**Molecular marker** **in Phylogenetic Studies**

A **molecular marker** is a [molecule](https://en.wikipedia.org/wiki/Molecule) contained within a sample taken from an [organism](https://en.wikipedia.org/wiki/Organism) ([biological markers](https://en.wikipedia.org/wiki/Biological_markers)) or other matter. It can be used to reveal certain characteristics about the respective source.

[DNA](https://en.wikipedia.org/wiki/DNA), for example, is a molecular marker containing information about [genetic disorders](https://en.wikipedia.org/wiki/Genetic_disorder), [genealogy](https://en.wikipedia.org/wiki/Genealogy) and the [evolutionary history of life](https://en.wikipedia.org/wiki/Evolutionary_history_of_life). Specific regions of the DNA ([genetic markers](https://en.wikipedia.org/wiki/Genetic_marker)) are used to diagnose the [autosomal recessive](https://en.wikipedia.org/wiki/Dominance_%28genetics%29) genetic disorder [cystic fibrosis](https://en.wikipedia.org/wiki/Cystic_fibrosis), [taxonomic affinity](https://en.wikipedia.org/wiki/Affinity_%28taxonomy%29) ([phylogenetics](https://en.wikipedia.org/wiki/Phylogenetics)) and [identity](https://en.wikipedia.org/wiki/Genetic_identity) ([DNA Barcoding](https://en.wikipedia.org/wiki/DNA_Barcoding)). Further, [life forms](https://en.wikipedia.org/wiki/Life_forms) are known to shed unique chemicals, including [DNA](https://en.wikipedia.org/wiki/DNA), into the [environment](https://en.wikipedia.org/wiki/Natural_environment) as evidence of their presence in a particular location. Other [biological markers](https://en.wikipedia.org/wiki/Biological_markers), like [proteins](https://en.wikipedia.org/wiki/Proteins), are used in [diagnostic](https://en.wikipedia.org/wiki/Molecular_diagnostics) tests for complex [neurodegenerative disorders](https://en.wikipedia.org/wiki/Neurodegeneration), such as [Alzheimer's disease](https://en.wikipedia.org/wiki/Alzheimer%27s_disease). Non-biological molecular markers are also used, for example, in [environmental](https://en.wikipedia.org/wiki/Natural_environment) studies.

Uses of molecular markers in the phylogenetic studies of various organisms have become increasingly important in recent times. This lecture gives an overview of different molecular markers employed by researchers for the purpose of phylogenetic studies. Availability of fast DNA sequencing techniques along with the development of robust statistical analysis methods, provided a new momentum to this field. In this context, utility of different nuclear encoded genes (like 16S rRNA, 5S rRNA, 28S rRNA) mitochondrial (cytochrome oxidase, mitochondrial 12S, cytochrome b, control region) and few chloroplast encoded genes are discussed.

**Nuclear ribosomal genes**

Ribosomal RNA is considered as the best target for studying phylogenetic relationship because, it is universal and is composed of highly conserved as well as variable domains . The ribosomes consist of rRNA and proteins. In all organisms the ribosome consists of two subunits, the small ribosomal subunit (SSU) contains a single RNA species (the 18S rRNA in eukaryotes and the 16S rRNA in others). In Bacteria and Archaea, the large subunit (LSU) contains two rRNA species (the 5S and 23S rRNAs); in most eukaryotes the large subunit contains three RNA species (the 5S, 5.8S and 25S/28S rRNAs). The core structures of the SSU and LSU rRNAs contain 10 and 18 such variable regions, respectively. Moreover, rRNA genes are evolving more slowly than protein encoding genes and are particularly important for the phylogenetic analysis of distantly related species . In particular, secondary-structure models of RNA molecules have been based almost exclusively on comparative sequence analysis.

**16S rRNA**

It was in 1960s that Dubnau et al. observed the conservation in the 16S rRNA gene sequence among Bacillus species . But, it was only after the classic work done by Woese, that these gene sequences were used for bacterial taxonomy . The 16S rRNA gene is conserved, which does not mean that it evolves at a same rate in all organisms. This important property helps researchers to distinguish among different bacterial groups . The 16S rRNA gene is about 1550 bp long and contains both variable and conserved regions with characteristic oligonucleotide signature sequences (unique to a particular phylogenetic group). Using primers of the conserved regions, the in-between variable region can be amplified. This is sufficient to differentiate organisms using statistically valid measurements .As 16S gene is present in all bacteria, one can measure relationships among all bacterial species. Comparing 16S sequences of unknown bacteria with already deposited sequence will assist in marking those bacteria in a particular group . Studying 16S and 23S rRNA are the backbone of bacterial taxonomy, especially for identification of nonculturable bacteria.



**5S rRNA**

Ribosomal 5S RNA, a ~120 nucleotide long RNA, is found in virtually all ribosomes with the exception of mitochondria of some fungi, higher animals and most protists . The nucleotide sequence of 5S rRNA is highly conserved throughout nature and phylogenetic analysis alone provided an initial model for its secondary structure. The primary structure of these rRNA molecules are sufficiently constrained that on the whole they have not changed rapidly in time. In fact, there are reports that 5S rRNA sequence data do not have sufficient resolving power to contribute significantly to our understanding of phylogenetic relationships at any taxonomic level.



**28S rRNA**

Phylogenetic analyses based on molecular sequences must come from genes encoding larger molecules than the 120 bp 5S rRNA . The 28S rRNA gene is about 811 bp in length. 28S rRNA gene sequences for many major metazoan groups have become available in the recent years. Also, efforts to align sequences according to the secondary-structure model for 28S rRNA of these organisms have become commonplaceمألوف for the purpose of phylogenetic analyses.



**Mitochondrial genes (mtDNA)**

Mitochondrial DNA data can be very powerful in resolving species-level phylogenies. The order of genes in the mitochondrion is variable, and they are separated by large regions of noncoding DNA. The mitochondrial genome rearranges itself frequently so that many rearranged forms can occur in the same cell. The use of mtDNA has become increasingly popular in phylogenetics and population genetic studies because of i) developments in methodology for mtDNA isolation, ii) use of restriction enzymes to detect nucleotide differences, iii) the developments of PCR methodologies and iv) applicability of universal primers for amplification of DNA.

**Cytochrome oxidase I/II (COI/II)**

The enzyme cytochrome c oxidase is a very well known protein of electron transport chain and is found in both bacteria and mitochondria. The COI and COII genes code for two of seven polypeptide subunits in the cytochrome c oxidase complex. The COI gene consists of approximately 894 bp COI and COII have been used for species and population analyses of parasitoids and COI has recently been suggested as a potential ‘barcode’ for insect identification in general.

**Mitochondrial 12S**

Mitochondrial 12S rRNA gene sequence analysis is extensively used in molecular taxonomy and phylogeny. Earlier, mitochondrial 12S rRNA gene sequence was used for species determination in wild-life forensic biology. It has been postulated earlier that 12S gene sequences are useful for the determination of moderate to long divergence times. The length of this gene is about 450 bp and it can be amplified by universal primers.

**Cytochrome-b**

 Cytochrome-b gene (~1,143 bp) is reported as the most useful marker in recovering phylogenetic relationships among closely related taxa but can lose resolution at deeper nodes. Although the Cytochrome-b gene has proven useful in recovering phylogenetically useful information at a variety of taxonomic levels, strength of its utility can be lineage-dependent and declines with evolutionary depth.

**Control region for replication of mitochondrial DNA**

 The only major non-coding area of the mtDNA is the control region, typically 1 kb, involved in the regulation and initiation of mtDNA replication and transcription and is responsible for the regulation of heavy (H) and light (L) strand transcription and of H-strand replication.



**Chloroplast genes**

Many plant phylogenetic studies are based on chloroplast DNA (cpDNA). In plants, cpDNA is smallest as compared to mitochondria and nuclear genome. It is assumed to be conserved in its evolution in terms of nucleotide substitution with very little rearrangements which permits the molecule to be used in resolving phylogenetic relationships especially at deep levels of evolution . However, selection of a gene of sufficient length and appropriate substitution rate is a crucial step. Currently used cpDNA genes include rbcL, ndhF, rpl16, matK, atpB and many more.