Journal of Phylogenetics & Evolutionary Biology

Patwardhan et al., J Phylogen Evolution Biol 2014, 2:2
DOI: 10.4172/2329-9002.1000131

Review Article Open Access

Molecular Markers in Phylogenetic Studies-A Review

Anand Patwardhan[1#], Samit Ray[2] and Amit Roy[1#*]

[1]Department of Biotechnology, Siksha-Bhavana, Visva-Bharati University, Santiniketan 731 235, India
[2]Department of Botany, Siksha-Bhavana, Visva-Bharati University, Santiniketan 731 235, India
[#]Equally contributed for the manuscript

Abstract

Uses of molecular markers in the phylogenetic studies of various organisms have become increasingly important in recent times. This review 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 (like rbcL, matK, rpl16) are discussed. Criteria for choosing suitable molecular markers and steps leading to the construction of phylogenetic trees have been discussed. Although widely practised even now, traditional morphology based systems of classification of organisms have some limitations. On the other hand it appears that the use of molecular markers, though relatively recent in popularity and are not free entirely of flaws, can complement the traditional morphology based method for phylogenetic studies.

Keywords: Molecular phylogeny; Phylogenetic tree; Molecular marker; Molecular clock; Bar code of Life

Introduction

Phylogeny is the history of descent of a group of taxa such as species from their common ancestors including the order of branching and sometimes the times of divergence. The term "Phylogeny" is derived from a combination of Greek words. Phylon stand for "tribe" or "clan" or "race" and genesis means "origin" or "source". The term can also be applied to the genealogy of genes derived from a common ancestral gene. In molecular phylogeny, the relationships among organisms or genes are studied by comparing homologues of DNA or protein sequences. Dissimilarities among the sequences indicate genetic divergence as a result of molecular evolution during the course of time. In brief, while classical phylogenetic approach relies on morphological characteristics of an organism, the molecular approaches depend on nucleotide sequences of RNA and DNA and sequences of amino acids of a protein which are determined using modern techniques. By comparing homologous molecules from different organisms it is possible to establish their degree of similarity thereby establishing or revealing a hierarchy of relationship a phylogenetic tree. Both the classical morphology based methods and molecular analysis based methods are of importance as the basic bio-molecular framework of all organisms are similar and morphology of an organism is actually the manifestations of its genome, proteome and transcriptome profiles. A combination of the morphological based methods and molecular analysis based methods thus strengthens the exercise of the determination of phylogenetic relationships of organisms to a great extent.

The job of determination of phylogenetic relationship of various organisms is a difficult one as the living world exhibits unimaginable diversity with respect to its species content. This diversity is not only reflected in phenotypic characters but also in ultra-structural, biochemical and molecular features. Phenotypically similar organisms may have contrasting biochemical and molecular features. A rough estimate of the number of described species is 1.4 to 1.8 million [1,2] of which arthropods, (especially insects), molluscs, and vascular plants account for more than 80%. Still there are millions of species which are unknown and unclassified. The field of taxonomy deals with classification, nomenclature and identification of unknown organisms i.e., the process of determining whether an organism belongs to one of the units defined previously, and if it does not belong to the any of the established taxonomic units, then categorize it as a new taxon. The task of describing, naming and classifying the organism is a part of systematics. Some terminologies related to molecular phylogeny are presented in Box 1.

Since every organism is the result of an evolutionary process, one has to know its evolutionary history to understand and express it in biological terms. For the purpose of determination of evolutionary history, three types of information are necessary. The first one is phenotypic, i.e. the information gained from expressed features including both internal and external morphology, proteins and biochemical markers. The second one is genotypic i.e. the knowledge obtained from the genetic

Cladogram: A phylogenetic tree in which the branch lengths are not proportional to the number of evolutionary changes and thus have no phylogenetic meaning
Homoplasy: Observed sequence similarity that is a result of convergence or parallel evolution, but not direct evolution
Internal transcribed spacers (ITS): The rRNA genes are transcribed as a single transcript separated by ITS, which are subsequently spliced out and serve no further purpose
Monophyletic: The taxa on the phylogenetic tree that are descended from a single common ancestor
Paraphyletic: Includes taxa that are not descendent from a common ancestor
Phylogeny: study of evolutionary relationships between organisms by using treelike diagrams as representations
Polyphyletic: Includes groups that resemble some members outside their groups
Phylogram: a phylogenetic tree in which the branch lengths represent the amount of evolutionary divergence
Outgroup: Taxon or a group of taxa in a phylogenetic tree known to have diverged earlier than the rest of the taxa in the tree and used to determine the position of the root
Synonymous substitution: Nucleotide changes in a protein coding sequence that do not result in amino acid sequence changes, for the encoded protein because of redundancy in the genetic code

Box 1: Important terms related to molecular phylogeny.

***Corresponding author:** Amit Roy, Department of Biotechnology, Visva-Bharati University, Santiniketan 731235, India, Tel: +91-9433144948; E-mail: amit.roy@visva-bharati.ac.in

Received July 27, 2014; **Accepted** August 21, 2014; **Published** August 29, 2014

Citation: Patwardhan A, Ray S, Roy A (2014) Molecular Markers in Phylogenetic Studies – A Review. J Phylogen Evolution Biol 2: 131. doi:10.4172/2329-9002.1000131

material inside the cell. Lastly, when the homologies between DNA and proteins are compared, we get information about the phylogeny of that organism and the knowledge gained can be represented in the graphical form of a phylogenetic tree. It is to be noted, however, that phylogenetic trees have also been constructed in early days, long before the advent of techniques employing molecular markers, from studies on external morphology of organisms by noted evolutionary biologists.

One of the most exciting developments in the past decade has been the application of powerful and ultra rapid nucleic acid sequencing techniques to the problems of phylogenetic studies. Rapid availability of large amounts of sequence data called for developments of robust mathematical and statistical analysis tools for explaining the process of evolution and this acute need ultimately gave rise to the science of molecular systematics. While molecular phylogeny, in a really broad way, may be a domain of the biology, the molecular systematics might be viewed as more of a statistical science in which powerful computation based simulation experiments are used to infer phylogenetic trees from these biological data obtained from a study of molecular markers. The idea of this review is mainly to focus on the molecular markers currently in use today and is divided into three sections; 1) the first section deals with history and general information on molecular phylogeny followed by 2) a section on typical molecular markers (e.g. 16S and 18S rRNA, matK etc.) used for this types of studies and 3) a very brief section on evolutionary tree building methods without which the review will remain incomplete. A general flow chart of various steps involved in studying molecular phylogeny using molecular markers is depicted in Figure 1.

General Information on Molecular Phylogeny

Classical and modern methods of phylogenetic studies

Long time back Aristotle (384-322 B.C.) did extensive morphological and embryological studies to classify marine organisms. Following this, in the 18th century Linnaeus developed binomial system of nomenclature. He not only gave birth to the field of taxonomy but was the first to draw a phylogenetic tree. Later Charles Darwin added the occurrence of two important processes in phylogeny, mainly, branching and subsequent divergence. Early proponents of molecular phylogeny claimed that molecular data were more likely to reflect the true phylogeny than morphological data, chiefly because they reflected gene-level changes, which were thought to be less subject to convergence and parallelism than were morphological traits. This early theory now appears to be inaccurate and molecular data are in fact subject to scores of the same problems that morphological data are. Additionally, in case of unicellular organisms like bacteria morphology, physiology and many other properties are not informative enough to be used as phylogenetic markers. Thus, bacterial classification remained a determinative one, despite the efforts of microbiologists to figure out a natural bacterial classification. Moreover, there are many bacteria that cannot be cultured in the laboratory and their identification solely relies on molecular data. Recent adoption of polyphasic approaches (discussed in brief later) appear to have solved these difficulties.

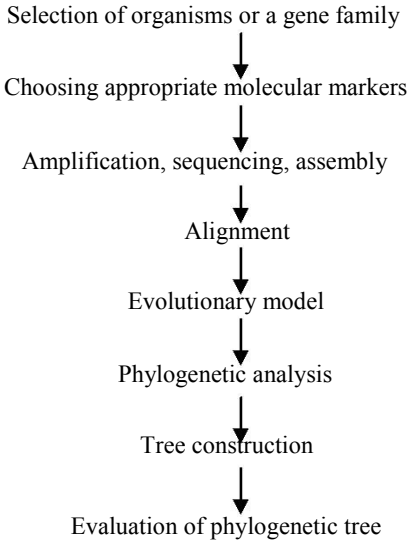

Figure 1: General steps in studying molecular phylogeny.

In recent years molecular phylogeny entered a rapidly expanding area with great improvements in the techniques and analyses of nucleic acid and protein sequencing. Early research using rRNA involved direct reverse transcriptase mediated sequencing of portion of both the small and large subunits of ribosome [3,4]. As rRNA are the major portion of total cellular RNAs, it was relatively easy to obtain enough RNA for sequencing. It is to be noted, however, that sequences generated from direct sequencing of rRNA by reverse transcriptase have been found to be far more more error-prone than DNA sequences generated directly from the nuclear genes encoding ribosomal DNA (rDNA) [5]. In general, the methods utilizing DNA isolation, PCR, automated sequencing and then comparing these DNA or protein sequences are more preferred these days. In summary, molecular phylogenetic studies have been and remains technique driven and as a corollary, dominates the modern taxonomic studies.

Molecular clock and the phylogenetics

Zuckerkandl and Pauling [6] were the first to study amino acid sequences of haemoglobin among different species and their results were remarkable. They found that haemoglobin molecules from horse and human differed by only 18 amino acids; mouse and human haemoglobins differed by 16 amino acids while mouse and horse hemoglobins differed only by 22 residues; but between humans and sharks there were differences in 79 amino acids in this molecule. These important observations seemed to suggest that there is a constant rate of amino acid substitution over time. To explain these results Zuckerkandl and Pauling [6] proposed the so called molecular clock hypothesis. The concept is based on a steady rate of change in DNA sequences over time and provided a basis for dating the time of divergence of lineages. It suggests that these amino acid differences correlate with the evolutionary time scale. As explained above, amino acid differences between mammals are less compared to that between mammals and shark. Thus, a biomolecule was acting like a molecular clock. Further they are distanced from each other in the evolutionary timescale, greater would be the differences in their molecular sequences and vice versa. Similarly the molecular clock hypothesis was used to propose that humans and apes diverged approximately 5 million years ago [7]. Although informative, the hypothesis has been questioned many times because biomolecules are subjected to changes at different rates.

The phylogeny concluded from a single marker gene or protein sequence only reflects evolution of that particular gene. But use of a single marker can lead to interpretation problems, because other genes in the organism may show different rates of evolution or even show different evolutionary history if horizontal gene transfer has taken place. Vertical gene transfer is the normal passage of genes from parent to offspring. Horizontal or lateral gene transfer happens when genes transfer between unrelated organisms, a common phenomenon in bacteria e.g. acquired antibiotic resistance leading to multidrug

resistant bacterial species. There have also been well-known cases of horizontal gene transfers between eukaryotes. Horizontal gene transfer has complicated the determination of phylogenies of organisms. Inconsistencies in phylogeny have been reported among specific groups of organisms depending on the marker genes used to construct evolutionary trees. The only way to determine which genes have been acquired vertically and which one horizontally is to assume that the largest set of genes that have been inherited together have been inherited vertically. This requires analyzing a large number of genes as opposed to studying a single marker gene. So only when one considers the evolution of multiple genes in a genome, one can get more convincing conclusions about the evolutionary status of an organism.

Molecular markers are favoured over morphological data

The underlying fact useful for molecular systematics is that different genes accumulate mutations at different rates. This difference depends on how much change a gene can tolerate without losing its function. For example, histone molecules may become non-functional if some of its amino acids are replaced with different ones. On the other hand internal transcribed spacers (ITS) of ribosomal RNA can still fold properly if many of its nucleotides are changed. Thus, ITS can accumulate mutations more rapidly than histones, reflecting the different functional constraints on their gene product. The advantages of using molecular data is obvious - molecular data are more numerous than fossil records and easier to obtain. There is no sampling bias involved, which helps to correct the gaps in real fossil records. A more clear and robust phylogenetic tree can be constructed with the molecular data. On the other hand parameters for morphological data on many occasions are limited in number and become insufficient to distinguish two organisms at phyla, class, order and family levels. When variation in morphological data become insufficient to distinguish two organisms-at phyla class, order, family etc. levels, analysis of the biomolecules are considered, which are large in number and occur in various forms in organisms. Therefore, the biomolecular markers have become favourite and sometimes the only information available for researchers to reconstruct evolutionary history. The big difference is that there are simply many more molecular characters available, and their interpretation is generally easier. Another advantage of molecular data is that all known life forms are based on nucleic acids and, each nucleotide position, in theory, can be considered as a character and assumed to be independent. The morphological adaptations of an organism, in any case, are mirrored in its biomolecules and vice versa.

Potential of a gene in resolving phylogenetic relationship

The biomolecule based reconstruction of ancient phylogenetic history first requires the discovery and analysis of slowly evolving nucleotide or amino acid sequences. Not all genes or macromolecules are suitable phylogenetic markers and not all marker molecules are useful for the analysis of a given group of organisms. The method of screening molecular sequences for their ability to resolve relationships within a particular group include studies which assess the ability of a gene to recover well-established phylogenetic relationships within clades of similar age and the construction of fossil-based pair wise difference curves, which estimate the rate of potentially informative character changes during the geological interval when a clade underwent phylogenetic divergence [8,9]. For example, to establish the utility of mitochondrial COI and COII (cytochrome oxidase I & II) genes for the purpose of phylogeny studies, Caterino and Sperling used these genes to study phylogeny of *Papilio sp*. and after that they examined the phylogenetic placements of several lineages which have proven difficult in previous studies [10]. Such genes serve as molecular fossils and through comparative analysis of the molecular fossils from a number of related organisms, the evolutionary history of the genes and even the organisms can be revealed.

Properties of ideal marker genes

The properties that should be possessed by an ideal marker are as follows [11]:

(a) A single-copy gene may be more useful than multiple-copy gene; this condition is satisfied by the mitochondrial and nuclear genes; (b) As marker gene sequences are aligned prior to phylogenetic analysis, their alignment should be easy. The length of the same gene can vary among different members of taxa due to insertions or deletions because of which aligning their sequences may be difficult. However, regions with ambiguous alignments can be avoided specifically or secondary structure information may be applied [12]; (c) The substitution rate should be optimum so as to provide enough informative sites. A gene evolving too fast may reach a state of saturation due to multiple substitutions. This problem can be enhanced by base composition bias since this makes it more likely that the second mutation at a particular site will be a reversion to the original state. For protein coding genes it may be the case that the synonymous substitution rate is too high even though very few non-substitutions have occurred; (d) Primers should be available to selectively amplify the marker gene. However, the primer should not be too universal as in that case it would lead to amplification of non-specific genes present as contaminants or contributed by symbionts [13]; (e) A too much of base variation among the taxa, is not preferable which may not reflect the true ancestry [14]. The breakthrough in the study of the phylogeny of prokaryotes was achieved by Carl Woese and co-workers in the seventies [15,16]. They introduced rapid methods of comparative 16S rRNA sequence analysis and phylogenetic tree reconstruction. The results of these efforts provided, for the first time, insight into the phylogeny of prokaryotes and also established the three domains of life, popularly known as– "The Universal Tree of Life" – Archaea (formerly archaebacteria), Bacteria (formerly eubacteria) and Eukarya (eukaryotes) [16,17]. So far, these molecular studies of divergence have drawn on DNA or amino acid sequence data for highly conserved genes, particularly the structural ribosomal genes 18S/16S/5S/28S, the nuclear protein-coding gene elongation factor-1a (EF-1α) and the slowly evolving mitochondrial gene cytochrome c oxidase I (COI), histone H3, U2 snRNA and many more genes which are widely distributed. Some of the very popular markers being used widely in phylogenetic studies are described below in some detail.

Molecular Markers

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 [16,18]. 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 [19]. In particular, secondary-structure models of RNA molecules have been based almost exclusively on comparative sequence analysis [20].

16S rRNA: It was in 1960s that Dubnau et al. observed the conservation in the 16S rRNA gene sequence among *Bacillus* species [21]. But, it was only after the classic work done by Woese, that these gene sequences were used for bacterial taxonomy [16]. 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 [16,20,22]. 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 [23,24]. 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 [22]. Studying 16S and 23S rRNA are the backbone of bacterial taxonomy, especially for identification of non-culturable 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 [25]. The nucleotide sequence of 5S rRNA is highly conserved throughout nature and phylogenetic analysis alone provided an initial model for its secondary structure [15,18]. The primary structure of these rRNA molecules are sufficiently constrained that on the whole they have not changed rapidly in time [18]. Some of the first molecular sequence data available for green algae came from nuclear 5S ribosomal RNA. Troitskii et al. [26] derived complete or partial nucleotide sequences of five different rRNAs from a number of seed plants and discussed the angiosperm origins and early stages of land plant evolution based on phylogenetic dendrograms using the compatibility [27] and parsimony methods from the PHYLIP package [28]. However, the reliability of hypothesis based on this molecule were questioned because the 5S rRNA molecule is only 120 bases long with too few informative sites that can be used in analysis of close relatives. It is a rapidly evolving molecule, so that in the positions that do vary, there are so many substitutions that the number of potentially informative sites is too small to allow reliable analysis for studying ancient divergences. 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 [29].

28S rRNA: Phylogenetic analyses based on molecular sequences must come from genes encoding larger molecules than the 120 bp 5S rRNA [29]. 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 commonplace for the purpose of phylogenetic analyses. For example, *Encarsia,* which is a large genus of minute parasitic wasps, only a few of the species-groups are defined unambiguously on the basis of morphological characters alone. Phylogenetic relationships within this genus still are largely unresolved; only recently attempts have been made to use molecular data to underpin the taxonomy based on morphological characters and to resolve phylogenetic relationships. All molecular studies conducted so far have used the D2 expansion region of the 28S ribosomal RNA; there has been comparatively little information about the suitability of other gene regions to inferring phylogenetic relationships or to defining species limits in this group [30,31].

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 [32].

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/or COII sequences have been applied to phylogenetic problems at a wide range of hierarchical levels in insects, from closely related species to genera and subfamilies, families, and even orders. The COI gene is slowly evolving compared to other protein coding mitochondrial genes and is widely used for estimating molecular phylogenies [33] and is a good performer in recovering an expected tree [34]. So sequencing both the genes represents one of the largest sequence data sets generated for phylogenetic study of any group and also fulfils the putative phylogenetic accuracy. The combination of COI and 12S rRNA is appropriate to distinguish the taxa of interest at different taxonomic level. 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. Zhang and Sota reported that the COI sequence of mitochondrial data had higher sequence divergence than four other nuclear genes, in beetles [35].

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. The 355 bp sequence of this gene was used for identification, phylogenetic relationships and calculation of divergence time of the Indian leopards [36]. Chaolun et al. used the 12S gene to infer the evolutionary history of 28 species of certain coral groups [37]. They found out that phylogenetic analyses using mitochondrial 12S rRNA gene data did not support the current view of phylogeny for this group of corals based upon skeletal morphology and fossil records. Allard and Honeycutt reported that the 12S rRNA gene is not evolving at a higher rate within certain rodent lineages [38].

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. Bradley et al. [39] concluded that, although the Cytochrome-b data contain considerable phylogenetic signal, definition of content and resolution of the phylogeny of genus *Peromyscus* (deer mice) needs other additional information [39]. The patterns of speciation and trait evolution in *Tragopan*, a genus of five Indo-Himalayan bird species, were examined using sequences of the mitochondrial cytochrome b gene (CYB) and its control region (CR) [40].

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. The approximate mutation rate in mtDNA is 10^{-8}/site/year compared to 10^{-9}/site/year in nuclear genes. Most differences between mtDNA sequences are point mutations, with a strong bias for transitions over transversions [32]. Rogaev et al. reported the presence of variable number of tandem repeats (VNTR) in the control region which are characterized by high somatic hypervariability in some mammoth [41]. The evolution of the control region of mammalian mtDNA shows some features such as strong rate heterogeneity among sites, the presence of tandem repeated elements, a high frequency of nucleotides insertion/ deletion, and lineage specificity [42].

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 [43]. 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 (some of them are described below).

rbcL: Ribulose 1, 5-bisphosphate carboxylase/oxygenase (rubisco) is the first enzyme of C3 cycle in plants. It is the most abundant and most important protein on the planet and central to the global carbon cycle [44]. The rbcL gene is located on cp genome as a single copy gene and has an enormous phylogenetic utility. The rbcL gene is ~1428 bp long and is universal to all plants (except in some parasites). It is very convenient to study, easy to align and its secondary structure is known and present in many copies with less insertions and deletions. The rbcL gene encodes the large subunit of rubisco, while the small subunit is encoded by rbcS gene in nucleus. The rbcL gene was one of the first plant genes to be sequenced [45] and is still among the most frequently sequenced segments of plant DNA. This gene has been used widely in systematic studies of land plants, angiosperms in particular [44]. About 500 rbcL sequences were used to address phylogenetic relationships within angiosperms and secondarily among extant seed plants [44]. Although there is length variation between plants and algal genes, their alignment is easy. However many researchers prefer 18S rDNA for sampling than rbcL sequence because of the more rapid rate of evolution in the latter molecule. Although rbcL is conserved and readily alignable across divergent taxa, this molecule exhibits a higher substitution rate than the 18S rDNA. Mc Court et al. tentatively concluded that although rbcL sequences may be inappropriate in phylogenetic studies of ancient branching events (unless and until more thorough taxon sampling is possible), the use of this gene within green algal groups appears to be appropriate [46]. For example, rbcL does not contain enough information for resolving relationships between closely related genera e.g. Hordeum, Triticum, and Aegilops. In such cases the non-coding regions of chloroplast DNA, which are supposed to evolve more rapidly than coding regions are also analyzed. Palmer et al. have shown that the 16S rRNA gene as the most conserved of chloroplast genes followed by 23S rRNA [47]. So, they are more useful phylogenetically at the higher hierarchical levels than the rbcL gene, which codes for a protein.

matK: The matK (maturase) gene is approximately 1500 base pairs (bp), located within the intron of the chloroplast gene trnK (lysine tRNA), and encodes a maturase involved in splicing type II introns from RNA transcripts [48,49]. Recent studies have shown the usefulness of this gene in resolving intergeneric or interspecific relationships among flowering plants. The matK gene is known to have relatively high rates of substitution compared with other genes used in grass systematics, possesses high proportions of transversion mutations, and the 3 section of its coding region has been proven quite useful for constructing phylogenies at the subfamily level in the Poaceae [47]. Sequences from noncoding regions of the chloroplast genome are often used in systematics because such regions tend to evolve relatively rapidly.

ndhF: This gene codes for subunit F of NADP dehydrogenase and is about 1100 bp in length and present in the small single-copy region. Givnish et al. used ndhF sequence variation to reconstruct relationships across 282 taxa representing 78 monocot families [49]. Moreover, they showed that relationships within orders are consistent with those based on rbcL, alone or in combination with atpB and 18S rDNA, and generally better supported and ndhF contributes more than twice as many informative characters as rbcL and nearly as many as rbcL, atpB, and 18S rDNA combined. Kim and Jansen did an extensive sequence comparison of the chloroplast ndhF gene from all major clades of the largest flowering plant family (Asteraceae) and showed that this gene provides ~3 times more phylogenetic information than rbcL [50]. This is because it is substantially longer and evolves twice as fast. The 5' region (1380 bp) of ndhF is very different from the 3' region (855 bp) and is similar to rbcL in both the rate and the pattern of sequence change.

rpl16: Zhang used chloroplast noncoding rpl16 intron (1059 bp) sequences to reconstruct the phylogeny of the grass family [51]. He reported that the rpl16 intron sequence data confirmed three traditional herbaceous bamboo tribes, Streptochaeteae, Anomochloeae, and Phareae, as the most basal lineages in the extant grasses. Zhang also showed that the comparisons of the nucleotide divergence and the genetic distance between the chloroplast noncoding rpl16 intron and the ndhF gene among the major groups of the grass family showed that the rpl16 intron sequences had a lower transition/transversion ratio but higher nucleotide divergence and genetic distance [51]. Earlier studies indicated that noncoding sequences had a much more complicated evolution pattern and more frequent insertion and deletion events than to coding regions [44]. The rpl16 intron sequences show similar results in many reports. Comparison between the ndhF gene and the rpl16 intron sequences done by Zhang indicated that the sequence divergence in the rpl16 intron was 1.40 times of that in the ndhF gene [51]. Some other additional marker genes are mentioned in Table 1.

Phylogenetic Tree Construction Methods

The result of a molecular phylogenetic analysis can be represented in a diagram in the form of a phylogenetic tree. Phylogeny is an abstract phenomenon and it cannot be observed directly. It is something that happened in the past and must be reconstructed using available evidence. By studying a phylogenetic tree it is possible to obtain a quick overall idea about the given species and its relation to other species phylogenetically close to it. As large numbers of potential trees are possible, finding out a tree which perfectly reflects the evolutionary history is very difficult. A tree can also be rooted or unrooted. There is an exponential relationship between the possible number of trees for 'n' taxa, given by, for rooted tree

$$N = (2n-3)!\ 2n^{-2}(n-s)!$$

and for unrooted tree ,

$$N=(2n-5)!/2n^{-3}(n-3)!.$$

Gene	Description	Reference
EF-1α	Elongation factor-1α, Role in protein synthesis.	[52]
rpoA gene	Encoding the alpha subunit of RNA polymerase	[53]
atpB	Encode the beta subunit of ATP synthase	[54]
dnaA	involved in DNA synthesis initiation	[55]
ftsZ	Role in cell division	[56]
gapA	Codes for glyceraldehyde phosphate dehydrogenase	[57]
groEL	Encodes bacterial heat shock protein.	[58]
gltA	Encoding citrate synthase	[59]
ITS	Piece of non-functional RNA situated between structural	[60]
	ribosomal RNAs precursor transcript.	
lux Gene	encode proteins involved in luminescence	[61]
PEPCK	Codes for phosphoenolpyruvate carboxykinase	[62]
pyrH genes	Codes for uridine monophosphate (UMP) kinases	[63]
recA	Role in recombination	[64]
U2 snRNA	Component of the spliceosome	[65]
Wsp gene	Encodes a major cell surface coat protein	[66]
Nuclear H3	Codes for protein which is associated with DNA	[67]
trnH-psbA	Non-coding intergenic spacer region located in plastid genome	[68]
rpoB, rpoC1	Coding region located in plastid genome	[69]

Table 1: List of some other molecular markers used in phylogeny research.

Thus, even for ten taxa under study, there are millions of possible tree topologies available. So, there are various methods to select an optimal tree. The trees can be drawn in different ways such as cladogram or a phylogram. As depicted in Figure 1, a phylogenetic tree construction goes through essentially five steps: a) Selection of molecular markers; b) Performing multiple sequence alignments; c) Choosing an evolutionary model; d) Determining a tree building method and lastly e) Assessing tree reliability [52-70].

Selection of molecular markers

The molecular data can either be obtained from nucleotide or protein sequence data. This often depends upon the closeness of the organisms under study. Nucleotide sequence is preferred while studying closely related organisms, slowly evolving genes are used for widely divergent groups, whereas non-coding mitochondrial DNA is a choice while studying individuals of a population. Protein sequences are more conserved due to codon degeneracy, while the third position of a codon in nucleotide sequence may show variation. Some of the widely used molecular markers preferred by the investigators engaged in molecular phylogenetic research have already been described in section 2.

Multiple Sequence Alignment

Once the markers to be studied have been determined, the DNA sequence of the selected marker genes of the target organism needs to be experimentally determined. For this, total DNA is isolated from the appropriate tissue of the organism. In most instances total cellular DNA may be isolated using many of the well established DNA isolation protocols. The chosen markers are then amplified using the isolated DNA as template and marker specific oligonucleotides as primers by PCR method. For many of the markers discussed in this article, well known universal primers are already described in the literature. Alternatively, the primer can be designed depending upon the specific need of the project. The amplified PCR products are then sequenced. As the DNA sequence of the marker genes are obtained, wholly or in part, the next step is to align the sequence with the DNA sequence of the same markers of closely known species. Multiple alignment is possibly the most critical step in the procedure because it establishes positional correspondence in evolution [70]. Only a successful sequence alignment produces a genealogically related tree. Multiple alignments can be done using various very well known alignment programs like ClustalW, T-coffee, Multialin etc. to mention a few. Secondary structure information may also assist alignment. Praline is one such program which extracts the information of secondary structure for the purpose of alignment. Some programs (Rascal, NorMD, and Gblocks) can improve the alignment by correcting the errors or by removing poorly aligned positions.

Choosing an evolutionary model

The next step is to select a proper substitution model that provides the researcher with ideas of the evolutionary process by taking into account multiple substitution events. However, the observed number of substitutions may not represent the true evolutionary process that actually occurred at the locus of interest. When a mutation is detected as G replaced by T, the nucleotides may have actually undergone a number of transitional steps to become T in the sequence G→A →C→ T. Similarly a back mutation could have taken place also when a mutated nucleotide changed back to the original nucleotide such that A→T→A. Additionally, an identical nucleotide observed in the alignment may be due to parallel mutations; such multiple substitutions and convergence at individual positions obscure the estimation of the true evolutionary distances between the sequences. This effect is known as homoplasy which needs to be corrected for the generation of a true evolutionary tree. To correct homoplasy, statistical models known as substitution models or evolutionary models, are needed to infer the true evolutionary distances between sequences. Following are the two important substitution models [70].

Jukes-Cantor model: Jukes-Cantor model assumes that purines as well as pyrimidines are substituted with equal probability. This model can only analyse reasonably closely related sequences.

Kimura model: In contrast, Kimura two-parameter model [71] assumes that transition mutations should occur more often that transversion. This is a model that takes in to account the differential mutation rates of transitions & transversions and is more realistic. For protein sequences, the evolutionary distances from an alignment can be corrected using a PAM or JTT amino acid substation matrix. Alternatively, protein equivalents of Jukes-Cantor model & Kimura models can be used to correct evolutionary distances.

Tree building method: Next step is the evolutionary tree building. There are several methods available [71] and it is generally recommended to perform exhaustive experiments using one or more

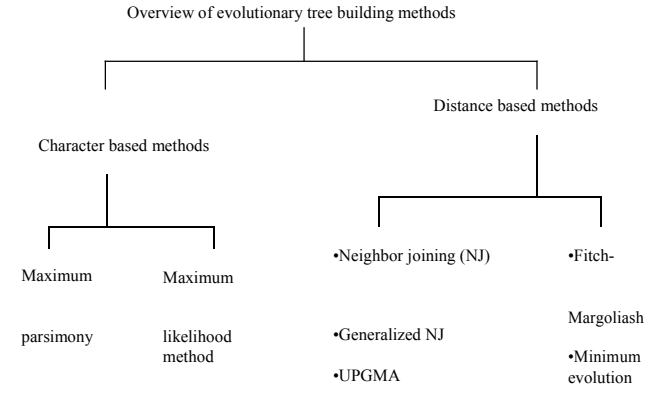

Figure 2: Summary of most commonly used tree building methods.

model. However, it may be a time consuming task when number of taxa increases drastically. Figure 2 shows the summary of different methods which are routinely used. Here we will discuss them in brief as detail explanation of each method is out of scope of this review.

Methods based on characters: Such methods take into account the mutational events accumulated on the sequences and thus avoid loss of information. It easily provides information regarding homoplasy and ancestral states. It produces more accurate trees than the distance based methods. Two most popular character based methods are maximum parsimony and maximum likelihood.

Methods based on distance: A true evolutionary distance between sequences can be calculated from observed distance after correction using different models. They are subdivided as optimality based and clustering based algorithms.

Phylogenetic tree evaluation method

Having constructed the tree, its validity needs to be checked. Different statistical test are used to evaluate the reliability of the constructed tree. Bootstrapping and Jackknifing are employed to check the reliability of the tree while, Kishino-Hasegawa test, Bayesian analysis, and Shimodaira-Hasegawa test are used to confirm whether the tree is better than any other tree. In bootstrapping technique, randomly sized and positioned pieces of sequence form the same part of the molecule are sampled randomly and a new phylogenetic analysis is performed to produce a tree. To determine the robustness of the tree it is generally recommended that a phylogenetic tree should be bootstrapped 500-1000 times, thus making the process time consuming. The bootstrap results are compared to the original approximated tree. Branch point scores around 90% suggest that the predicted tree is accurate. However, controversies can still arise. In Jackknifing half of the data set is subjected to phylogenetic tree construction using the same method as of original. The Bayesian simulation test uses Markov chain MonteCarlo (MCMC) procedure which is very fast and involved thousands of steps of resampling. Kishino-Hasegawa test is especially used for maximum parsimony trees, a t-value is calculated, which is used for evaluation against the t-distribution to see whether the values falls within the significant range (e.g. <0.05),

$$t=Pa-Pt/SD/\sqrt{n}$$

where,

n is the number of informative sites, the degree of freedom is n-1, t is the test statistical value, Pa is the average site-to-site difference between the two trees, SD is the standard deviation, and Pt is the total difference of branch lengths of the two trees. Shimodaira-Hasegawa (SH) test is frequently used for Maximum likelihood trees; it tests the goodness of fit using $\chi 2$ test [70].

DNA Barcode in Animals and Plants

While in case of most animal species cytochrome oxidase (COI) has been described as a relatively accurate system for cost effective species identification purpose, even in the recent past there has not been a generally accepted DNA barcode standard for the plant kingdom as the performance of different loci combinations remains inadequate among different plant families. DNA barcoding, a relatively new term, is defined as a method for identifying species by using short DNA sequences, known as DNA barcodes, to facilitate biodiversity studies and enhance forensic analyses etc. So the researchers designed family specific primers and came closer to accepted phylogeny using this approach. In 2009, a large consortium of researchers, the "Consortium for the Barcode of Life (CBOL) Plant Working Group" proposed portions of two coding regions from the plastid (chloroplast) genome—molecular markers rbcL and matK—as a core barcode for plants, to be supplemented with additional regions as required. This recommendation was accepted by the international Consortium for the Barcode of Life, but with the rider that further sequencing of additional markers should be undertaken. This was driven by concerns that routine use of a third (or even a fourth) marker may be necessary to obtain adequate discriminatory power and to guard against sequencing failure for one of the markers [69,72].

Polyphasic Approach for Bacterial Taxonomy

Over the last 25 years, a much broader range of taxonomic studies of bacteria has gradually replaced the former reliance upon morphological, physiological, and biochemical characterization [73]. The polyphasic taxonomy includes all available phenotypic and genotypic data and integrates them into a system of classification, derived from 16S rRNA sequence analysis. It is conjectured that as more and more parameters become available in future, the polyphasic classification will gain increasing stability. Bacterial taxonomists did not have a clearly set array of rules for species definition, mainly because in unicellular organisms like bacteria morphology, physiology and many other properties are not informative enough to be used as phylogenetic markers. This has a telling effect on bacterial taxonomy problems. This problem is faced in polyphasic taxonomy, which does not depend on a theory, a hypothesis, or a set of rules and presents a pragmatic approach to a consensus type of taxonomy, integrating all available data maximally. In future, polyphasic taxonomy will have to cope with (i) enormous amounts of data, (ii) large numbers of strains, and (iii) data fusion (data aggregation), which will demand efficient and centralized data storage. Thus taxonomic studies will require collaborative efforts by specialized laboratories even more than now is the case [73,74].

Discussion

Although there are large numbers of phylogenetic markers available, the researcher should not be limited only to these genes. In fact, there is a need for developing additional markers for phylogenetic analysis. The number of genes used for phylogenetic analysis over plants, animals and microorganisms should be increased through nuclear genome sequencing and EST (expressed sequence tag) projects. Also, need of markers over large group of organisms is very crucial. Future effort should be directed towards improving the algorithms for various analysis softwares. The power of genes involved with the physiology of organisms such as the cell division (cdc) genes, salt tolerance genes, heat shock genes, homeotic genes, receptor genes etc. to mention a few, should also be explored as they show great homology over a large range of organisms. At the same time, efforts of classical biologists who have been basing their phylogeny analyses on morphological studies of both external and internal features of an organism should be encouraged. In combination with studies using molecular genetic markers and morphology, relatively full proof systems can be devised for the phylogenetic studies of Archaea and Eukarya groups, much in line with the polyphasic approaches described for bacteria.

As time passes more data will become available, more novel organisms will be detected and software development will need to take into account the combination and linking of the different databases. We will also have increasing access to the genome and DNA sequences from many organisms will be available because of the repaid advances in the sequencing technologies. The most challenging task will definitely be

to process this mass of information into a useful classification concept. Discovery of newer molecular markers for the purpose of phylogenetic studies will have to keep pace with the progress in the downstream techniques and analysis procedure as they generate the raw data on the basis of which the analyses are carried out.

Acknowledgements

This work has been supported in parts by grants from CSIR (India), UGC (India) to AR and DBT-HRD (India) grant to the Department of Biotechnology, Visva-Bharati University, India.

References

1. Wilson E O (1994) The diversity of life. Cambridge: Harvard University Press. pp. 205-260.
2. May RM (1990) How many species? Phil Trans R Soc Lond B 330: 293-304.
3. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ. Evolving genes and proteins. Academic Press, New York. pp. 97-166.
4. Sarich VM, Wilson AC (1967) Immunological time scale for hominid evolution. Science 158: 1200-1203.
5. Zimmer EA (1989) Ribosomal RNA phylogenies and flowering plant evolution. In: Fernholm B, Bremer K, Jornvall H. The hierarchy of life. Elsevier, Amsterdam.
6. Buchheim MA, Turmel M, Zimmer EA, Chapman RL (1990) Phylogeny of Chlamydomonas (Chlorophyta) based on cladistic analysis of nuclear 18S rRNA sequence data. J Phycol 26: 689-699.
7. Medlin L, Elwood HJ, Stickel S, Sogin ML (1988) The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene 71: 491-499.
8. Mindell DP, Honeycutt RL (1990) Ribosomal RNA in vertebrates: Evolution and phylogenetic applications. Annu Rev Ecol Syst 21: 541-566.
9. Graybeal A (1994) Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. Syst Biol 43: 174-193.
10. Caterino MS, Sperling FA (1999) Papilio phylogeny based on mitochondrial cytochrome oxidase I and II genes. Mol Phylogenet Evol 11: 122-137.
11. Cruickshank R (2002) Molecular markers for the phylogenetic of mites and ticks. Sys App Acarol 7: 3-14.
12. Kjer KM (1995) Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Mol Phylogenet Evol 4: 314-330.
13. Yli-Mattila T, Paavanen-Huhtala S, Fenton B, Tuovinen T (2000) Species and strain identification of the predatory mite Euseius finlandicus by RAPD-PCR and ITS sequences. Exp Appl Acarol 24: 863-880.
14. Galtier N, Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92: 11317-11321.
15. Fox GE, Woese CR (1975) The architecture of 5S rRNA and its relation to function. J Mol Evol 6: 61-76.
16. Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221-271.
17. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87: 4576-4579.
18. Goerge EF, Pechman CR, Woese CR (1977) Comparative cataloging of 16S ribosomal ribonucleic acid: Molecular approach to prokaryotic systematics. Int J Syst Bacteriol 27: 44-57.
19. Moritz C, Dowling TE, Brown WM (1987) Evolution of Animal mitochondrial DNA-revelence for population biology and systematic. Ann Rev Ecol Syst 18: 269-292.
20. Gutell RR, Gray MW, Schnare MN (1993) A compilation of large subunit (23S and 23S-like) ribosomal RNA structures: 1993. Nucleic Acids Res 21: 3055-3074.
21. Dubnau D, Smith I, Morell P, Marmur J (1965) Gene conservation in Bacillus species. I. Conserved genetic and nucleic acid base sequence homologies. Proc Natl Acad Sci U S A 54: 491-498.
22. Clarridge JE (2004) Impact of 16S rRNA Gene Sequence Analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 17: 840–862.
23. Pfister P, Hobbie S, Vicens Q, Böttger EC, Westhof E (2003) The molecular basis for A-site mutations conferring aminoglycoside resistance: relationship between ribosomal susceptibility and X-ray crystal structures. Chembiochem 4: 1078-1088.
24. Chen K, Neimark H, Rumore P, Steinman CR (1989) Broad range DNA probes for detecting and amplifying eubacterial nucleic acids. FEMS Microbiol Lett 48: 19-24.
25. Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. Science 283: 1476-1481.
26. Troitsky AV, Melekhovets YuF, Rakhimova GM, Bobrova VK, Valiejo-Roman KM, et al. (1991) Angiosperm origin and early stages of seed plant evolution deduced from rRNA sequence comparisons. J Mol Evol 32: 253-261.
27. Estabrook GF (1983) The causes of character incompatibility. In: Felsenstein J Numerical taxonomy. Springer-Verlag, Berlin. pp. 279-295.
28. Felsenstein J (1989) Phylip-Phylogeny inference package version 3.2
29. Steele KP, Holsinger KE, Jansen RK, Taylor DW (1991) Assessing the reliability of 5S rRNA sequence data for phylogenetic analysis in green plants. Mol Biol Evol 8: 240-248.
30. Manzari S, Polaszek A, Belshaw R, Quicke DL (2002) Morphometric and molecular analysis of the Encarsia inaron species-group (Hymenoptera: Aphelinidae), parasitoids of whiteflies (Hemiptera: Aleyrodidae). Bull Entomol Res 92: 165-176.
31. Schmidt S, Driver F, Barro PD (2006) The phylogenetic characteristics of three different 28S rRNA gene regions in Encarsia (Insecta, Hymenoptera, Aphelinidae). Org Divers Evol 6: 127-139.
32. Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18: 225-239.
33. Russo CA, Takezaki N, Nei M (1996) Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol Biol Evol 13: 525-536.
34. Zardoya R, Meyer A (1996) Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol Biol Evol 13: 933-942.
35. Zhang AB, Sota T (2007) Nuclear gene sequences resolve species phylogeny and mitochondrial introgression in Leptocarabus beetles showing trans-species polymorphisms. Mol Phyl Evol 45: 534-546.
36. Pandey PK, Dhotre DP, Dharne MS, Khadse AN, Hiremath UI, et al. (2007) Evaluation of mitochondrial 12S rRNA gene in the identification of Panthera pardus fusca (Meyer, 1794) from field-collected scat samples in the Western Ghats, Maharashtra, India. Curr Sci 92: 1129-1133.
37. Chen CA, Wallace CC, Wolstenholme J (2002) Analysis of the mitochondrial 12S rRNA gene supports a two-clade hypothesis of the evolutionary history of scleractinian corals. Mol Phylogenet Evol 23: 137-149.
38. Allard MW, Honeycutt RL (1992) Nucleotide sequence variation in the mitochondrial 12S rRNA gene and the phylogeny of African mole-rats (Rodentia: Bathyergidae). Mol Biol Evol 9: 27-40.
39. Bradley RD, Durish ND, Rogers DS, Miller JR, Engstrom MD, et al. (2007) toward a molecular phylogeny for peromyscus: evidence from mitochondrial cytochrome-b sequences. J Mammal 88: 1146-1159.
40. Randi E, Lucchini V, Hennache A, Kimball RT, Braun EL, Ligon JD (2001) Evolution of the mitochondrial DNA control-region and cytochrome b genes, and the inference of phylogenetic relationships in the avian genus Lophura (Galliformes). Mol Phyl Evol 19: 187–201.
41. Rogaev EI, Moliaka YK, Malyarchuk BA, Kondrashov FA, Derenko MV, et al. (2006) Complete mitochondrial genome and phylogeny of Pleistocene mammoth Mammuthus primigenius. PLoS Biol 4: e73.
42. Pesole , Gissi C, De Chirico A, Saccone C (1999) Nucleotide substitution rate of mammalian mitochondrial genomes. J Mol Evol 48: 427-434.
43. Zurawski G, Clegg MT (1987) Evolution of higher plant chloroplast DNA-encoded genes: Implications for structure– function and phylogenetic studies. Annu Rev Plant Physiol 38: 391–418.
44. Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, et al. (1993) Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene rbcL. Ann Miss Bot Gard 80: 528-580.

45. Zurawski G, Clegg MT, Brown AH (1984) The Nature of Nucleotide Sequence Divergence between Barley and Maize Chloroplast DNA. Genetics 106: 735-749.

46. McCourt RM, Karol KG, Guerlesquin M, Feist M, et al. (1996) Phylogeny of extant genera in the family Characeae (Charales, Charophyceae) based on rbcL sequence and morphology. Am J Bot 83: 125-131.

47. Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR (1988) Chloroplast DNA variation and plant phylogeny. Ann Miss Bot Gard 75: 1180-1206.

48. Liang H, Hili KW (1996) Application of the matK gene sequences to grass systematics. Canadian J Bot 74: 125-134.

49. Givnish TJ, Pires JC, Graham SW, McPherson MA, Prince LM, et al. (2006) Phylogenetic relationships of monocots based on the highly informative plastid gene ndhF: Evidence for widespread concerted convergence. Aliso 22: 28-51.

50. Kim KJ, Jansen RK (1995) ndhF sequence evolution and the major clades in the sunflower family. Proc Natl Acad Sci U S A 92: 10379-10383.

51. Zhang W (2000) Phylogeny of the grass family (Poaceae) from rpl16 intron sequence data. Mol Phylogenet Evol 15: 135-146.

52. Friendlander TP, Regier JC, Mitter C (1994) Phylogenetic information content of five nuclear gene sequences in animals: Initial assessment of character sets from concordance and divergence studies. Syst Biol 43: 511-525.

53. Fox MG, Sorhannus UM (2003) RpoA: a useful gene for phylogenetic analysis in diatoms. J Eukaryot Microbiol 50: 471-475.

54. Hoot SB, Douglas AW (1998) Phylogeny of the Proteaceae based on atpB and atpB-rbcL intergenic spacer region sequences. Australian Syst Bot 11: 301 – 320.

55. Mitra S, Stallions DR (1976) The role of Escherichia coli dna A gene and its integrative suppression in M13 coliphage DNA synthesis. Eur J Biochem 67: 37-45.

56. Tsagkarakou A, Guillemaud T, Rousset F (1996) Molecular identification of a Wolbachia endosymbiont in a Tetranychus urticae strain (Acari: Tetranychidae). Insect Mol Biol 5: 217-221.

57. Lawrence JG, Ochman H, Hartl DL (1991) Molecular and evolutionary relationships among enteric bacteria. J Gen Microbiol 137: 1911-1921.

58. Georgopoulos C, Ang D (1990) The Escherichia coli groE chaperonins. Semin Cell Biol 1: 19-25.

59. Casiraghi M, Bordenstein SR, Baldo L, Lo N, Beninati T, et al. (2005) Phylogeny of Wolbachia pipientis based on gltA, groEL and ftsZ gene sequences: clustering of arthropod and nematode symbionts in the F supergroup, and evidence for further diversity in the Wolbachia tree. Microbiology 151: 4015-4022.

60. Persson C (2000) Phylogeny of the Neotropical Alibertia group (Rubiaceae), with emphasis on the genus Alibertia, inferred from ITS and 5S ribosomal DNA sequences. Am J Bot 87: 1018-1028.

61. Urbanczyk H, Ast JC, Kaeding AJ, Oliver JD, Dunlap PV (2008) Phylogenetic analysis of the incidence of lux gene horizontal transfer in Vibrionaceae. J Bacteriol 190: 3494-3504.

62. Leys R, Cooper SJB, Schwarz MP (2002) Molecular phylogeny and historical biogeography of the large carpenter bees, genus Xylocopa (Hymenoptera: Apidae). Biol J Linn Soc 77: 249 – 266.

63. Sakamoto H, Landais S, Evrin C, Laurent-Winter C, Bârzu O, et al. (2004) Structure-function relationships of UMP kinases from pyrH mutants of Gram-negative bacteria. Microbiology 150: 2153-2159.

64. Stine OC, Sozhamannan S, Gou Q, Zheng S, Morris JG Jr, et al. (2000) Phylogeny of Vibrio cholerae based on recA sequence. Infect Immun 68: 7180-7185.

65. Edgecombe GD (2000) Arthropod cladistics: combined analysis of histone H3 and U2 snRNA sequences and morphology. Cladistics 16: 155–203.

66. Van Meer MM, Witteveldt J, Stouthamer R (1999) Phylogeny of the arthropod endosymbiont Wolbachia based on the wsp gene. Insect Mol Biol 8: 399-408.

67. Kjer MK, Carle FL, Litman J, Ware J (2006) A molecular phylogeny of Hexapoda. Arthropod Syst Phyl 64: 35-44.

68. Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS One 2: e508.

69. Chase (2007) A proposal for a standardised protocol to barcode all land plants. Taxon 56(2), 295 - 299.

70. Xiong J (2006) Essential Bioinformatics. Cambridge University Press. pp. 127 – 168.

71. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120.

72. Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, et al. (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60: 407-438.

73. Gillis M, Vandamme P, De Vos P, Swings, J and Kersters K (2005) Polyphasic taxonomy. 43-48.

74. Hollingsworth PM (2011) Refining the DNA barcode for land plants. Proc Natl Acad Sci U S A 108: 19451-19452.