

ANALYSIS OF PHYLOGENETIC RELATIONSHIP AMONG CARANGOIDES SPECIES USING MEGA 6

NIVETHA SARAH EBENEZER¹, JOYCE PRIYAKUMARI C^{2*}

¹Research Assistant, Bioinformatics Infrastructure Facility (BIF), BTIS Net Centre, Department of Zoology, Madras Christian College, Tambaram (East), Chennai-600059, Tamil Nadu, India

²Assistant Professor, Bioinformatics Infrastructure Facility (BIF), BTIS Net Centre, Department of Zoology, Madras Christian College, Tambaram (East), Chennai-600059, Tamil Nadu, India

Abstract - The Phylogenetic analysis of Cytochrome b, mitochondrion (partial) protein sequences of Carangoides species having the length of 380 amino acids was performed using the tool Molecular Evolutionary Genetics Analysis (Mega6). The evolutionary relationship among seven species was determined. A Maximum Likelihood tree with the length of the branch along with a traditional straight branch tree was constructed. Multiple sequence alignment for the protein sequences was performed using Clustal Omega and their Percentage Identity Matrix was calculated showing the identity between the species.

		trevally	
5.	<i>Carangoides oblongus</i>	Coach whip trevally	46 cm
6.	<i>Carangoides orthogrammus</i>	Island trevally	75cm
7.	<i>Carangoides uii</i>	Japanese trevally	40 cm

Key Words: Phylogenetics, *Carangoides*, Carangidae, Maximum likelihood tree, Cytochrome B.

1. INTRODUCTION

Carangoides is a genus of tropical and subtropical marine fishes in the jack family, Carangidae. They are small to large-sized, deep-bodied fish characterized by certain gill raker and jaw morphology, often appearing very similar to jacks in the genus *Caranx* and are widely distributed in all tropical and subtropical regions of the Indian, Atlantic and Pacific oceans, mostly occupying coastal areas, including reefs, bay sand estuaries. They are also distributed along Madagascar, East Africa, Red Sea, Taiwan, and Japan. The genus *Carangoides* was first erected by Pieter Bleeker in 1851 for an unknown taxon and currently containing 21 species [1]. The carangids are categorized under five main sub groups as black pomfrets, queen fishes, trevallies, scads, and pompanos. Seven species of the Carangidae family have been used for the Phylogenetic analysis:

Table -1: Seven species of *Carangoides*

S.NO	SCIENTIFIC NAME	COMMON NAME	SIZE(Length)
1.	<i>Carangoides dinema</i>	Shadow trevally	85cm
2.	<i>Carangoides equula</i>	White fin trevally	37 cm
3.	<i>Carangoides ferdau</i>	Blue trevally	70 cm
4.	<i>Carangoides malabaricus</i>	Malabar	60 cm

These fishes are predatory in nature, consuming a variety of smaller fishes, crustaceans and cephalopods as prey.

Phylogenetic relationships are discovered through the methods of the phylogenetic inference which evaluate heritable traits, such as DNA sequences.

The software called Molecular Evolutionary Genetics Analysis (MEGA) is developed for comparative analysis of protein and nucleotide sequences to infer the molecular evolutionary patterns of genes, genomes, and species over time [2]. This software qualifies the given data and produces a result where the evolutionary relationship among organisms or the history of an individual organism is represented in the form of a tree.

2. MATERIALS AND METHODS

2.1 Retrieval of protein sequences:

The protein Cytochrome B is found in the mitochondria of eukaryotic cells which is the main subunit of cytochrome bc1 and bf6 complexes. It functions in cellular respiration involving electron transport chain for the generation of ATP [3].

Cytochrome B (mitochondrial) protein sequence from all the seven species was retrieved from the National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>) in the FASTA format, a format which represents the protein sequence. The length of the sequence of the seven species was 380 amino acids.

2.2 Multiple Sequence alignment:

Multiple sequence alignment for the seven species was performed using online software known as Clustal Omega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>).

2.3 Construction of Phylogenetic Tree:

The retrieved FASTA sequences were uploaded into the MEGA 6 software. The Maximum-likelihood tree was constructed by computing and analysis of the given data. A traditional straight branch tree was also constructed to illustrate a clear view of evolutionary relationship.

3. RESULTS AND DISCUSSION

Alignment of multiple sequences highlights areas of similarity which may be related with specific features that are more highly conserved than other regions. These regions are used to classify sequences. Multiple sequence alignment is an important step for phylogenetic analysis, which aims to model the alterations that have occurred over time and derive the evolutionary relationships between sequences.

```

YP_009024969.1[C.malabaricus]  MANLRHTPMLLIIVNSLIDLPAPSNISAWNFQSLALCLATQITGLFLAMHTSDIA
AHC57660.1[C.uui]              MANLRHTPMLLIIVNSLIDLPAPSNISAWNFQSLALCLATQITGLFLAMHTSDIA
AHC57661.1[C.dinema]          MANLRHTPMLLIIVNSLIDLPAPSNISAWNFQSLALCLATQITGLFLAMHTSDIA
YP_009108298.1[C.equula]      MANLRHTPMLLIIVNSLIDLPAPSNISAWNFQSLALCLATQITGLFLAMHTSDIA
AHC57658.1[C.ferdau]          MANLRHTPMLLIIVNSLIDLPAPSNISAWNFQSLALCLATQITGLFLAMHTSDIA
AHC57659.1[C.orthogrammus]    MANLRHTPMLLIIVNSLIDLPAPSNISAWNFQSLALCLATQITGLFLAMHTSDIA

YP_009024969.1[C.malabaricus]  TAFTSVAHCIDVNVGILINHWANGSFFFCICVYLIHGRLYVGSVLYEETHTGVVLL
AHC57660.1[C.uui]              TAFTSVAHCIDVNVGILINHWANGSFFFCICVYLIHGRLYVGSVLYEETHTGVVLL
AHC57661.1[C.dinema]          TAFTSVAHCIDVNVGILINHWANGSFFFCICVYLIHGRLYVGSVLYEETHTGVVLL
YP_009108298.1[C.equula]      TAFTSVAHCIDVNVGILINHWANGSFFFCICVYLIHGRLYVGSVLYEETHTGVVLL
AHC57658.1[C.ferdau]          TAFTSVAHCIDVNVGILINHWANGSFFFCICVYLIHGRLYVGSVLYEETHTGVVLL
AHC57659.1[C.orthogrammus]    TAFTSVAHCIDVNVGILINHWANGSFFFCICVYLIHGRLYVGSVLYEETHTGVVLL

YP_009024969.1[C.malabaricus]  LLLMSTAFVGVYVPLNGGIFSGVATVITNLLSAPVYVNTLVQIVGGFSDNATLTFRFA
AHC57660.1[C.uui]              LLLMSTAFVGVYVPLNGGIFSGVATVITNLLSAPVYVNTLVQIVGGFSDNATLTFRFA
AHC57661.1[C.dinema]          LLLMSTAFVGVYVPLNGGIFSGVATVITNLLSAPVYVNTLVQIVGGFSDNATLTFRFA
YP_009108298.1[C.equula]      LLLMSTAFVGVYVPLNGGIFSGVATVITNLLSAPVYVNTLVQIVGGFSDNATLTFRFA
AHC57658.1[C.ferdau]          LLLMSTAFVGVYVPLNGGIFSGVATVITNLLSAPVYVNTLVQIVGGFSDNATLTFRFA
AHC57659.1[C.orthogrammus]    LLLMSTAFVGVYVPLNGGIFSGVATVITNLLSAPVYVNTLVQIVGGFSDNATLTFRFA

YP_009024969.1[C.malabaricus]  FHFLLPFIAAVFIIHLVFLHETGSSNPTGLNSDDIISFHPYFVYDILGFAALLTALA
AHC57660.1[C.uui]              FHFLLPFIAAVFIIHLVFLHETGSSNPTGLNSDDIISFHPYFVYDILGFAALLTALA
AHC57661.1[C.dinema]          FHFLLPFIAAVFIIHLVFLHETGSSNPTGLNSDDIISFHPYFVYDILGFAALLTALA
YP_009108298.1[C.equula]      FHFLLPFIAAVFIIHLVFLHETGSSNPTGLNSDDIISFHPYFVYDILGFAALLTALA
AHC57658.1[C.ferdau]          FHFLLPFIAAVFIIHLVFLHETGSSNPTGLNSDDIISFHPYFVYDILGFAALLTALA
AHC57659.1[C.orthogrammus]    FHFLLPFIAAVFIIHLVFLHETGSSNPTGLNSDDIISFHPYFVYDILGFAALLTALA

YP_009024969.1[C.malabaricus]  SLALFSPNLLGGDFNTPANPLVTFPHIPEVYFVAVAILISPHLGGVALLFSLIV
AHC57660.1[C.uui]              SLALFSPNLLGGDFNTPANPLVTFPHIPEVYFVAVAILISPHLGGVALLFSLIV
AHC57661.1[C.dinema]          SLALFSPNLLGGDFNTPANPLVTFPHIPEVYFVAVAILISPHLGGVALLFSLIV
YP_009108298.1[C.equula]      SLALFSPNLLGGDFNTPANPLVTFPHIPEVYFVAVAILISPHLGGVALLFSLIV
AHC57658.1[C.ferdau]          SLALFSPNLLGGDFNTPANPLVTFPHIPEVYFVAVAILISPHLGGVALLFSLIV
AHC57659.1[C.orthogrammus]    SLALFSPNLLGGDFNTPANPLVTFPHIPEVYFVAVAILISPHLGGVALLFSLIV

YP_009024969.1[C.malabaricus]  LNVVPLHTSQSLTFRFPITQFLFVTLVADVHMLTWGGPVPVHFIIIGQVAVLYFL
AHC57660.1[C.uui]              LNVVPLHTSQSLTFRFPITQFLFVTLVADVHMLTWGGPVPVHFIIIGQVAVLYFL
AHC57661.1[C.dinema]          LNVVPLHTSQSLTFRFPITQFLFVTLVADVHMLTWGGPVPVHFIIIGQVAVLYFL
YP_009108298.1[C.equula]      LNVVPLHTSQSLTFRFPITQFLFVTLVADVHMLTWGGPVPVHFIIIGQVAVLYFL
AHC57658.1[C.ferdau]          LNVVPLHTSQSLTFRFPITQFLFVTLVADVHMLTWGGPVPVHFIIIGQVAVLYFL
AHC57659.1[C.orthogrammus]    LNVVPLHTSQSLTFRFPITQFLFVTLVADVHMLTWGGPVPVHFIIIGQVAVLYFL

YP_009024969.1[C.malabaricus]  LFLVLTPLAGVENHMLENT
AHC57660.1[C.uui]              LFLVLTPLAGVENHMLENT
AHC57661.1[C.dinema]          LFLVLTPLAGVENHMLENT
YP_009108298.1[C.equula]      LFLVLTPLAGVENHMLENT
AHC57658.1[C.ferdau]          LFLVLTPLAGVENHMLENT
AHC57659.1[C.orthogrammus]    LFLVLTPLAGVENHMLENT
    
```

Fig (1). Sequence alignment of Carangoides species using Clustal Omega.

Here, the polar amino acids are marked in green, the non-polar amino acids are marked in red and the electrically charged amino acids (negative & hydrophilic) are marked in blue. The "*" (asterisk) indicates positions having a single, fully conserved region. The ":" (colon) indicates conserved region between groups of strongly similar properties. The "." (period) indicates conserved region between groups of weakly similar properties.

Table (2). Percentage Identity Matrix

Percent Identity Matrix							
1: YP_009024969.1[C.malabaricus]	100.00	98.16	98.42	98.16	97.11	98.16	98.42
2: AHC57660.1[C.uui]	98.16	100.00	98.68	98.42	96.84	97.89	98.16
3: AHC57661.1[C.dinema]	98.42	98.68	100.00	99.21	97.63	98.68	98.95
4: AHC57662.1[C.oblongus]	98.16	98.42	99.21	100.00	97.37	98.42	98.68
5: YP_009108298.1[C.equula]	97.11	96.84	97.63	97.37	100.00	98.95	98.16
6: AHC57658.1[C.ferdau]	98.16	97.89	98.68	98.42	98.95	100.00	99.21
7: AHC57659.1[C.orthogrammus]	98.42	98.16	98.95	98.68	98.16	99.21	100.00

The percentage identity matrix showing the identity between the seven Carangoides species is given in (Table 2). The sequences are most likely to be similar with minute differences in their sequence. This is a matrix system. According to the matrix, C. malabaricus (1) is 98.16% similar to C. uui (2), 98.42% similar to C. dinema (3), 98.16% similar to C. oblongus (4), 97.11% similar to C. equula (5), 98.16% similar to C. ferdau (6), 98.42% similar to C. orthogrammus (7). Correspondingly, the identity matrix percentage is read in the same method for all the given species.

The Cytochrome B sequences from Carangoides dinema (AHC57661), Carangoides equula (YP_009108298), Carangoides ferdau (AHC57658), Carangoides malabaricus (YP_009024969), Carangoides oblongus (AHC57662), Carangoides orthogrammus (AHC57659) and Carangoides uui (AHC57660) having the amino acid length as 380 were computed in MEGA 6 software.

A sequence explorer displays the data of the aligned sequences. It also provides a number of valuable functionalities to explore the statistical attributes of the data. For example, it shows the conserved and variable domain of the sequences.

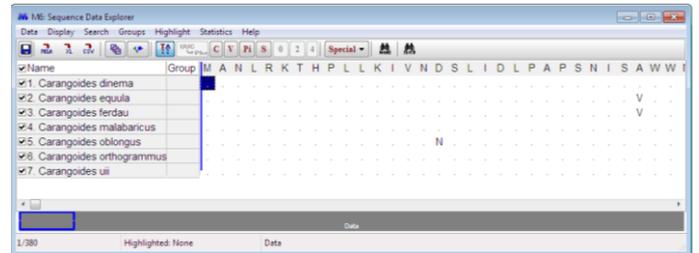


Fig (2). The sequence data explorer.

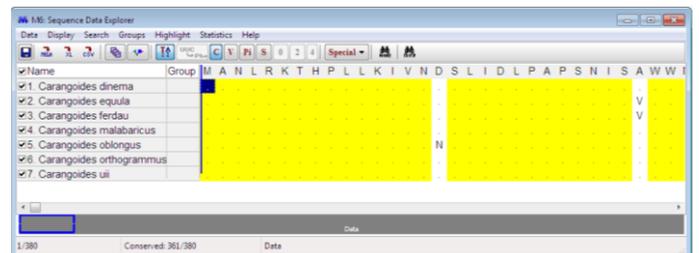


Fig (3). The conserved domain of the sequences.

Conserved domains are regions which stay constant or which are not altered in due course of evolution. These regions are the unique regions of the ancestors that are passed on to their progeny. Here the highlighted regions are

the conserved domains. The conserved region of the sequences together is 361 out of the 380 amino acids.

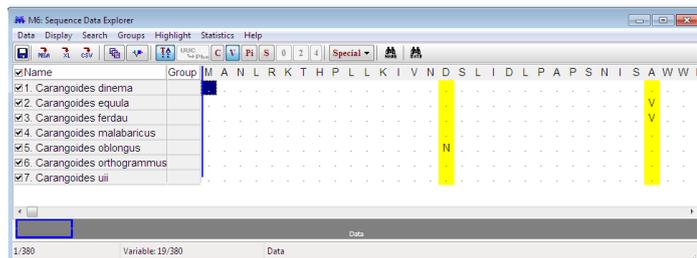


Fig (4). The variable domain of the sequences.

Variable domains are regions in a sequence which may be altered in the due course of evolution. The alteration may be due to genetic mutations. The highlighted regions are the variable domain. The variable region of the sequences together is 19 out of 380 amino acids.

Maximum likelihood is a general statistical method for estimating unknown parameters of a probability model. A parameter is some descriptor of the model. A familiar model might be the normal distribution of a population with two parameters: the mean and variance. In phylogenetics there are many parameters, including rates, differential transformation costs, and, most important, the tree itself.

A tree is a graphical representation showing the evolutionary relationships among organisms. There are individual sources of sequences which are phylogenetically distinct units on the tree known as taxa. The tree is composed of nodes (representing taxa), a point where branches bifurcate. The branching (speciation) indicates the evolutionary relationship among the organisms. Two or more daughter lineages are formed from a single lineage when speciation occurs. Two lineages arising from the same branch point are called as sister taxa. A branch having two or more lineages is called a polytomy. The sister taxa and polytomy do not share an ancestor and they may have split at a specific branch point. A grouping that includes a common ancestor and all its descendants is known as a Clade, which forms a nested hierarchy.

A tree can be distinguished into a rooted, an unrooted and a bifurcating tree. A rooted tree has a single lineage, representing a common ancestor that connects all organisms presented in the phylogenetic tree. An unrooted tree specifies the relationships among organisms, but not their evolutionary paths. A bifurcating tree is one where each ancestral lineage gives rise to two descendent lineages.

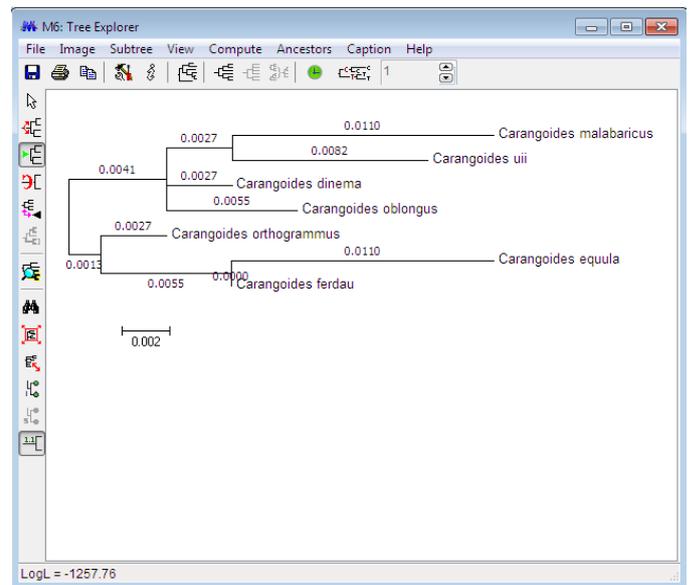


Fig (5). Maximum Likelihood Tree with the length of each of the branch.

The constructed tree is inferred as a bifurcating rooted tree. The horizontal lines or branch, represent evolutionary lineages changing over time. The longer the branch is the greater is the amount of change. The line at the bottom of the figure is the scale for the branches. Here the line segment with the number '0.002' shows the length of branch. This represents the amount of genetic change of 0.002. The given units of branch length are usually nucleotide substitutions per site (the number of changes per 100 nucleotide sites). The branch length of *C. malabaricus* (0.0110) means to expect an average of 0.0110 substitutions per site. This applies to all the branches of the tree. The branch length also represents a measure of support for the node. The maximum support for a node is 1.

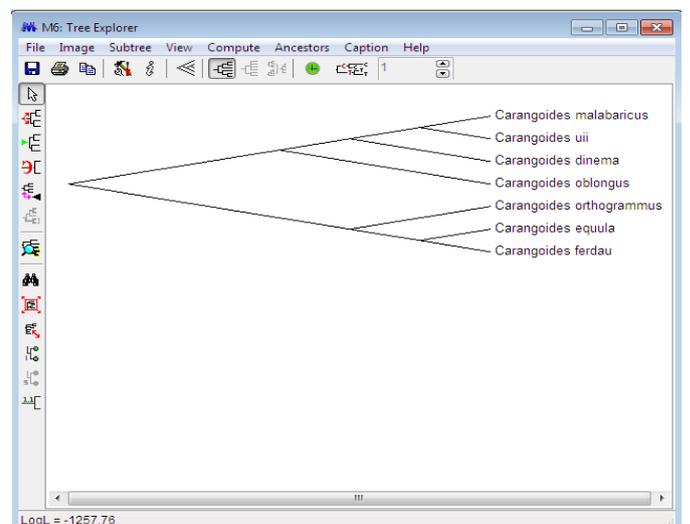


Fig (6). Traditional Straight branch style tree under Maximum Likelihood Tree

The Maximum Likelihood method based on the JTT (Jones DT, Taylor WR, and Thornton JM) matrix-based model was used to infer the evolutionary history. The tree with the highest log likelihood (-1257.76) is shown in (Fig 6). Initial trees for the heuristic search were obtained automatically using a JTT model, and then were obtained by selecting the topology with the highest log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 7 amino acid sequences of the *Carangoides* species. All the positions containing gaps and missing data were eliminated. The total position in the final data set was 380.

The tree infers that *C. equula*, *C. ferdau* and *C. orthogrammus* are closely related. *C. dinema* and *C. oblongus* are branched from different nodes from a common ancestor. This common ancestor also deviated into another branch, which gives rise to *C. malabaricus* and *C. uii*. Understanding the length of the branches, the precise phylogenetic lineage and relationship among the species can be procured.

4. CONCLUSION

In this study, phylogenetic analysis of Cytochrome B protein sequences of *Carangoides* species was accomplished. Having a common ancestor, *C. equula*, *C. ferdau*, *C. orthogrammus* and *C. dinema*, *C. oblongus*, *C. malabaricus*, *C. uii* arise to from two separate lineages (from a node). According to the length of the branches of the maximum likelihood tree, it can be concluded that *C. orthogrammus* might have been evolved much earlier than the other six species. *C. malabaricus* and *C. equula* would have been evolved much later. Multiple sequence alignment performed by using Clustal Omega indicated the identical regions and percentage identity of the sequences. The percentage identity of six species of *Carangoides* was relatively identical. The percentage identity of *C. uii* slightly differed from the other *Carangoides* species.

ACKNOWLEDGEMENT

I thank Biotechnology information system network (BTISNET) centre sponsored by Department of Biotechnology, Ministry of Science and Technology and Government of India for supporting my work.

REFERENCES

[1] Persis M., Chandra Sekhar Reddy A., Rao L. M., Khedkar G. D., Ravinder K, Nasruddin K(2008). COI (cytochrome oxidase-I) sequence based studies of Carangid fishes from Kakinada coast, India. *Molecular Biology Reports*. Mol Biol Rep 36:1733–1740. DOI 10.1007/s11033-008-9375-4

[2] Koichiro Tamura, Glen Stecher, 3 Daniel Peterson, Alan Filipki and Sudhir Kumar(2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Oxford journals – Molecular Biology and Evolution*. 30(12):2725–2729 doi:10.1093/molbev/mst197.

[3] Mahender Singh*, Ashish Gupta and Lakra W.S (2012). In silico 3-D structure prediction of cytochrome b protein of sisorid catfish *Glyptothorax gapang*. *Indian Journal of Biotechnology*. Vol 11, April 2012, pp 156-162.

[4] Min Liab, ZirongHuanga & Zuozhi Chena (2016) ;27(1):378-9. doi:10.3109/19401736.2014.895994. Epub 2014 Mar 11. Characterization of the mitochondrial genome of the Malabar trevally *Carangoides malabaricus* and related phylogenetic analyses.

[5] Pai-Lei Lin and Kwang- Tsao Shao*(1999). A Review of the Carangid Fishes (Family Carangidae) from Taiwan with Descriptions of Four New Records. Institute of Zoology, Academia Sinica, Taipei, Taiwan 115, R.O. C. *Zoological Studies* 38(1): 33-68.

[6] Seishi Kimura, Keiichi Matsuura, Sasanti R.Suharti, and TeguhPeristiwady. Fishes of Bitung. D VIII-I, 21-23; A II-I, 18-19; Scutes 16-26; GR 5-7+15-18 www.kahaku.go.jp/research/db/zoology/Fishes_of_Bitung/data/p076_02b.html

[7] Jones D.T., Taylor W.R., and Thornton J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8: 275-282

[8] Tamura K., Stecher G., Peterson D., Filipki A., and Kumar S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* 30: 2725-2729

[9] Mar'ai D. Ganfornina, Gabriel Gutie'rrez, Michael Bastiani, and Diego Sa'nchez (2000). A Phylogenetic Analysis of the Lipocalin Protein Family. *Oxford journals – Molecular Biology and Evolution*: 17(1):114–126. ISSN: 0737-4038

[10] Barry G. Hall. Building Phylogenetic Trees from Molecular Data with MEGA(2013). *Oxford journals – Molecular Biology and Evolution*. Bellingham Research Institute, Bellingham, Washington. doi:10.1093/molbev/mst012.

[11] F. Williams, P. C. Heemstra, and A. Shameem(1980). Notes on Indo-Pacific Carangid fishes of the genus *Carangoides* Bleeker II. The *Carangoides armatus* group bulletin of Marine Science. 30(1): 13-20, 1980.

[12] Gareth Palidwor, Emmanuel G Reynaud and Miguel A Andrade-Navarro(2006). Taxonomic colouring of phylogenetic trees of protein sequences. *BMC Bioinformatics* 2006, 7:79 doi: 10.1186/1471-2105-7-79.

[13] John S. Gunn. A Revision of Selected Genera of the Family Carangidae (Pisces) from Australian Waters. Division of Fisheries, CSIRO Marine Laboratories, Australia.

[14] "Phylogenetic Trees." Boundless Biology. Boundless, 26 May 2016. <https://www.boundless.com/biology/textbooks/boundless-biology-textbook/phylogenies-and-the-history-of-life-20/organizing-life-on-earth-133/phylogenetic-trees-539-11748/>.

BIOGRAPHIES

Nivetha Sarah Ebenezer completed her B.Sc. degree in Advanced Zoology and Biotechnology from Stella Maris College (2014) and M.Sc. degree in Zoology from Madras Christian College in (2016). Currently she is working as a Research Assistant in Bioinformatics Infrastructure Facility (BIF), of BTIS Net centre at Madras Christian College.



Dr. C. Joyce Priyakumari, an Assistant Professor, is also the co-coordinator of Bioinformatics Infrastructure Facility (BIF) of BTIS Net centre at Madras Christian College. She holds a PhD from University of Madras and currently teaches Molecular biology, Biotechnology and Bioinformatics at Madras Christian College, Chennai.