DNA barcoding and molecular taxonomy of marine organisms

Srinivas Raghavan, V., Lijo John, Reynold Peter and K. K. Vijayan

Marine Biotechnology Division, CMFRI, Cochin - 682 018, vetvsr@yahoo.com

Taxonomical awareness on species, the intraspecific and interspecific relations etc, are the necessary pre requisites in different fields of applied biology, such as in biological resources management and conservation, in mariculture ventures, in pest/pathogen control and identification, etc. Generally, there are two contexts in which significant problems arise in species identification, where the molecular approaches are applicable. The first concerns the very fundamentals of taxonomy, in identifying the species, the phylogenetic status and the identification at subspecies and hybrids level. The second concerns to ecologists are the issue of identifying the species sex, or identity of individuals under circumstances where simple morphology cannot be relied upon. Biodiversity studies require species level analyses to assess the community structures. In conventional taxonomic practice, organisms were classified based on their fine morphological characters, comparative osteology, etc., which are more time consuming, laborious and expensive. Moreover, in some situations, morphological variation arises due to the environmental factors rather than genetic causes and is therefore not heritable. In this context, classification of organisms based on the molecular variations using the modern molecular tools has got a wide acceptance globally.

In molecular taxonomy, variations in protein profile and the nucleic acid structure among organisms has been utilised in determining the evolutionary relationship of the species. Applications of molecular taxonomy and molecular systematics were pioneered by Charles G. Sibley (birds), Herbert C. Dessauer (herpetology), and Morris Goodman (primates), followed by Allan C. Wilson, Robert K. Selander and John C. Avise (who studied in various groups) during 1960's and 70's. Allozymes, RFLP, RAPD, microsatellite, mitochondrial DNA sequence analysis, etc. are the most popular genetic markers commonly used in molecular taxonomic studies. Eventhough, proteins and mitochondrial, nuclear and chloroplast DNA are used for molecular taxonomic analyses, applications involving the protein are far fewer than those with DNA markers due to several reasons. Most importantly, protein polymorphism is usually much less than that detectable in DNA, thus greatly limiting the resolving power of protein methods. Protein analysis generally requires relatively large amounts of tissue, which will be a limiting factor when working with small organisms. Proteins are also often differentially expressed both in space (tissue specificity) and time (developmental regulation), which limits its availability. Further, proteins are difficult to store in non-denatured state than DNA under field conditions, which may or may not matter according to the type of analysis involved.

Despite the limitations recounted above, protein can be a useful marker in certain circumstances. 'Protein profiling' using the sodium dodecyl sulfate - polyacrylamide gel electrophoresis (SDS-PAGE)

is an established molecular tool in identifying/differentiating organisms. This is a simple procedure and protein sample can be collected and stored in ethanol, just like DNA, because denaturation is inherent in the protocol. Allozyme analysis is rarely used in identification studies because of its low resolving power. Allozyme study requires non-denatured protein (usually by deep freezing) and also more expensive and time consuming to generate the results. This situation paved the way to establish the DNA sequencing as a reliable and acceptable tool in recent times. Hence, these are considered as a superior marker for evolutionary studies, because the actions of evolution are ultimately reflected in DNA sequences. As there are millions of species and life stage transformations, DNA based identification aid in resolution and strengthen the classical taxonomical identification system.

DNA barcoding is a taxonomic method that uses a short DNA sequence of the organism's gene to identify the species. In 2003, Paul Hebert, researcher at the University of Guelph in Ontario, Canada, proposed "DNA barcoding" as a way to identify species. DNA barcoding uses a short genetic sequence from a standard part of the organism's genome as the way a supermarket scanner distinguishes products using the black stripes of the Universal Product Code (UPC). Partial sequence of the mitochondrial cytochrome c oxidase subunit I (*cox1*- usually referred to as COI in barcoding studies) gene are considered as a potential 'barcode'. The intent of DNA barcoding is to use large-scale screening of one or a few reference genes in order to (i) assign unknown individuals to species and (ii) enhance discovery of new species. The purpose of DNA barcoding is to identify a species with a piece of DNA with which a biologist could run several biotic surveys without the need of morphological keys.

What is the need for DNA barcoding?

Until now, biological specimens were identified using morphological features like the shape, size and color of body parts. In some cases a trained individual could make routine identifications using morphological "keys" but in most cases an experienced professional taxonomist is needed. If a specimen is damaged or is in an immature stage of development, even taxonomists may be unable to make identifications. Molecular markers can be used to solve these problems which offer a wide range of options for identifying species, but the question arises as to whether a general, universally applicable method might be found. Now comes the role of DNA barcoding, could provide just such an approach. The idea here is to select one or a few genes that are shared by most, if not all organisms on earth and which show large interspecific but small intraspecific levels of variation. The sequences of such genes could then become the equivalent of species-specific barcodes.

A DNA barcode is not just any DNA sequence, it is a rigorously standardized sequence of a minimum length and quality from an agreed-upon gene, deposited in a major sequence database, and attached to a voucher specimen whose origins and current status are recorded. The primary goal is to develop an accurate, rapid, cost-effective, and universally accessible DNA-based system for species identification. It could also be applied where traditional methods are unrevealing, for instance identification of eggs and immature forms, and analysis of stomach contents or excreta to determine food webs. The Barcode of Life Data System (BoLD), provides an online interface (http://www.BOLDsystems.org) allowing the scientists and researchers to work together and share information on a global scale. Hundreds of different barcoding projects around the world are contributing to a barcode library that contains hundreds of thousands of specimen barcodes from around 70,000 species. DNA barcoding has the potential to increase access to taxonomic knowledge

in all regions of the world. Databases of reference barcodes are connecting specimens to their correct species names, providing a direct route to species information associated with those names. The Consortium for the Barcode of Life (CBOL) is working with GenBank and its partner DNA repositories (European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ)) to construct a global library of reference barcode sequences. Once fully implemented, this system will revolutionize access to biological information and will exert broad impacts on research, policy, pest and disease control, food safety, resource management, conservation, and many other areas in which society interacts with wild biodiversity.

A DNA barcoding system will enable the prompt diagnosis of invasive species, thereby allowing quarantine and eradication efforts to begin years earlier with massive reductions in cost and increased chances of success. The same strategy extends to the selection of optimal control strategies for pest/ pathogen species impacting the varied natural resource sectors. Similarly, it can, as well, play a critical role in regulating trade in endangered or protected species or products. DNA barcoding can also assist with one of the most intractable practical problems in ecology, notably the study of predator–prey interactions. Particularly with small organisms, identifying prey and especially prey range can be a daunting task. Mostly this has to be done by the manual examination of gut contents. Unfortunately digestive processes rapidly destroy many morphological clues to prey identity, but molecules or fragments of them may persist for longer. DNA-based methods offer a useful alternative and further the barcoding gene is from mtDNA has the advantage in this kind of study of high copy number, thus maximizing the chances of detection by PCR before its complete destruction by digestive enzymes. DNA barcoding seeks merely to aid in delimiting species - to highlight genetically distinct groups exhibiting levels of sequence divergence suggestive of species status – but are never sufficient to describe new species.

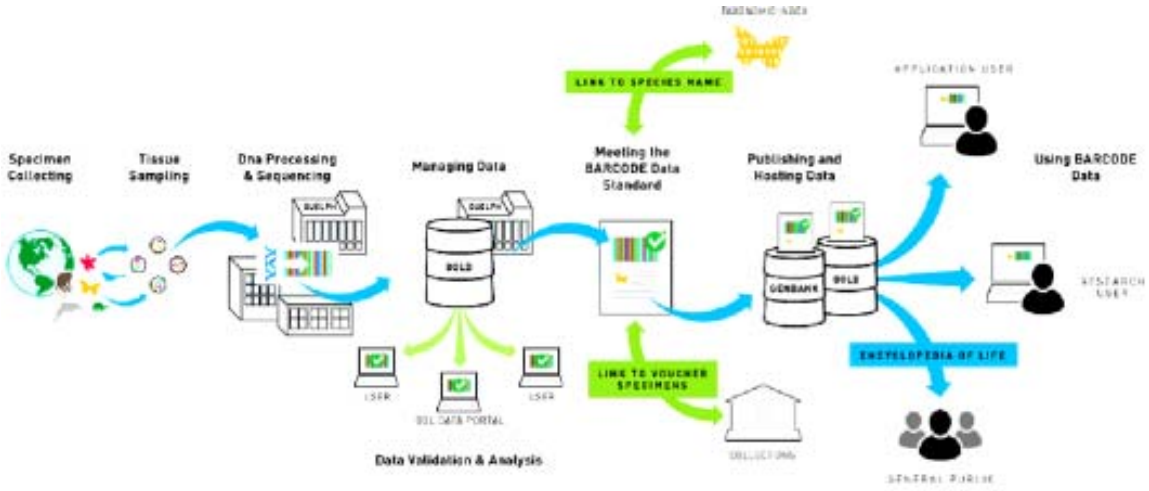

Fig. 1. The barcoding pipeline (source: CBOL website)

Marker of Choice:

DNA barcoding is based on a relatively simple concept. Most eukaryote cells contain mitochondria and mitochondrial DNA (mtDNA), therein observed relatively higher mutation rate. This results in significant variation in mtDNA sequences even between the species and in principle, a comparatively less variation within species. Desirable locus for DNA barcoding should be standardized, preferably

present in all taxa of interest and sequenceable without species specific PCR primers, short enough to be easily sequenced with the available technology and expected to provide large variation between species and a relatively less variation within the species. Although several loci have been suggested, for animals and many other eukaryotes, partial region of the mitochondrial cytochrome c oxidase subunit I (COI) gene was proposed as a potential 'barcode' and widely accepted internationally.

Even though the mitochondrial genes encoding ribosomal DNA (12S, 16S) are widely used as marker in phylogenetic analysis, their utility in taxonomic analyses are found to be limited due to the prevalence of insertions and deletions (indels) which in turn complicate sequence alignments. There are different protein-coding genes in the animal mitochondrial genome which are better target due to the rarity/absence of indels as this will lead to frame shift. The important advantages of COI as the gene for DNA barcoding are the availability of universal primers for amplifying this gene in different animal phyla and it possesses a greater range in phylogenetic signal than the other mitochondrial gene. This gene has some other desirable properties too, to be used as the gene for barcoding. The COI gene being mitochondrial it is usually present in high copy number per cell and is a necessary gene in all aerobic organisms. Partial COI sequence of about 650 bp has turned out to have, in a wide variety of organisms from insects to birds, high interspecific but low intraspecific variation. However, there are some significant difficulties with this locus. Anaerobic organisms are excluded and interspecific variation of mtDNA (including COI) in plants other than algae is often too low to be useful. Prokaryotes, which include most of the earth's biodiversity, are essentially excluded. There is also a risk of errors with any single locus from lineage sorting in recently diverged species where reciprocal monophyly has not yet been achieved. It looks as if at the very least barcoding will have to include sequences from several different genes to be universally applicable. In the case of plants, plastid loci such as ribulose bisphosphate carboxylase (*rbcL*) and megakatyocyte-associated tyrosine kinase (*matK*) have been proposed. Nuclear ribosomal DNA intragenic spacer (ITS) regions have also been proposed for species identification in several other eukaryotes such as protozoa.

The selection of COI as a target gene for DNA barcoding is supported by published and ongoing work, which demonstrates that barcoding via COI, will meet the goals for a wide diversity of animal taxa (http: //www.barcodinglife.com//). A model COI profile, based upon the analysis of a single individual from each of 200 closely allied species of lepidopterans, was 100% successful in correctly identifying subsequent specimens (Hebert *et al*., 2003a). Two hundred and seven species of fish, mostly Australian marine fish, were sequenced (barcoded) for a 655 bp region of the mitochondrial cytochrome oxidase subunit I gene (cox1) as reported by Ward *et al*. (2005). Spies *et al*. (2006) examined the variation at the mtDNA COI gene in 15 species of North Pacific skates and indicated that, a DNA-based barcoding approach may be useful for species identification. DNA barcoding reveals a likely second species of Asian sea bass (barramundi) (*Lates calcarifer*); Ward *et al*., (2008a) strongly suggest that barramundi from Australia and from Myanmar are different species based on the sequencing of 650 base pair region of the mitochondrial COI gene. Out of fifteen fish species barcoded from Northern (Atlantic and Mediterranean) and Southern (Australasian) Hemisphere waters using COI sequences, Ward *et al*. (2008b) observed significant evidence of spatial genetic differentiation for this gene in two fishes; the silver scabbardfish (*Lepidopus caudatus*) and John dory (*Zeus faber*). These observations further supported the scope of barcoding in identifying species and also the geographical variations expected within species.

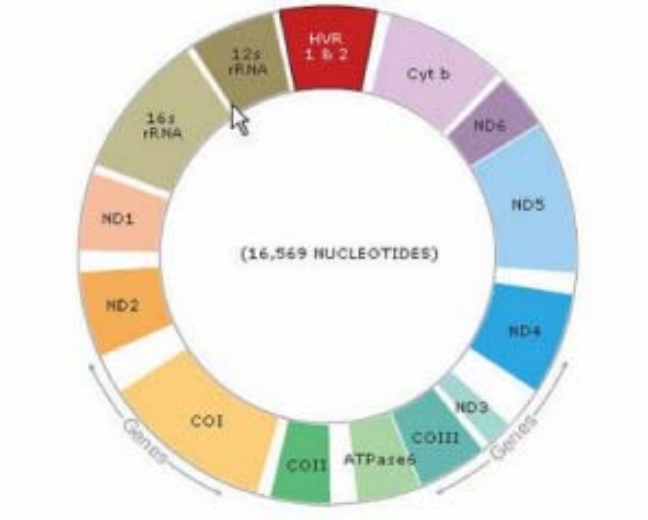

Fig. 2. Diagrammatic representation of vertebrate mitochondrial genome

Components of Barcoding:

Any barcoding project has 4 components

1. The Specimen which forms the treasure trove for DNA barcoding.
2. DNA analysis involving standard laboratory protocol.
3. Database to assign unknown specimens to known ones. The main data base is Barcode of Life database (FISH BOL, BOLD) created and maintained by University of Guelph, Ontario, Canada which offers researchers to collect analyze and manage barcode data. It's an international collaboration which helps in assembling the CO1 sequences of different species. The FISH BOL helps in curating the barcode sequence, electropherograms, voucher specimens, GIS data of different collection sites and aids in interpretation of data.
4. Data analysis which helps in identifying the unknown species with closely matching species in the data base.

Steps in DNA barcoding:

The main steps involved in DNA barcoding of samples are,

1. Collection and field identification of specimens and its taxonomic data.
2. Maintaining voucher specimen with voucher data (catalogue number and institution storing), collection records (collector, collection date and location with GPS coordinates) and photographs of the specimen.
3. Extraction of DNA from tissue samples using standardized protocol.
4. Amplification of mtDNA genes like COI by polymerase chain reaction (PCR).
5. Qualification and quantification of PCR products through Agarose gel electrophoresis.
6. Purification of PCR products and sequencing.

7. Analysis and validation of sequence data.
8. Submission of morphological data, specimen voucher data, sequence information and photographs in FISH BOL database.

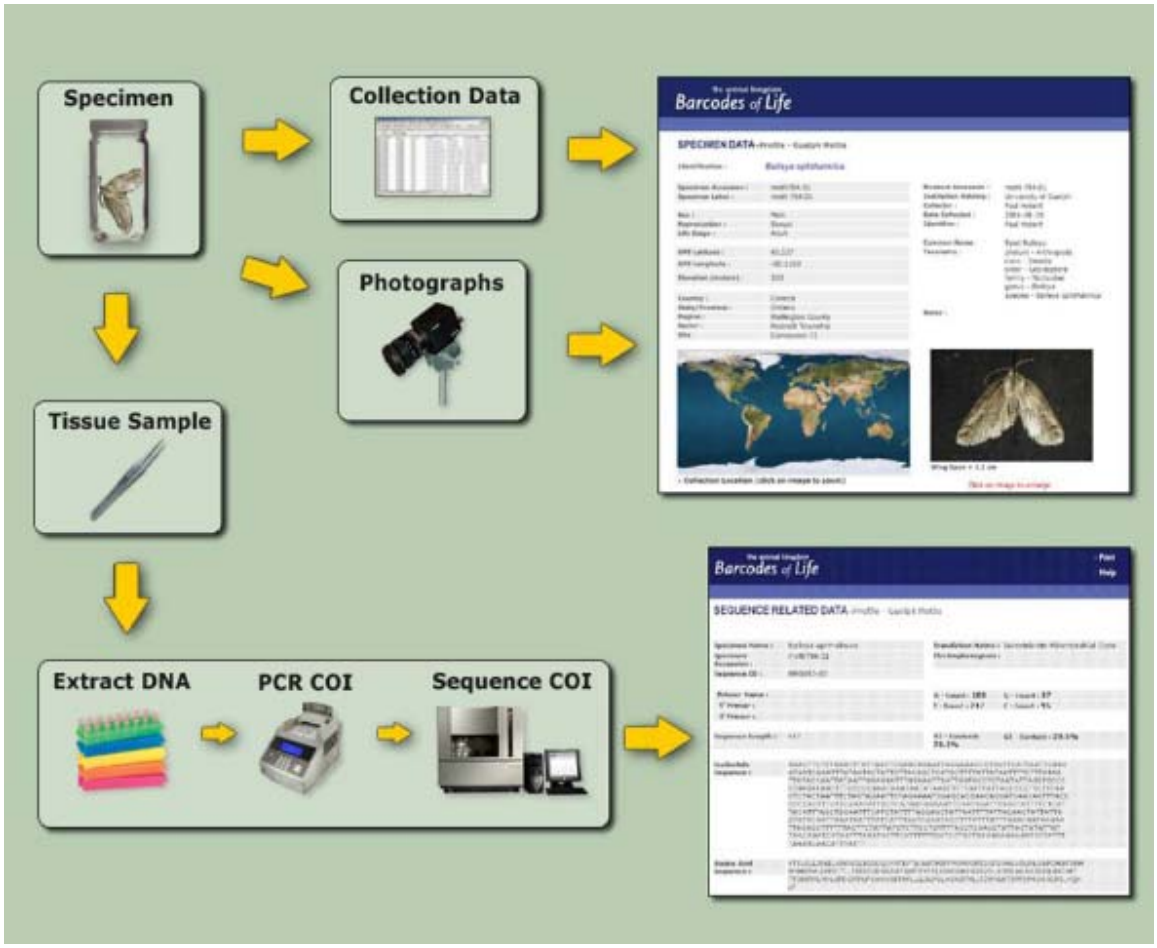

Fig 3. Barcoding: Process & database (Source: www.barcodinglife.org)

Utilities of DNA barcoding:

1. *Consumer protection*: DNA barcoding mainly involve consumer, agricultural, health and environmental protection. Apart from, protection of ecosystem, DNA barcoding also helps in identification of catch and by-catch on commercial vessels and at the dock, better understanding of the food chain through analysis of gut contents and improved fish stock assessments, based on identification of larvae as well as juveniles and adults. The Consortium for the Barcode of Life (CBOL) generates barcodes for economically relevant and potentially hazardous fish species. The barcoding method would improve species identification, which is essential in determining associated hazards, addressing economic fraud issues and aiding in food-borne illness outbreak investigations.

2. *Conservation of Biodiversity*: A library of barcodes will enable researchers to identify species, about its nativity, abundance and endangering status both locally and globally. The molecular taxonomy is very critical in developing strategies to preserve different genetic entities or species to enable the species identification of eggs, larvae and tissues. DNA barcoding is a combination of molecular techniques and traditional taxonomy, cost-effective to investigate the unexplored species and biodiversity.

3. *Cataloguing of extinct species*: Molecular taxonomy can identify species even with tiny, damaged, old and ancient specimens. DNA barcoding helps in identifying illegally obtained wildlife species and poorly preserved samples. Barcoding forms a bridge between creating a reference library of barcodes of species already known to science which in turn will help us to understand the molecular pattern of speciation.

Limitations of DNA barcodes:

DNA barcodes increase our ability to identify species accurately. But the major draw back is that, only a representative sample from a population of a particular species is taken for study. This sample may not be reflecting the entire diversity present in a particular species leading to inaccurate identification of a species with unusual genetic variation patterns. Barcoding only identify species but hardly throws light on the evolutionary pattern of a species. The mtDNA would not reflect upon the evolutionary relationships between different groups and will not provide sufficient information on the taxa of species leading to flaws in identification. As the mtDNA is inherited maternally, hybrids cannot be identified using COI based DNA barcodes. Occurrence of nuclear mitochondrial pseudogenes (numts), which are nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms can be easily co-amplified with orthologous mtDNA by using conserved universal primers forms another major limiting factor for DNA barcoding. Many of which are highly divergent from orthologous mtDNA sequences, and DNA barcoding analysis may incorrectly overestimates the number of unique species based on this and can introduce serious ambiguity into DNA barcoding.

Conclusion:

The barcoding research not only provides an increased understanding of biodiversity but also helps to understand the basic evolutionary process. The DNA barcoding would not replace the Linnaean system of classification. The large scale sequencing when integrated with the traditional taxonomical identification will contribute to the challenge of identifying species and enhance the rate of discovering the biological diversity. The real challenge lies in defining the boundaries of using barcoding technique with a large phylogeographic structure. DNA barcoding generates a plethora of information which does not compete with the traditional taxonomy but in turn supplements the taxonomists for accurate identification of species using both morphological keys and molecular markers. If mtDNA alone is used for barcoding a species, then we may end up with identifying only a few species in such a large global diversified conditions. So, an integrated approach using different markers should be carried out while using molecular taxonomy for categorizing a species to particular taxa.

Suggested Readings:

Hebert, P. D. N., A. Cywinska, S. L. Ball and J. R. deWaard. 2003*a*. Biological identifications through DNA barcodes. *Proceeding Royal Society London Series B.,* 270**:**313–321.

Hebert, P. D. N., S. Ratnasingham and J. R. deWaard. 2003*b*. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Royal Society London Series B* 270(Supplement):S96–S99.

Moritz, C., and C. Cicero. 2004. DNA barcoding: promise and pitfalls. *PLoS Biology.,* 2**:**1529–1531.

Spies, I.B., S. Gaichas, D. E. Stevenson, J.W. Orr and M. F. Canino, 2006. DNA-based identification of Alaska skates (*Amblyraja, Bathyraja* and *Raja*:Rajidae) using cytochromec oxidase subunit I (coI) variation. *Journal of Fish Biology*, 69(B): 283–292.

Ward, R. D., T. S. Zemlak, B. H. Innes, P. R. Last and P. D. N. Hebert, 2005. DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London B*. doi:10.1098/rstb.2005.1716 Published online.

Ward, R. D., B. H. Holmes, and G. K. Yearsley, 2008a. DNA barcoding reveals a likely second species of Asian sea bass (barramundi) (*Lates calcarifer*). *Journal of Fish Biology*, 72: 458–463.

Ward, R. D., F. O. Costa, B. H. Holmes and D. Steinke, 2008b. DNA barcoding of shared fish species from the North Atlantic and Australasia: minimal divergence for most taxa, but *Zeus faber* and *Lepidopus caudatus* each probably constitute two species. *Aquatic Biology*, 3: 71–78.

Waples, R. S. 1991. Pacific salmon, *Oncorhynchus* spp., and the definition of "species" under the Endangered Species Act. *Marine Fisheries Review.,* 53**:**11–22.

Lipscomb, D., N. Platnick, and Q. Wheeler. 2003. The intellectual content of taxonomy: a comment on DNA taxonomy. *Trends in Ecology& Evolution.,*18**:** 64–66.