International Journal of Engineering Research & Management Technology
ISSN: 2348-4039
March-2017 Volume 4, Issue-2
Email: editor@ijermt.org
www.ijermt.org

# ROLE OF BIOINFORMATICS IN RECONSTRUCTION OF PHYLOGENETIC TREE OF DIFFERENT GROUPS OF FISHES ON THE BASIS OF MITOCHONDRIAL GENE SEQUENCE DATA

Dr. SudhaGarg
Department of Zoology
Hindu College Moradabad

## INTRODUCTION

There has been an explosive growth in recent years in biological sequence data. There is a great need for the storage and managing of this data to be protected and secure. Bioinformatics is the science that deals with storing, extracting, organizing, analyzing, interpreting and utilizing information. Basically, bioinformatics is the combination of using applications of computer technology in management of biological information.The central challenge of bioinformatics is the rationalization of the mass of sequence information, with a view not only to deriving more efficient means of data storage but also to designing more incisive analysis tools. The imperative derives that analytical process is the need to convert sequence information into biochemical and biophysical knowledge, to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences. (Attawood and Perry-Smith, 2006).

A living cell is a biological system where cellular components such as genome, the gene transcript and the proteins interact with each other, and these interactions determine the fate of the cell, e.g., whether a stem cell is going to become a liver cell or cancer cell. The characterization of these three types of components and the associated development of analytical methods lead to the establishments of three closely related branches of bioinformatics:

1. **Genomics,**
2. **Transcriptomics and**
3. **Proteomcs.**

The primary database like Protein Data Bank (PDB), Cambridge structural database (CSD), Bio Meg Res Bank (BMRB), Nucleic Acid structure Database (NDB) are main constituent of structural bioinformatics. They work as the repositories of knowledge regarding the structure of a molecule as well as experimental (raw/unprocessed) data used for deducing the structures. Developing analytical tools to discover or decipher the knowledge in data is another aspect of bioinformatics. The ultimate goal of analytical bioinformatics is to develop predictive methods that allow scientists to model the function and phenotype of an organism based only on its protein and genome sequences data. It includes following tasks-

1. Biological data research on the web.
2. Sequence analysis, pair-wise alignment.
3. Multiple sequence alignment, trees and profiles.
4. Visualizing protein structures and computing structural properties.
5. Predicting protein structure and function from sequence.
6. Tools for genomics and proteomics.

The scope of bioinformatics lies in the fact that it has empowered the research and development in the biological field, right from the level of single nucleotide to the vast classification. Online floras provide description about plants and simply by inserting the characters user can get the genus to which the plant belongs. Days are not far when a world of flora will be available and any local plant can be easily identified. Also one can access required data using such online information. The concept of e-herbarium has been already introduced and has generated overwhelming response (Attawood and Perry-Smith, 2006). Work on virtual taxonomy is in the pipeline where in a 360$^0$ view of herbarium along with its description using simple mouse would be quite possible. Decoding of genomes of various organisms will prove vital in comparing and assigning taxon to each individual. Moreover it will play decisive role in understanding evolution migration and adaptations of various organisms. Data of DNA barcodes has a potential for the development of e-museum for identification of particular animals.

Bioinformatics is a lot more than just a useful tool in biological research and the development of drugs. It is quickly becoming indispensable to all researchers and scientists alike. The technology of bioinformatics is both versatile and is able to be applied wherever research is being done on genetics, proteins and cells for the use of discovering and development of new drugs, evaluation of drug toxicology, pharmaceutical products and clinical trials have also been able to benefit from the use and technology of bioinformatics.

## MOLECULAR PHYLOGENY AND EVOLUTION

Every living organism contains DNA, RNA and Proteins. Closely related organisms generally have a high degree of agreement in the molecular structure of these substances, while the molecules of organisms distantly related usually show a pattern of dissimilarity. Conserved sequences, such as mitochondrial DNA, are expected to accumulate mutations over time, and assuming a constant rate of mutation and provide a molecular clock for dating divergence. Molecular phylogeny uses such data to build a "relationship tree" that shows the probable evolution of various organisms. Not until recent decades, however, has it been possible to isolate and identify these molecular structures. Early attempts at molecular systematic were also termed as chemotaxonomy and made use of proteins enzymes, carbohydrates and other molecules which were separated and characterized using techniques such as chromatography. These have been largely replaced in recent times by DNA sequencing which produces the exact sequences of nucleotides or *bases* in either DNA or RNA segments extracted using different techniques. These are generally considered superior for evolutionary studies since the actions of evolution are ultimately reflected in the genetic sequences.

The most common approach is the comparison of homologous sequences for genes using sequence alignment techniques to identify similarity. Another application of molecular phylogeny is in DNA barcoding, where the species of an individual organism is identified using small sections of mitochondrial DNA. Another application of the techniques that make this possible can be seen in the very limited field of human genetics, such as the ever more popular use of genetic testing to determine a child's paternity, as well as the emergence of a new branch of criminal forensic focused on evidence known as genetic fingerprinting.

Computational phylogenetics is the application of computational algorithms, methods and programs to phylogenetic analyses. The goal is to assemble a phylogenetic tree representing a hypothesis about the evolutionary ancestry of a set of gene, species, or other taxa. For example, these techniques have been used to explore the family tree of hominid species (1) and the relationships between specific genes shared by many types of organisms.(2) Traditional phylogenetics relies on morphological data obtained by measuring and quantifying the phenotypic properties of representative organisms, while the more recent field of molecular phylogenetics uses nucleotide sequences encoding genes or amino acid sequences encoding proteins as the basis for classification. Many forms of molecular phylogenetics are closely related to and make extensive use of sequence alignment in constructing and refining phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species. The phylogenetic trees constructed by computational methods are unlikely to perfectly reproduce the evolutionary tree that represents the historical relationships between the species being analyzed. The historical species tree may also differ from the historical tree of an individual homologous gene shared by those species.

Producing a phylogenetic tree requires a measure of homology among the characteristics shared by the taxa being compared. In morphological studies, this requires explicit decisions about which physical characteristics to measure and how to use them to encode distinct states corresponding to the input taxa. In molecular studies, a primary problem is in producing a multiple sequence alignment (MSA) between the genes or amino acid sequences of interest. Progressive sequence alignment methods produce a phylogenetic tree by necessity because they incorporate new sequences into the calculated alignment in order of genetic distance.

## MITOCHONDRIAL DNA (mtDNA) AND PHYLOGENETIC RESEARCH

With the development of DNA sequencing methods and the extensive sequencing experiments undertaken in the last two decades in a wide variety of organisms, the order of the genes in the mitochondrial DNA molecule has begun to be disclosed. A great number of phylogenetic studies using mitochondrial gene sequences have been done and reported, some dealing with the use of mitochondrial genes in the establishment of different levels of phylogenetic relationships (Kumazawa and Nishida, 1993; Zardoya and Meyer, 1996). For example, the control region of the mitochondrial genome is frequently used in population studies due to the high variability in its nucleotide sequence, while protein-coding genes, such as cytochrome b (Cyt*b*), are generally used for phylogenetic analysis of taxa above the species level. The success of mtDNA sequences in phylogenetic studies is due to several characteristics:

(a) Compact gene packing, with little noncoding intergenic nucleotides and some nucleotide overlapping between genes encoded in opposite strands (Cantatore and Saccone, 1987; and other genomes completely sequenced by several researchers);

(b) Lack of recombination (Clayton, 1982, 1992; Hayashi *et al*., 1985);

(c) Mainly maternal inheritance (Kondo *et al*., 1990; Gyllestein*et al*., 1991);

(d) Faster sequence evolution as compared to nuclear sequences, perhaps due to repair inefficiency

(Brown *et al*., 1979), and

(e) Multicity status in a cell (Michaels *et al*., 1982; Robin and Wong, 1988).

Until recently this approach was not used in vertebrate phylogeny due to the belief that there was a "conserved gene order" for the vertebrate mitochondrial genome. The reason for this is that the first complete mitochondrial genome sequences found in vertebrate taxa had no variation in the position of the genes along the molecule. This has been seen in taxa as diverse as *Xenopuslaevis*(Roe *et al*., 1985) humans and other mammals (Bibb *et al*., 1981; Anderson *et al*., 1981, 1982; Brown *et al*., 1982; Gadaleta*et al*., 1989; Árnason and Johnsson, 1992; Árnason and Gullberg, 1993), and some fish species (Johansen *et al*., 1990; Tzeng*et al*., 1992; Chang *et al*., 1994; Zardoya*et al*., 1995; Zardoya and Meyer, 1996). Thus, the typical vertebrate mitochondrial genome was assumed to have a noncoding control region, 13 protein-coding genes, two ribosomal RNAs (rRNA), and 22 transfer RNAs (tRNA) spread along the circular DNA molecule (Anderson *et al*., 1981; Bibb *et al*., 1981; Roe *et al*., 1985; Gadaleta*et al*., 1989; Johansen *et al*., 1990, Figure 1).

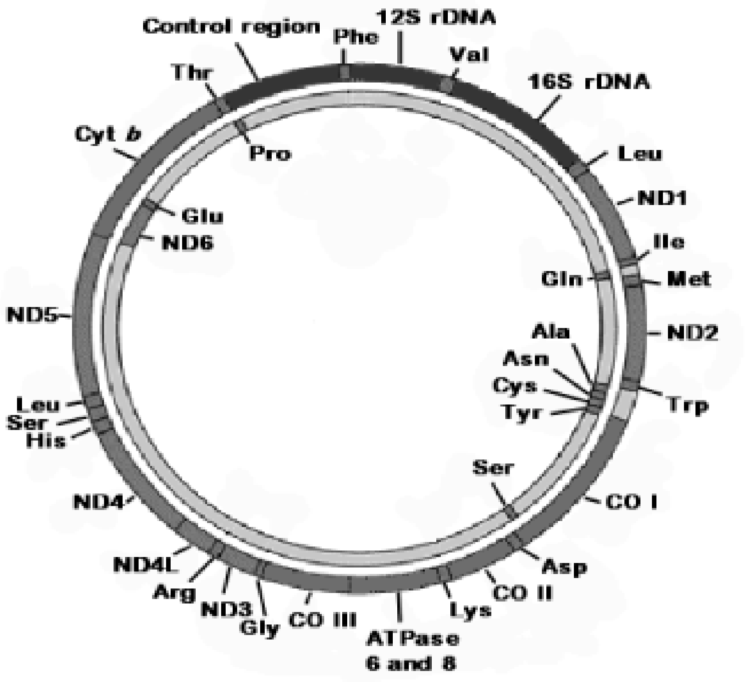

**Figure 1** - Schematic representation of the circular molecule of the "conserved" vertebrate mitochondrial genome. Genes outside and inside the circle are transcribed in the H and L strands, respectively. Protein-coding genes are represented as follows: Cyt *b* - cytochrome *b*; CO I, CO II and CO III - subunits I, II and III of the cytochrome oxidase; ND1-6 - subunits 1 to 6 of the NADH reductase. tRNA are represented by their three-letter amino acid abbreviations.

## PHYLOGENY OF FISHES (PISCES)

All modern fishes except cyclostomes belong to either Osteichthyes or Chondrichtyes. Fishes are diverse in morphology and worldwide in distribution. They outnumber all other vertebrates combined and one of the most successful group of animals.In the case of fishes, molecular approaches have not been widely used in phylogenetic studies. So far, only mitochondrial DNA sequences (Kocher et al. 1989; Meyer et al. 1990; Meyer and Wilson 1990; Normark et al. 1991; Martin and Palumbi 1992) or nuclear ribosomal sequences (Stock et al. 1991; Bernardi et al. 1992) have been used for recon- structing some fish phylogenies (in fact, either for several closely related taxa, or for a small number of taxa separated by large evolutionary distances). Indeed, very few genes coding for proteins have been sequenced in fishes (less than 200 for all fishesvs over 3,000 for man). Among them, the coding sequences of the growth hormone gene are the most widely known in primary structure and are potentially useful for phylogenetic studies covering broad phylogenetic spectra both in cold- and warm- blooded vertebrates.

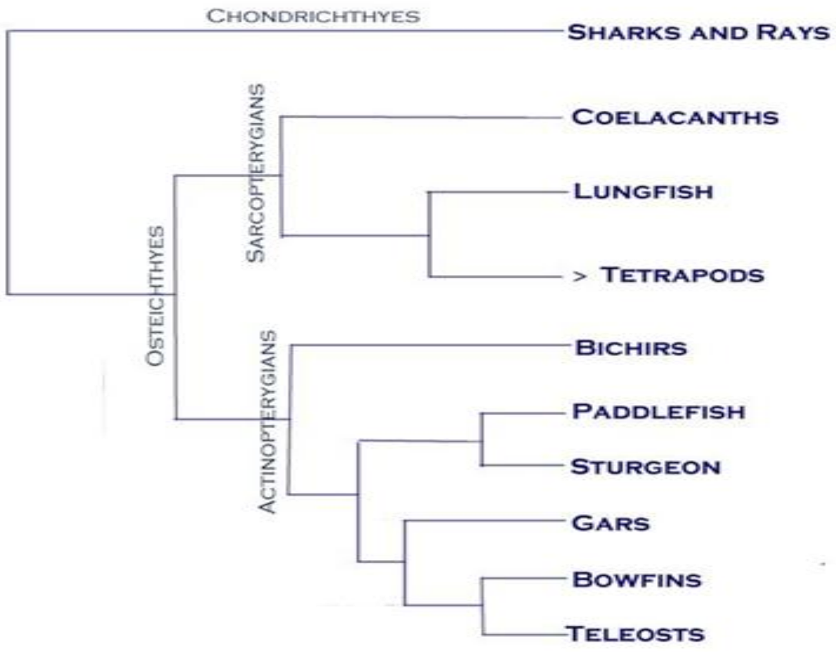

**Fig. 2.Phylogram of all groups of modern fishes showing their evolution from a common ancestor**

Phylogenetic reconstruction is an attempt to discern the ancestral relationship of a set of sequences. It involves the construction of a tree, where the nodes indicate separate evolutionary paths, and the lengths of the branches give an estimate of how distantly related the sequences represented by those branches are.Genes from different species may not have the same phylogenetic history as the species from which those sequences are taken have (although they do, obviously, have the same evolutionary history).Only shared and derived (synapomorphic) characters can be used to clearly establish a phylogenetic relationship.

There are several methods of constructing phylogenetic trees - the most common are:

1. Distance based methods
2. Character based methods

All these methods can only provide estimates of what a phylogenetic tree might look like for a given set of data. Most good methods also provide an indication of how much variation there is in these estimates.

## DISTANCE BASED METHODS:

1. This method is preferred for work with immunological data, frequency data, or data with some impreciseness in its methods. It is very rapid, and easily permits statistical tests e.g. bootstrapping. It derives some measure of similarity or difference between the input sequences. 1. UPGMA: One algorithm for inferring a tree from a distance matrix is a progressive clustering method (much like those used for sequence alignment described above) known as the unweighted pair group method with arithmetic mean (UPGMA) algorithm. This method constructs a tree by identifying the shortest distance (D) in the matrix, clustering those two taxa into a single OTU for use in all subsequent calculations, calculating a new distance matrix, and then repeating these steps.

2. Neighbor Joining (NJ): Many other distance algorithms have been created that attempt to infer trees accurately, even in the face of the vagaries of evolution such as the unequal rates problem outlined in the discussion of UPGMA above. In this case we can make use of an alternative method—neighbor joining. This method resembles the UPGMA clustering method but has some unique properties. Most importantly, it allows for unequal rates of evolution in different branches of the tree. Furthermore, if the distance matrix is

an accurate reflection of the real tree, neighbor joining will always infer the true tree. It corrects several assumptions made in the UPGMA method. It yields an unrooted tree.

## CHARACTER BASED METHODS:

It is a popular method for reconstructing ancestral relationships.

1. Maximum parsimony:It evaluates all possible trees for that given character and infers the number of evolutionary events implied by a particular topology. The most likely tree is then one that requires the minimum number of evolutionary changes needed to explain the observed data. Problems: Most parsimonious tree may not be unique; difficult to make valid statistical statements if there are many steps in a tree; branches with particularly rapid rates of change tend to attract one another, especially when the sequence lengths are small.
2. Maximum likelihood: it is a very slow method and preferred mostly when homoplasies (Convergences of a particular character at a site) are expected to be concentrated in a few sites only, whose identities are known in advance. The method works by estimating, for all nucleotide positions in a sequence, what the probability of having a particular nucleotide at a particular site is, based on whether or not its ancestors had it (and the transition/transversion ratio). These probabilities are summed over the whole sequence, for both branches of a bifurcating tree. The product of the two probabilities gives you the likelihood of the tree up to this point. With more sequences, the estimation is done recursively at every branch point. Since each site evolves independently, the likelihood of the phylogeny can be estimated at every site. This process can only be done in a reasonable amount of time with four sequences. If there are more than four sequences, basic trees can be made for sets of four sequences, and then extra sequences added to the tree and the process of finding the maximum likelihood re-estimated. The order in which the sequences are added and the initial sequence chosen to start the process critically influences the resulting tree. To prevent any bias, the whole process is done multiple times with random choices for the order of the sequences. A majority rule consensus tree is then chosen as the final tree.

## REFERENCES

1. **Atteson K (1997).**"The performance of neighbor-joining algorithms of phylogeny reconstruction", pp. 101–110.*In* Jiang, T., and Lee, D., eds., *Lecture Notes in Computer Science, 1276*, Springer-Verlag, Berlin.COCOON '97.
2. **IP Farias, G Ortí, I Sampaio, H Schneider, A Meyer(2001)**The cytochrome b gene as a phylogenetic marker: the limits of resolution for analyzing rel. J MolEvol 53: 89-103.
3. **Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijin, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J.H., Standen, R.** and **Young, I.G.** (1981). Sequence and organization of the human mitochondrial genome.*Nature 290*: 457-465
4. **Árnason, U.** and**Gullberg, A.** (1993). Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. *J. Mol. Evol. 37*: 312-322
5. **Árnason, U.** and**Johnsson, E.** (1992).The complete mitochondrial DNA sequences of the harbor seal, *Phocavitulina. J. Mol. Evol. 34*: 493-505.
6. **Bernardi, G., Sordino, P., Powers, DA., (1992)** Nucleotide sequence of the 18S ribosomal ribonucleic acid gene from two teleosts and two sharks and their molecular phylogeny. Mol Mar BiolBiotechnol 1:187–194
7. **Bibb, M.J., Van Etten, R.A., Wright, C.T., Walberg, M.W.** and**Calyton, D.A.** (1981). Sequence and organization of mouse mitochondrial DNA. *Cell 26*: 167-180
8. **Brown, W.M., George, M.** and **Wilson, A.C.** (1979).Rapid evolution of animal mitochondrial DNA.*Proc. Natl. Acad. Sci USA 76*: 1967-1971.
9. **Brown, W.M., Prager, E.M., Wang, A.** and **Wilson, A.C.** (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol. 18*: 225-239
10. **Cantatore, P.** and**Saccone, C.** (1987).Organization, structure, and evolution of mammalian mitochondrial genes.*Int. Rev. Cytol. 108*: 149-208
11. **Chang, Y.-S., Huang, F.-C.**and **Lo, T.-B.** (1994). The complete nucleotide sequence and gene organization of carp (*Cyprinuscarpio*) mitochondrial genome. *J. Mol. Evol. 38*: 138-155
12. **Clayton, D.A.** (1982). Replication of animal mitochondrial DNA.*Cell 28*: 693-705.
13. **Clayton, D.A.** (1992). Structure and function of the mitochondrial genome.*J. Inherited Metab. Dis. 15*: 439-447

14. **Gadaleta, G., Pepe, D., De Candia, G., Quagliariello, C., Sbisà, E.** and**Saccone, C.** (1989). The complete nucleotide sequence of the *Rattusnovergicus* mitochondrial genome: cryptical signals revealed by comparative analysis between vertebrates. *J. Mol. Evol. 28*: 497-516
15. **Gyllestein, U., Wharton, D., Josefsson, A.** and **Wilson, A.C.** (1991).Paternal inheritance of mitochondrial DNA in mice.*Nature 352*: 255-257
16. **Hayashi, J.I., Tagashira, Y.** and **Yoshida, M.C.** (1985).Absence of extensive recombination between interspecies and intraspeciesmitochodrialDNA in mammalian cells.*Exp. Cell Res. 160*: 387-395
17. **Johansen, S., Guddal, P.H.** and **Johansen, T.** (1990).Organization of the mitochondrial genome of Atlantic cod, *Gadusmorhua*.*Nucleic Acids Res. 18*: 411-419.
18. **Kondo, R., Satta, Y., Matsuura, E.T., Ishiwa, H., Takahata, N.** and**Chigusa, S.I.** (1990).Incomplete maternal transmission of mitochondrial DNA in *Drosophila.Genetics 126*: 657-663
19. **Kocher, TD., Thomas, WK/, Meyer, A., Edwards, SV/, Paabo, S., Villablanca, FX., Wilson, AC., (1989)** Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. ProcNatlAcadSci USA 86: 6196–6200
20. **Kumazawa, Y.** and **Nishida, M.** (1993).Sequence evolution of mitochondrial tRNA genes and deep-branch animal phylogenetics.*J. Mol. Evol. 37*: 380-398
21. **Martin, AP., Naylor, GJP.,Palumbi, SR., (1992)** Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. Nature 357:153–155
22. **Meyer, A., Wilson, AC., (1990)** Origin of tetrapods inferred from their mitochondrial DNA affiliation to lungfish. J MolEvol 31:359
23. **Meyer, A., Kocher, TD.,Basasibuaki, P., Wilson, AC., (1990)** Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. Nature 247:550–553
24. **Michaels, G.S., Hauswirth, W.W.** and**Laipis, P.J.(1982).**Mitochondrial DNA copy number in bovine oocytes and somatic cells.*Dev. Biol. 94*: 246-251.
25. **Normark, BB., McCune, AR., Harrison, RG., (1991)** Phylogenetic relationships of Neopterygian fishes, inferred from mitochondrial DNA sequences. MolBiolEvol 8:819–834
26. **Robin, E.D.** and **Wong, R.(1988).**Mitochondrial DNA molecules and virtual number of mitochondrial per cell in mammalian cells.*J. Cell. Physiol. 136*: 507-513
27. **Roe, B.A., Ma, D.-P., Wilson, R.K.** and **Wong, F.-H.(1985).**The complete nucleotide sequence of the *Xenopuslaevis*mitochondrial genome. *J. Biol. Chem. 260*: 9759-9774
28. **Stock, DW.,Moberg, KD., Maxson, LR., Whitt, GS., (1991)** A phylogenetic analysis of the 18S ribosomal RNA sequence of the coelacanth Latimeriachalumnae. Environmental Biology of Fishes 32:99–117
29. **Strait, D. S., and Grine, F. E., (2004).**Inferring hominoid and early hominid phylogeny using craniodental characters: the role of fossil taxa. J Hum Evol 47(6):399-452.
30. **Tzeng, C.-S., Hui, C.-F., Shen, S.-C.**and **Huang, P.C.(1992).** The complete nucleotide sequence of the *Crossostomalacustre* mitochondrial genome: conservation and variations among vertebrates. *Nucleic Acids Res. 20*: 4853-4858
31. **Zardoya, R.** and **Meyer, A.** (1996). The complete nucleotide sequence of the mitochondrial genome of the lungfish (*Protopterusdolli*) supports its phylogenetic position as a close relative of land vertebrates. *Genetics 142*: 1249-1263.
32. **Zardoya, R., Garrido-Pertierra, A.** and **Bautista, J.M.(1995).**The complete nucleotide sequence of the mitochondrial DNA genome of the rainbow trout, *Oncorhynchusmykiss*.*J. Mol. Evol. 41*: 942-951