rapid biodiversity assessment and food chain analysis), conservation biology (monitoring of protected species), biosecurity (early identification of invasive pest species), medicine (identification of medically important pathogens and their vectors) and pharmacology (identification of active compounds). However, it is important that the limitations of DNA barcoding are understood and techniques continually adapted and improved as this young science matures (Fišer Pečnikar and Buzan 2014).