Textbook of Phylogenetics

Joshua Krieger

First Edition, 2012

ISBN 978-81-323-4723-1

Published by:
The English Press
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

Introduction to Phylogenetics

In biology, **phylogenetics** is the study of evolutionary relatedness among various groups of organisms (for example, species or populations), which is discovered through molecular sequencing data and morphological data matrices. The term *phylogenetics* is of Greek origin from the terms *phyle/phylon* (φυλή/φῦλον), meaning "tribe, race" and *genetikos* (γενετικός), meaning "relative to birth" from *genesis* (γένεσις, "birth"). Taxonomy, the classification, identification and naming of organisms, has been richly informed by phylogenetics but remains methodologically and logically distinct.

The fields overlap however in the science of phylogenetic systematics – often called "cladism" or "cladistics" – where only phylogenetic trees are used to delimit taxa, which represent groups of lineage-connected individuals. In biological systematics as a whole, phylogenetic analyses have become essential in researching the evolutionary tree of life.

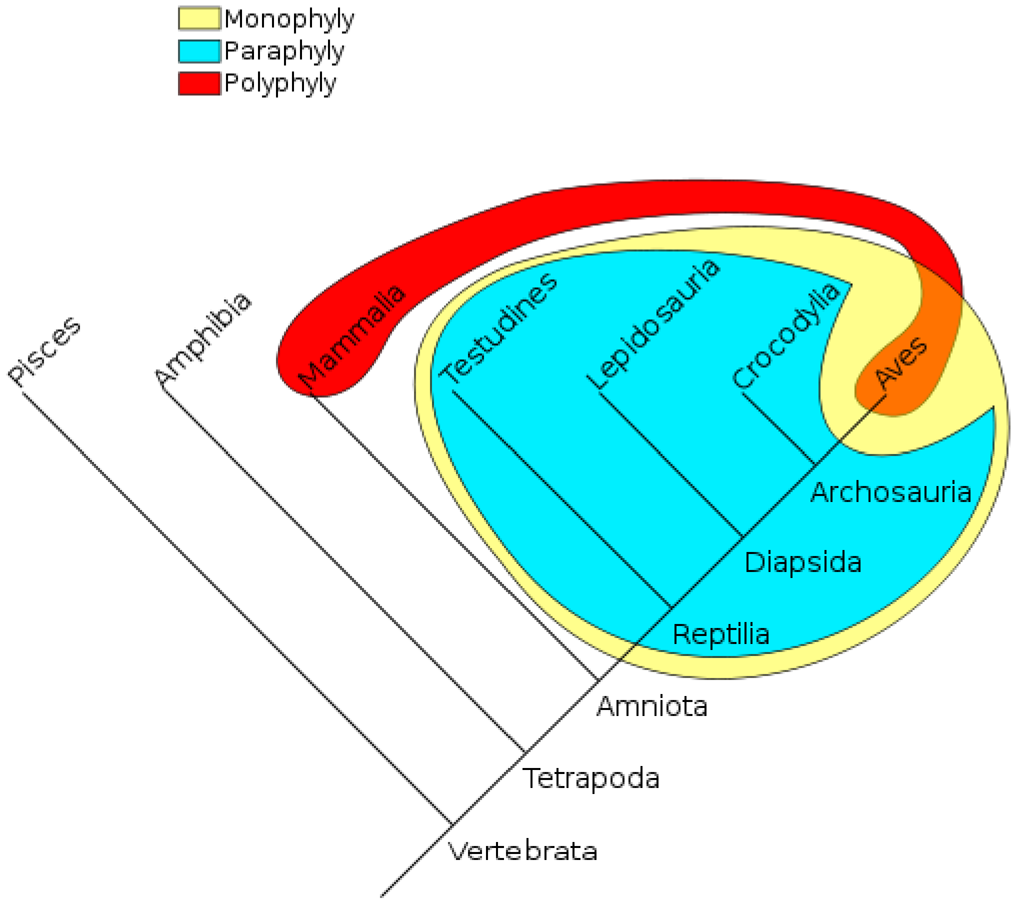
Construction of a phylogenetic tree

Evolution is regarded as a branching process, whereby populations are altered over time and may speciate into separate branches, hybridize together, or terminate by extinction. This may be visualized in a phylogenetic tree.

The problem posed by phylogenetics is that genetic data are only available for living taxa and the fossil records (osteometric data) contains less data and more-ambiguous morphological characters. A phylogenetic tree represents a hypothesis of the order in which evolutionary events are assumed to have occurred.

Cladistics is the current method of choice to infer phylogenetic trees. The most commonly-used methods to infer phylogenies include parsimony, maximum likelihood and MCMC-based Bayesian inference. Phenetics, popular in the mid-20th century but now largely obsolete, uses distance matrix-based methods to construct trees based on overall similarity, which is often assumed to approximate phylogenetic relationships. All methods depend upon an implicit or explicit mathematical model describing the evolution of characters observed in the species included and are usually used for molecular phylogeny, wherein the characters are aligned nucleotide or amino acid sequences.

Grouping of organisms

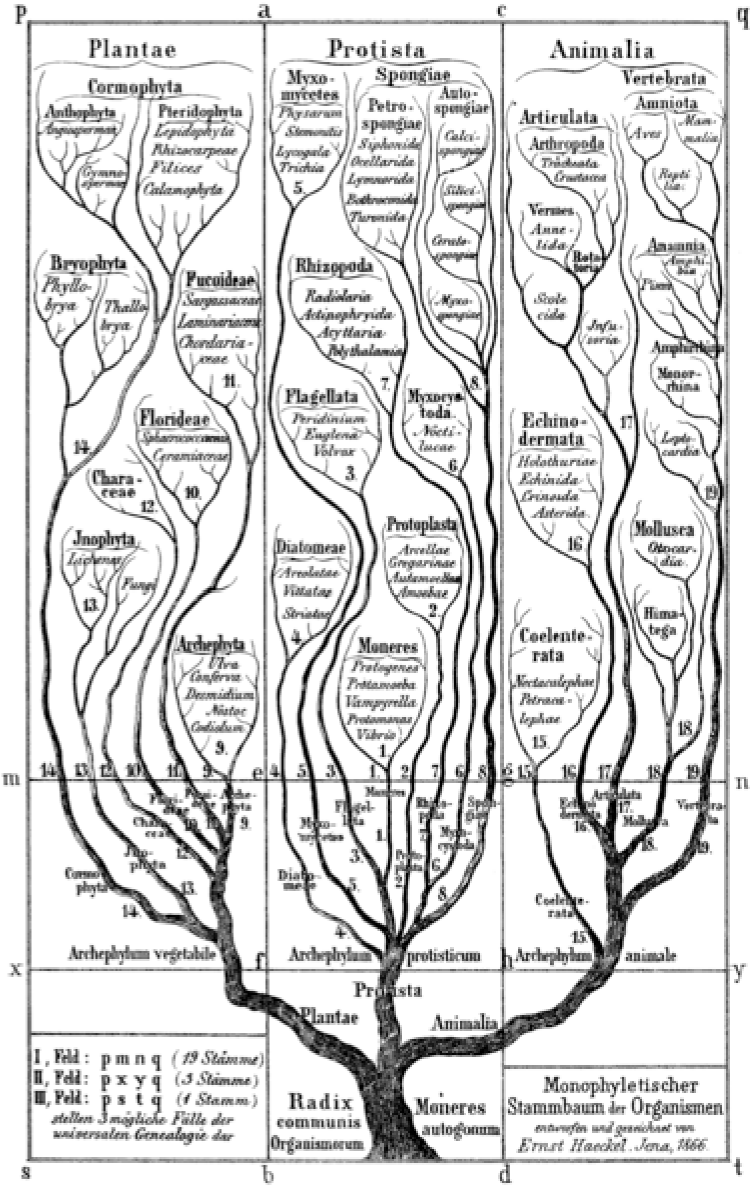


Phylogenetic groups, or *taxa*, can be monophyletic, paraphyletic, or polyphyletic

There are some terms that describe the nature of a grouping in such trees. For instance, all birds and reptiles are believed to have descended from a single common ancestor, so this taxonomic grouping (yellow in the diagram below) is called monophyletic. "Modern reptile" (cyan in the diagram) is a grouping that contains a common ancestor, but does not contain all descendants of that ancestor (birds are excluded).

This is an example of a paraphyletic group. A grouping such as warm-blooded animals would include only mammals and birds (red/orange in the diagram) and is called polyphyletic because the members of this grouping do not include the most recent common ancestor.

Ernst Haeckel's recapitulation theory



Genealogical tree suggested by Haeckel (1866)

During the late 19th century, Ernst Haeckel's recapitulation theory, or biogenetic law, was widely accepted. This theory was often expressed as "ontogeny recapitulates phylogeny", i.e. the development of an organism exactly mirrors the evolutionary development of the species. Haeckel's early version of this hypothesis [that the embryo mirrors *adult* evolutionary ancestors] has since been rejected and the hypothesis amended as the embryo's development mirroring *embryos* of its evolutionary ancestors. He was accused by five professors of falsifying his images of embryos. Most modern biologists recognize numerous connections between ontogeny and phylogeny, explain them using evolutionary theory, or view them as supporting evidence for that theory. Donald I.

Williamson suggested that larvae and embryos represented adults in other taxa that have been transferred by hybridization (the larval transfer theory). However, Williamson's views do not represent mainstream thought in molecular biology and there is a significant body of evidence against the larval transfer theory.

Gene transfer

In general, organisms can inherit genes in two ways: vertical gene transfer and horizontal gene transfer. Vertical gene transfer is the passage of genes from parent to offspring and horizontal gene transfer or lateral gene transfer occurs when genes jump between unrelated organisms, a common phenomenon in prokaryotes.

Horizontal gene transfer has complicated the determination of phylogenies of organisms and inconsistencies in phylogeny have been reported among specific groups of organisms depending on the genes used to construct evolutionary trees.

Carl Woese came up with the three-domain theory of life (eubacteria, archaea and eukaryotes) based on his discovery that the genes encoding ribosomal RNA are ancient and distributed over all lineages of life with little or no horizontal gene transfer. Therefore, rRNAs are commonly recommended as molecular clocks for reconstructing phylogenies.

This has been particularly useful for the phylogeny of microorganisms, to which the species concept does not apply and which are too morphologically simple to be classified based on phenotypic traits.

Taxon sampling and phylogenetic signal

Owing to the development of advanced sequencing techniques in molecular biology, it has become feasible to gather large amounts of data (DNA or amino acid sequences) to infer phylogenetic hypotheses. For example, it is not rare to find studies with character matrices based on whole mitochondrial genomes (~16,000 nucleotides, in many animals). However, it has been proposed that it is more important to increase the number of taxa in the matrix than to increase the number of characters, because the more taxa the more robust is the resulting phylogenetic tree.

This may be partly due to the breaking up of long branches. It has been argued that this is an important reason to incorporate data from fossils into phylogenies where possible. Of course, phylogenetic data that include fossil taxa are generally based on morphology, rather than DNA data. Using simulations, Derrick Zwickl and David Hillis found that increasing taxon sampling in phylogenetic inference has a positive effect on the accuracy of phylogenetic analyses.

Another important factor that affects the accuracy of tree reconstruction is whether the data analyzed actually contain a useful phylogenetic signal, a term that is used generally to denote whether related organisms tend to resemble each other with respect to their genetic material or phenotypic traits. Ultimately, however, there is no way to measure whether a particular phylogenetic hypothesis is accurate or not, unless the "true" relationships among the taxa being examined are already known. The best result an empirical systematist can hope to attain is a tree with branches well-supported by the available evidence.

Importance of missing data

In general, the more data that is available when constructing a tree, the more accurate and reliable the resulting tree will be. Missing data is no less detrimental than simply having less data, although its impact is greatest when most of the missing data is in a small number of taxa. The fewer characters that have missing data, the better; concentrating the missing data across a small number of character states produces a more robust tree.

Role of fossils

Because many morphological characters involve embryological or soft-tissue characters that cannot be fossilized and the interpretation of fossils is more ambiguous than living taxa, it is sometimes difficult to incorporate fossil data into phylogenies. However, despite these limitations, the inclusion of fossils is invaluable, as they can provide information in sparse areas of trees, breaking up long branches and constraining intermediate character states; thus, fossil taxa contribute as much to tree resolution as modern taxa.

Molecular phylogenies can reveal rates of diversification, but in order to track rates of origination, extinction and patterns in diversification, fossil data must be incorporated. Molecular techniques assume a constant rate of diversification, which is rarely likely to be true; in some (but by no means all) cases, the assumptions inherent in interpreting the fossil record (e.g. a complete and unbiased record) are closer to being true than the assumption of a constant rate, making fossil insights more accurate than molecular reconstructions.

Homoplasy weighting

Certain characters are more likely to be evolved convergently than others; logically, such characters should be given less weight in the reconstruction of a tree. Unfortunately the only objective way to determine convergence is by the construction of a tree – a somewhat circular method. Even so, weighting homoplasious characters does indeed lead to better-supported trees. Further refinement can be brought by weighting changes in one direction higher than changes in another; for instance, the presence of thoracic wings almost guarantees placement among the pterygote insects, although because wings are often lost secondarily, their absence does not exclude a taxon from the group.

Chapter- 2

Molecular Phylogenetics

Molecular phylogenetics, also known as **molecular systematics** (a term likely discouraged to avoid confusion with molecular-biological system/structure-activity relationship), is the use of the structure of molecules to gain information on an organism's evolutionary relationships. The result of a molecular phylogenetic analysis is expressed in a phylogenetic tree.

History of molecular phylogenetics

The history of molecular evolution starts in the early 20th century with "comparative biochemistry", but the field of molecular evolution came into its own in the 1960s and 1970s, following the rise of molecular biology. The advent of protein sequencing allowed molecular biologists to create phylogenies based sequence comparison and to use the differences between homologous sequences as a molecular clock to estimate the time since the last common ancestor. In the late 1960s, the neutral theory of molecular evolution provided a theoretical basis for the molecular clock, though both the clock and the neutral theory were controversial, since most evolutionary biologists held strongly to panselectionism, with natural selection as the only important cause of evolutionary change. After the 1970s, nucleic acid sequencing allowed molecular evolution to reach beyond proteins to highly conserved ribosomal RNA sequences, the foundation of a reconceptualization of the early history of life.

Early history

Before the rise of molecular biology in the 1950s and 1960s, a small number of biologists had explored the possibilities of using biochemical differences between species to study evolution. Ernest Baldwin worked extensively on comparative biochemistry beginning in the 1930s and Marcel Florkin pioneered techniques for constructing phylogenies based on molecular and biochemical characters in the 1940s. However, it was not until the 1950s that biologists developed techniques for producing biochemical data for the quantitative study of molecular evolution.

The first molecular systematics research was based on immunological assays and protein "fingerprinting" methods. Alan Boyden—building on immunological methods of G. H. F.

Nuttall—developed new techniques beginning in 1954 and in the early 1960s Curtis Williams and Morris Goodman used immunological comparisons to study primate phylogeny. Others, such as Linus Pauling and his students, applied newly developed combinations of electrophoresis and paper chromatography to proteins subject to partial digestion by digestive enzymes to create unique two-dimensional patterns, allowing fine-grained comparisons of homologous proteins.

Beginning in the 1950s, a few naturalists also experimented with molecular approaches—notably Ernst Mayr and Charles Sibley. While Mayr quickly soured on paper chromatography, Sibley successfully applied electrophoresis to egg-white proteins to sort out problems in bird taxonomy, soon supplemented that with DNA hybridization techniques—the beginning of a long career built on molecular systematics.

While such early biochemical techniques found grudging acceptance in the evolutionary biology community, for the most part they did not impact the main theoretical problems of evolution and population genetics. This would change as molecular biology shed more light on the physical and chemical nature of genes.

Genetic load, the classical/balance controversy and the measurement of heterozygosity

At the time that molecular biology was coming into its own in the 1950s, there was a long-running debate—the classical/balance controversy—over the causes of heterosis, the increase in fitness observed when inbred lines are crossed. In 1950, James F. Crow offered two different explanations (later dubbed the *classical* and *balance* positions) based the paradox first articulated by J. B. S. Haldane in 1937: the effect of deleterious mutations on the average fitness of a population depends only on the rate of mutations (not the degree of harm caused by each mutation) because more-harmful mutations are eliminated more quickly by natural selection, while less-harmful mutations remain in the population longer. H. J. Muller dubbed this "genetic load".

Muller, motivated by his concern about the effects of radiation on human populations, argued that heterosis is primarily the result of deleterious homozygous recessive alleles, the effects of which are masked when separate lines are crossed—this was the *dominance hypothesis*, part of what Dobzhansky labeled the *classical position*. Thus, ionizing radiation and the resulting mutations produce considerable genetic load even if death or disease does not occur in the exposed generation and in the absence of mutation natural selection will gradually increase the level of homozygosity. Bruce Wallace, working with J. C. King, used the *overdominance hypothesis* to develop the *balance position*, which left a larger place for overdominance (where the heterozygous state of a gene is more fit than the homozygous states). In that case, heterosis is simply the result of the increased expression of heterozygote advantage. If overdominant loci are common, then a high level of heterozygosity would result from natural selection and mutation-induced radiation may in fact facilitate an increase in fitness due to overdominance. (This was also the view of Dobzhansky.)

Debate continued through 1950s, gradually becoming a central focus of population genetics. A 1958 study of *Drosophila* by Wallace suggested that radiation-induced mutations *increased* the viability of previously homozygous flies, providing evidence for heterozygote advantage and the balance position; Wallace estimated that 50% of loci in natural *Drosophila* populations were heterozygous. Motoo Kimura's subsequent mathematical analyses reinforced what Crow had suggested in 1950: that even if overdominant loci are rare, they could be responsible for a disproportionate amount of genetic variability. Accordingly, Kimura and his mentor Crow came down on the side of the classical position. Further collaboration between Crow and Kimura led to the infinite alleles model, which could be used to calculate the number of different alleles expected in a population, based on population size, mutation rate and whether the mutant alleles were neutral, overdominant, or deleterious. Thus, the infinite alleles model offered a potential way to decide between the classical and balance positions, if accurate values for the level of heterozygosity could be found.

By the mid-1960s, the techniques of biochemistry and molecular biology—in particular, electrophoresis—provided a way to measure the level of heterozygosity in natural populations: a possible means to resolve the classical/balance controversy. In 1963, Jack L. Hubby published an electrophoresis study of protein variation in *Drosophila*; soon after, Hubby began collaborating with Richard Lewontin to apply Hubby's method to the classical/balance controversy by measuring the proportion of heterozygous loci in natural populations. Their two landmark papers, published in 1966, established a significant level of heterozygosity for *Drosophila* (12%, on average). However, these findings proved difficult to interpret. Most population geneticists (including Hubby and Lewontin) rejected the possibility of widespread neutral mutations; explanations that did not involve selection were anathema to mainstream evolutionary biology. Hubby and Lewontin also ruled out heterozygote advantage as the main cause because of the segregation load it would entail, though critics argued that the findings actually fit well with overdominance hypothesis.

Protein sequences and the molecular clock

While evolutionary biologists were tentatively branching out into molecular biology, molecular biologists were rapidly turning their attention toward evolution.

After developing the fundamentals of protein sequencing with insulin between 1951 and 1955, Frederick Sanger and his colleagues had published a limited interspecies comparison of the insulin sequence in 1956. Francis Crick, Charles Sibley and others recognized the potential for using biological sequences to construct phylogenies, though few such sequences were yet available. By the early 1960s, techniques for protein sequencing had advanced to the point that direct comparison of homologous amino acid sequences was feasible. In 1961, Emanuel Margoliash and his collaborators completed the sequence for horse cytochrome c (a longer and more widely distributed protein than insulin), followed in short order by a number of other species.

In 1962, Linus Pauling and Emile Zuckerkandl proposed using the number of differences between homologous protein sequences to estimate the time since divergence, an idea Zuckerkandl had conceived around 1960 or 1961. This began with Pauling's long-time research focus, hemoglobin, which was being sequenced by Walter Schroeder; the sequences not only supported the accepted vertebrate phylogeny, but also the hypothesis (first proposed in 1957) that the different globin chains within a single organism could also be traced to a common ancestral protein. Between 1962 and 1965, Pauling and Zuckerkandl refined and elaborated this idea, which they dubbed the molecular clock and Emil L. Smith and Emanuel Margoliash expanded the analysis to cytochrome c. Early molecular clock calculations agreed fairly well with established divergence times based on paleontological evidence. However, the essential idea of the molecular clock—that individual proteins evolve at a regular rate independent of a species' morphological evolution—was extremely provocative (as Pauling and Zuckerkandl intended it to be).

The "molecular wars"

From the early 1960s, molecular biology was increasingly seen as a threat to the traditional core of evolutionary biology. Established evolutionary biologists—particularly Ernst Mayr, Theodosius Dobzhansky and G. G. Simpson, three of the founders of the modern evolutionary synthesis of the 1930s and 1940s—were extremely skeptical of molecular approaches, especially when it came to the connection (or lack thereof) to natural selection. Molecular evolution in general—and the molecular clock in particular—offered little basis for exploring evolutionary causation. According to the molecular clock hypothesis, proteins evolved essentially independently of the environmentally determined forces of selection; this was sharply at odds with the panselectionism prevalent at the time. Moreover, Pauling, Zuckerkandl and other molecular biologists were increasingly bold in asserting the significance of "informational macromolecules" (DNA, RNA and proteins) for *all* biological processes, including evolution. The struggle between evolutionary biologists and molecular biologists—with each group holding up their discipline as the center of biology as a whole—was later dubbed the "molecular wars" by Edward O. Wilson, who experienced firsthand the domination of his biology department by young molecular biologists in the late 1950s and the 1960s.

In 1961, Mayr began arguing for a clear distinction between *functional biology* (which considered proximate causes and asked "how" questions) and *evolutionary biology* (which considered ultimate causes and asked "why" questions) He argued that both disciplines and individual scientists could be classified on either the *functional* or *evolutionary* side and that the two approaches to biology were complementary. Mayr, Dobzhansky, Simpson and others used this distinction to argue for the continued relevance of organismal biology, which was rapidly losing ground to molecular biology and related disciplines in the competition for funding and university support. It was in that context that Dobzhansky first published his famous statement, "nothing in biology makes sense except in the light of evolution", in a 1964 paper affirming the importance of organismal biology in the face of the molecular threat; Dobzhansky characterized the molecular disciplines as "Cartesian" (reductionist) and organismal disciplines as "Darwinian".

Mayr and Simpson attended many of the early conferences where molecular evolution was discussed, critiquing what they saw as the overly simplistic approaches of the molecular clock. The molecular clock, based on uniform rates of genetic change driven by random mutations and drift, seemed incompatible with the varying rates of evolution and environmentally-driven adaptive processes (such as adaptive radiation) that were among the key developments of the evolutionary synthesis. At the 1962 Wenner-Gren conference, the 1964 Colloquium on the Evolution of Blood Proteins in Bruges, Belgium and the 1964 Conference on Evolving Genes and Proteins at Rutgers University, they engaged directly with the molecular biologists and biochemists, hoping to maintain the central place of Darwinian explanations in evolution as its study spread to new fields.

Gene-centered view of evolution

Though not directly related to molecular evolution, the mid-1960s also saw the rise of the gene-centered view of evolution, spurred by George C. Williams's *Adaptation and Natural Selection* (1966). Debate over units of selection, particularly the controversy over group selection, led to increased focus on individual genes (rather than whole organisms or populations) as the theoretical basis for evolution. However, the increased focus on genes did not mean a focus on molecular evolution; in fact, the adaptationism promoted by Williams and other evolutionary theories further marginalized the apparently non-adaptive changes studied by molecular evolutionists.

The neutral theory of molecular evolution

The intellectual threat of molecular evolution became more explicit in 1968, when Motoo Kimura introduced the neutral theory of molecular evolution. Based on the available molecular clock studies (of hemoglobin from a wide variety of mammals, cytochrome c from mammals and birds and triosephosphate dehydrogenase from rabbits and cows), Kimura (assisted by Tomoko Ohta) calculated an average rate of DNA substitution of one base pair change per 300 base pairs (encoding 100 amino acids) per 28 million years. For mammal genomes, this indicated a substitution rate of one every 1.8 years, which would produce an unsustainably high genetic load unless the preponderance of substitutions was selectively neutral. Kimura argued that neutral mutations occur very frequently, a conclusion compatible with the results of the electrophoretic studies of protein heterozygosity. Kimura also applied his earlier mathematical work on genetic drift to explain how neutral mutations could come to fixation, even in the absence of natural selection; he soon convinced James F. Crow of the potential power of neutral alleles and genetic drift as well.

Kimura's theory—described only briefly in a letter to *Nature*—was followed shortly after with a more substantial analysis by Jack L. King and Thomas H. Jukes—who titled their first paper on the subject "non-Darwinian evolution". Though King and Jukes produced much lower estimates of substitution rates and the resulting genetic load in the case of non-neutral changes, they agreed that neutral mutations driven by genetic drift were both real and significant. The fairly constant rates of evolution observed for individual proteins was not easily explained without invoking neutral substitutions (though G. G. Simpson and Emil Smith had tried). Jukes and King also found a strong correlation between the frequency of amino acids and the number different of codons for each; this pointed to amino acid sequences as largely the product of random genetic drift.

King and Jukes' paper, especially with the provocative title, was seen as a direct challenge to mainstream neo-Darwinism and it brought molecular evolution and the neutral theory to the center of evolutionary biology. It provided a mechanism for the molecular clock and a theoretical basis for exploring deeper issues of molecular evolution, such as the relationship between rate of evolution and functional importance. The rise of the neutral theory marked synthesis of evolutionary biology and molecular biology—though an incomplete one.

With their work on firmer theoretical footing, in 1971 Emile Zuckerkandl and other molecular evolutionists founded the *Journal of Molecular Evolution*.

The neutralist-selectionist debate and near-neutrality

The critical responses to the neutral theory that soon appeared marked the beginning of the *neutralist-selectionist debate*. In short, selectionists viewed natural selection as the primary or only cause of evolution, even at the molecular level, while neutralists held that neutral mutations were widespread and that genetic drift was a crucial factor in the evolution of proteins. Kimura became the most prominent defender of the neutral theory—which would be his main focus for the rest of his career. With Ohta, he refocused his arguments on the rate at which drift could fix new mutations in finite populations, the significance of constant protein evolution rates and the functional constraints on protein evolution that biochemists and molecular biologists had described. Though Kimura had initially developed the neutral theory partly as an outgrowth of the *classical position* within the classical/balance controversy (predicting high genetic load as a consequence of non-neutral mutations), he gradually deemphasized his original argument that segregational load would be impossibly high without neutral mutations (which many selectionists and even fellow neutralists King and Jukes, rejected).

From the 1970s through the early 1980s, both selectionists and neutralists could explain the observed high levels of heterozygosity in natural populations, by assuming different values for unknown parameters. Early in the debate, Kimura's student Tomoko Ohta focused on the interaction between natural selection and genetic drift, which was significant for mutations that were not strictly neutral, but nearly so. In such cases, selection would compete with drift: most slightly deleterious mutations would be eliminated by natural selection or chance; some would move to fixation through drift. The behavior of this type of mutation, described by an equation that combined the mathematics of the neutral theory with classical models, became the basis of Ohta's nearly neutral theory of molecular evolution.

In 1973, Ohta published a short letter in *Nature* suggesting that a wide variety of molecular evidence supported the theory that most mutation events at the molecular level are slightly deleterious rather than strictly neutral. Molecular evolutionists were finding that while rates of protein evolution (consistent with the molecular clock) were fairly independent of generation time, rates of noncoding DNA divergence were inversely proportional to generation time. Noting that population size is generally inversely proportional to generation time, Tomoko Ohta proposed that most amino acid substitutions are slightly deleterious while noncoding DNA substitutions are more neutral. In this case, the faster rate of neutral evolution in proteins expected in small populations (due to genetic drift) is offset by longer generation times (and vice versa), but in large populations with short generation times, noncoding DNA evolves faster while protein evolution is retarded by selection (which is more significant than drift for large populations).

Between then and the early 1990s, many studies of molecular evolution used a "shift model" in which the negative effect on the fitness of a population due to deleterious mutations shifts back to an original value when a mutation reaches fixation. In the early 1990s, Ohta developed a "fixed model" that included both beneficial and deleterious mutations, so that no artificial "shift" of overall population fitness was necessary. According to Ohta, however, the nearly neutral theory largely fell out of favor in the late 1980s, because the mathematically simpler neutral theory for the widespread molecular systematics research that flourished after the advent of rapid DNA sequencing. As more detailed systematics studies started to compare the evolution of genome regions subject to strong selection versus weaker selection in the 1990s, the nearly neutral theory and the interaction between selection and drift have once again become an important focus of research.

Microbial phylogeny

While early work in molecular evolution focused on readily sequenced proteins and relatively recent evolutionary history, by the late 1960s some molecular biologists were pushing further toward the base of the tree of life by studying highly conserved nucleic acid sequences. Carl Woese, a molecular biologist whose earlier work was on the genetic code and its origin, began using small subunit ribosomal RNA to reclassify bacteria by genetic (rather than morphological) similarity. Work proceeded slowly at first, but accelerated as new sequencing methods were developed in the 1970s and 1980s. By 1977, Woese and George Fox announced that some bacteria, such as methanogens, lacked the rRNA units that Woese's phylogenetic studies were based on; they argued that these organisms were actually distinct enough from conventional bacteria and the so-called higher organisms to form their own kingdom, which they called archaebacteria. Though controversial at first (and challenged again in the late 1990s), Woese's work became the basis of the modern three-domain system of Archaea, Bacteria and Eukarya (replacing the five-domain system that had emerged in the 1960s).

Work on microbial phylogeny also brought molecular evolution closer to cell biology and origin of life research. The differences between archaea pointed to the importance of

RNA in the early history of life. In his work with the genetic code, Woese had suggested RNA-based life had preceded the current forms of DNA-based life, as had several others before him—an idea that Walter Gilbert would later call the "RNA world". In many cases, genomics research in the 1990s produced phylogenies contradicting the rRNA-based results, leading to the recognition of widespread lateral gene transfer across distinct taxa. Combined with the probable endosymbiotic origin of organelle-filled eukarya, this pointed to a far more complex picture of the origin and early history of life, one which might not be describable in the traditional terms of common ancestry.

Techniques and applications

Every living organism contains DNA, RNA and proteins. Closely related organisms generally have a high degree of agreement in the molecular structure of these substances, while the molecules of organisms distantly related usually show a pattern of dissimilarity. Conserved sequences, such as mitochondrial DNA, are expected to accumulate mutations over time and assuming a constant rate of mutation provide a molecular clock for dating divergence. Molecular phylogeny uses such data to build a "relationship tree" that shows the probable evolution of various organisms. Not until recent decades, however, has it been possible to isolate and identify these molecular structures.

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

Theoretical background

Early attempts at molecular systematics were also termed as chemotaxonomy and made use of proteins, enzymes, carbohydrates and other molecules which were separated and characterized using techniques such as chromatography. These have been largely replaced in recent times by DNA sequencing which produces the exact sequences of nucleotides or *bases* in either DNA or RNA segments extracted using different techniques. These are generally considered superior for evolutionary studies since the actions of evolution are ultimately reflected in the genetic sequences. At present it is still a long and expensive process to sequence the entire DNA of an organism (its genome) and this has been done for only a few species. However it is quite feasible to determine the sequence of a defined area of a particular chromosome. Typical molecular systematic analyses require the sequencing of around 1000 base pairs. At any location within such a sequence, the bases found in a given position may vary between organisms. The particular sequence found in a given organism is referred to as its haplotype. In principle, since there are four base types, with 1000 base pairs, we could have 4^{1000} distinct haplotypes. However, for organisms within a particular species or in a group of related species, it has been found empirically that only a minority of sites show any variation at all and most of the variations that are found are correlated, so that the number of distinct haplotypes that are found is relatively small.

In a molecular systematic analysis, the haplotypes are determined for a defined area of genetic material; ideally a substantial sample of individuals of the target species or other taxon are used however many current studies are based on single individuals. Haplotypes of individuals of closely related, but supposedly different, taxa are also determined. Finally, haplotypes from a smaller number of individuals from a definitely different taxon are determined: these are referred to as an *out group*. The base sequences for the haplotypes are then compared. In the simplest case, the difference between two haplotypes is assessed by counting the number of locations where they have different bases: this is referred to as the number of *substitutions* (other kinds of differences between haplotypes can also occur, for example the *insertion* of a section of nucleic acid in one haplotype that is not present in another). Usually the difference between organisms is re-expressed as a *percentage divergence*, by dividing the number of substitutions by the number of base pairs analysed: the hope is that this measure will be independent of the location and length of the section of DNA that is sequenced.

An older and superseded approach was to determine the divergences between the genotypes of individuals by DNA-DNA hybridisation. The advantage claimed for using hybridisation rather than gene sequencing was that it was based on the entire genotype, rather than on particular sections of DNA. Modern sequence comparison techniques overcome this objection by the use of multiple sequences.

Once the divergences between all pairs of samples have been determined, the resulting triangular matrix of differences is submitted to some form of statistical cluster analysis and the resulting dendrogram is examined in order to see whether the samples cluster in the way that would be expected from current ideas about the taxonomy of the group, or not. Any group of haplotypes that are all more similar to one another than any of them is to any other haplotype may be said to constitute a clade. Statistical techniques such as bootstrapping and jackknifing help in providing reliability estimates for the positions of haplotypes within the evolutionary trees.

Limitations of molecular systematics

Molecular systematics is an essentially cladistic approach: it assumes that classification must correspond to phylogenetic descent and that all valid taxa must be monophyletic.

Molecular systematics often uses the molecular clock assumption that quantitative similarity of genotype is a sufficient measure of the recency of genetic divergence. Particularly in relation to speciation, this assumption could be wrong if either some genotypic modification acted to prevent interbreeding between two groups of organisms, or genetic modification proceeded at different rates in different subgroups of the organisms.

In animals, it is often convenient to use mitochondrial DNA for molecular systematic analysis. However, because in mammals mitochondria are inherited only from the mother, this is not fully satisfactory, because inheritance in the paternal line might not be detected.

Chapter- 3

Microbial Phylogenetics and Computational Phylogenetics

Microbial phylogenetics

Microbial phylogenetics is the study of the evolutionary relatedness among various groups of microorganisms. The molecular approach to microbial phylogenetic analysis, pioneered by Carl Woese in the 1970s and leading to the three-domain model (Archaea, Bacteria, Eucaryota), revolutionized our thinking about evolution in the microbial world. Phylogenetic analysis plays a central role in microbiology and the emerging fields of comparative genomics and phylogenomics require substantial knowledge and understanding of phylogenetic analysis and computational methods.

Historical overview

When at the end of the 19th century information began to accumulate about the diversity within the bacterial world, scientists started to include the bacteria in phylogenetic schemes to explain how life on Earth may have developed. Some of the early phylogenetic trees of the prokaryote world were morphology-based; others were based on the then-current ideas on the presumed conditions on our planet at the time that life first developed. Around 1950 many leading microbiologists had become pessimistic with respect to the possibility of ever reconstructing bacterial phylogeny. The concept of the prokaryote-eukaryote dichotomy did little to clarify phylogenetic relationships. The developing technology of nucleic acid sequencing, together with the recognition that sequences of building blocks in informational macromolecules (nucleic acids, proteins) can be used as 'molecular clocks' that contain historical information, led to the development of the three-domain model (Archaea - Bacteria - Eucaryota) in the late 1970s, primarily based on small subunit ribosomal RNA sequence comparisons. The information currently accumulating from complete genome sequences of an ever increasing number of prokaryotes are now leading to further modifications of our views on microbial phylogeny.

Methods and programs

The purpose of phylogenetic analysis is to understand the past evolutionary path of organisms. Even though we will never know for certain the true phylogeny of any organism, phylogenetic analysis provides best assumptions, thereby providing a framework for various disciplines in microbiology. Due to the technological innovation of modern molecular biology and the rapid advancement in computational science, accurate inference of the phylogeny of a gene or organism seems possible in the near future. There has been a flood of nucleic acid sequence information, bioinformatic tools and phylogenetic inference methods in public domain databases, literature and World Wide Web space. Phylogenetic analysis has long played a central role in basic microbiology, for example in taxonomy and ecology. In addition, more recently emerging fields of microbiology, including comparative genomics and phylogenomics, require substantial knowledge and understanding of phylogenetic analysis and computational skills to handle the large-scale data involved. Methods of phylogenetic analysis and relevant computer software tools lend accuracy, efficiency and availability to the task.

There are four steps in general phylogenetic analysis of molecular sequences: (i) selection of a suitable molecule or molecules (phylogenetic marker), (ii) acquisition of molecular sequences, (iii) multiple sequence alignment (MSA) and (iv) phylogenetic treeing and evaluation. The first step of phylogenetic analysis is to choose a suitable homologous part of the genomes to be compared. Mechanisms of molecular evolution include mutations, duplication of genes, reorganization of genomes and genetic exchanges such as recombination, reassortment and lateral gene transfer. Although all of this information can be used to infer phylogenetic relationships of genes or organisms, information on mutations, including substitution, insertion and deletion, is most frequently used in phylogeny reconstruction. The aim is to infer a correct organismal phylogeny, using orthologous genetic loci, in which common ancestry of two sequences can be traced back to a speciation event. Phylogeny using homologous genetic loci derived by gene duplication (paralogy) or related through lateral gene transfer (xenology), cannot reflect evolutionary history of organisms.

Once DNA sequence data are generated, they are subjected to a multiple sequence alignment process. This involves finding homologous sites, that is, positions derived from the same ancestral organism in the molecules under study. A set of sequences can be aligned with another by introducing "alignment gaps" (known in brief as "gaps"). In general, multiple sequence alignment starts by aligning a pair of sequences (pairwise alignment) and is then expanded to multiple sequences using various algorithms.

Many algorithms and computer programs have been developed in the last few decades for multiple sequence alignment, but the original Clustal series programs are still most widely used and produce reasonably good quality MSA for small data sets. For a large dataset, such as massive pyrosequencing reads, the MUSCLE program can generate good compromise between accuracy and speed. The MAFFT program utilizes several different algorithmic approaches and can be used for either small or very large datasets. There are also other computer programs developed for general multiple sequence alignment, but the above three have been most popular and are routinely used in publications in various microbiological disciplines.

Multilocus sequence analysis

Multilocus sequence analysis (MLSA) represents the novel standard in microbial molecular systematics. In this context, MLSA is implemented in a relatively straightforward way, consisting essentially in the concatenation of several sequence partitions for the same set of organisms, resulting in a "supermatrix" which is used to infer a phylogeny by means of distance-matrix or optimality criterion-based methods. This approach is expected to have an increased resolving power due to the large number of characters analyzed and a lower sensitivity to the impact of conflicting signals (i.e. phylogenetic incongruence) that result from eventual horizontal gene transfer events. The strategies used to deal with multiple partitions can be grouped in three broad categories: the total evidence, separate analysis and combination approaches. The concatenation approach that dominates MLSAs in the microbial molecular systematics literature is known to systematists working with plants and animals as the "total molecular evidence" approach and has been used to solve difficult phylogenetic questions such as the relationships among the major groups of cetaceans, that of microsporidia and fungi, or the phylogeny of major plant lineages. The total molecular evidence approach has been criticized because by directly concatenating all available sequence alignments, the evidence of conflicting phylogenetic signals in the different data partitions is lost along with the possibility to uncover the evolutionary processes that gave rise to such contradictory signals. The nature of these conflicts is varied, but in the microbial world the strongest conflicting signals often derive from the existence of horizontal gene transfer events in the dataset. If the individuals containing xenologous loci are not identified and removed from the supermatrix prior to phylogeny inference, the resulting hypothesis may be strongly distorted, since standard treeing methods assume a single underlying evolutionary history. Based on these arguments, the conditional data combination strategy is to be generally preferred in bacterial MLSA.

rRNA and other global markers

The introduction of comparative rRNA sequence analysis represents a major milestone in the history of microbiology. The current taxonomy of prokaryotes as well as modern probe and chip based identification methods are mainly based upon rRNA derived phylogenetic conclusions. Also of importance is single gene based phylogenetic inference and alternative global markers include elongation and initiation factors, RNA polymerase subunits, DNA gyrases, heat shock and recA proteins. Although the comparative analyses are hampered by the generally low phylogenetic information content and different resolution power and multiple copies of the individual markers, the domain and prokaryotic phyla concept is globally supported.

The conserved inserts or deletions (indels) in protein sequences provide particularly useful means for identifying different groups of microbes in clear molecular terms and for understanding how they have branched off from a common ancestor. Conserved indels and lineage-specific proteins can be useful for understanding microbial phylogeny at different phylogenetic depths.

The phyla of prokaryotes

There is no official classification of prokaryotes. For the higher taxa there even is no official nomenclature: the rules of the International Code of Nomenclature of Prokaryotes do not cover taxa above the rank of class. The most commonly accepted division of the prokaryotes in two "subkingdoms" or "domains" (Bacteria and Archaea) and the classification of their species with validly published names in respectively 27 and 2 "phyla" or "divisions" (as of November 2009) is primarily based on 16S rRNA sequence comparisons. This type of classification was adopted in the latest edition of Bergey's Manual of Systematic Bacteriology. Alternative classifications have been proposed as well, based e.g. on the structure of the cell wall. Some 16S rRNA sequence-based phyla unite prokaryotes of similar physiological properties (for example Cyanobacteria, Chlorobi, Thermotogae); others (Euryarchaeota, Proteobacteria, Flavobacteria) contain organisms with highly disparate lifestyles. Some phyla based on deep 16S rRNA lineages are currently represented by one or a few species only. Environmental genomics/metagenomics approaches suggest existence of many more phyla based on the deep lineages of 16S rRNA gene sequences recovered. To obtain the organisms harboring these sequences and to study their properties is a major challenge of microbiology today.

Horizontal gene transfer

Efforts to construct the tree of life take their conceptual motivation from Charles Darwin's theory of evolution. Until the advent of molecular biology, however, a universal tree of life was well beyond the scope of the data and methods of traditional organismal phylogeny. The rapid development of these methods and bodies of genetic sequence from the 1970s onwards resulted in major reclassifications of life and revived ambitions to represent all organismal lineages by one true tree of life. Subsequent realization of the significance of lateral gene transfer and other non-vertical processes has subtly reconceptualized and reoriented attempts to construct this universal phylogeny.

Horizontal gene transfer has affected the formation of groups of organisms. Gene transfer can make it more difficult to define and determine relationships. In those cases where many genes have been transferred between preferred partners, the majority of genes in a genome may reflect gene acquisition and as a consequence, if a coherent signal is detected, one nevertheless might not be sure that the signal is due to organismal shared ancestry. However, the presence of a particular transferred gene has been shown, in several cases, to constitute a shared derived character useful in classification. Gene transfer can put together new metabolic pathways that open up new ecological niches and consequently, the transfer of an adaptive gene might create a new group of organisms.

Computational phylogenetics

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

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

Types of phylogenetic trees

Phylogenetic trees generated by computational phylogenetics can be either *rooted* or *unrooted* depending on the input data and the algorithm used. A rooted tree is a directed graph that explicitly identifies a most recent common ancestor (MRCA), usually an imputed sequence that is not represented in the input. Genetic distance measures can be used to plot a tree with the input sequences as leaf nodes and their distances from the root proportional to their genetic distance from the hypothesized MRCA. Identification of a root usually requires the inclusion in the input data of at least one "outgroup" known to be only distantly related to the sequences of interest.

By contrast, unrooted trees plot the distances and relationships between input sequences without making assumptions regarding their descent. An unrooted tree can always be produced from a rooted tree, but a root cannot usually be placed on an unrooted tree without additional data on divergence rates, such as the assumption of the molecular clock hypothesis.

The set of all possible phylogenetic trees for a given group of input sequences can be conceptualized as a discretely defined multidimensional "tree space" through which search paths can be traced by optimization algorithms. Although counting the total number of trees for a nontrivial number of input sequences can be complicated by variations in the definition of a tree topology, it is always true that there are more rooted than unrooted trees for a given number of inputs and choice of parameters.

Coding characters and defining homology

Morphological analysis

The basic problem in morphological phylogenetics is the assembly of a matrix representing a mapping from each of the taxa being compared to representative measurements for each of the phenotypic characteristics being used as a classifier. The types of phenotypic data used to construct this matrix depend on the taxa being compared; for individual species, they may involve measurements of average body size, lengths or sizes of particular bones or other physical features, or even behavioral manifestations. Of course, since not every possible phenotypic characteristic could be measured and encoded for analysis, the selection of which features to measure is a major inherent obstacle to the method. The decision of which traits to use as a basis for the matrix necessarily represents a hypothesis about which traits of a species or higher taxon are evolutionarily relevant. Morphological studies can be confounded by examples of convergent evolution of phenotypes. A major challenge in constructing useful classes is the high likelihood of inter-taxon overlap in the distribution of the phenotype's variation. The inclusion of extinct taxa in morphological analysis is often difficult due to absence of or incomplete fossil records, but has been shown to have a significant effect on the trees produced; in one study only the inclusion of extinct species of apes produced a morphologically derived tree that was consistent with that produced from molecular data.

Some phenotypic classifications, particularly those used when analyzing very diverse groups of taxa, are discrete and unambiguous; classifying organisms as possessing or lacking a tail, for example, is straightforward in the majority of cases, as is counting features such as eyes or vertebrae. However, the most appropriate representation of continuously varying phenotypic measurements is a controversial problem without a general solution. A common method is simply to sort the measurements of interest into two or more classes, rendering continuous observed variation as discretely classifiable (e.g., all examples with humerus bones longer than a given cutoff are scored as members of one state and all members whose humerus bones are shorter than the cutoff are scored as members of a second state). This results in an easily manipulated data set but has been criticized for poor reporting of the basis for the class definitions and for sacrificing information compared to methods that use a continuous weighted distribution of measurements.

Because morphological data is extremely labor-intensive to collect, whether from literature sources or from field observations, reuse of previously compiled data matrices is not uncommon, although this may propagate flaws in the original matrix into multiple derivative analyses.

Molecular analysis

The problem of character coding is very different in molecular analyses, as the characters in biological sequence data are immediate and discretely defined - distinct nucleotides in DNA or RNA sequences and distinct amino acids in protein sequences. However, defining homology can be challenging due to the inherent difficulties of multiple sequence alignment. For a given gapped MSA, several rooted phylogenetic trees can be constructed that vary in their interpretations of which changes are "mutations" versus ancestral characters and which events are insertion mutations or deletion mutations. For example, given only a pairwise alignment with a gap region, it is impossible to determine whether one sequence bears an insertion mutation or the other carries a deletion. The problem is magnified in MSAs with unaligned and nonoverlapping gaps. In practice, sizable regions of a calculated alignment may be discounted in phylogenetic tree construction to avoid integrating noisy data into the tree calculation.

Distance-matrix methods

Distance-matrix methods of phylogenetic analysis explicitly rely on a measure of "genetic distance" between the sequences being classified and therefore they require an MSA as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches. Distance methods attempt to construct an all-to-all matrix from the sequence query set describing the distance between each sequence pair. From this is constructed a phylogenetic tree that places closely related sequences under the same interior node and whose branch lengths closely reproduce the observed distances between sequences. Distance-matrix methods may produce either rooted or unrooted trees, depending on the algorithm used to calculate them. They are frequently used as the basis for progressive and iterative types of multiple sequence alignments. The main disadvantage of distance-matrix methods is their inability to efficiently use information about local high-variation regions that appear across multiple subtrees.

Neighbor-joining

Neighbor-joining methods apply general data clustering techniques to sequence analysis using genetic distance as a clustering metric. The simple neighbor-joining method produces unrooted trees, but it does not assume a constant rate of evolution (i.e., a molecular clock) across lineages. Its relative, UPGMA (Unweighted Pair Group Method with Arithmetic mean) produces rooted trees and requires a constant-rate assumption - that is, it assumes an ultrametric tree in which the distances from the root to every branch tip are equal.

Fitch-Margoliash method

The Fitch-Margoliash method uses a weighted least squares method for clustering based on genetic distance. Closely related sequences are given more weight in the tree construction process to correct for the increased inaccuracy in measuring distances between distantly related sequences. The distances used as input to the algorithm must be normalized to prevent large artifacts in computing relationships between closely related and distantly related groups. The distances calculated by this method must be linear; the linearity criterion for distances requires that the expected values of the branch lengths for two individual branches must equal the expected value of the sum of the two branch distances - a property that applies to biological sequences only when they have been corrected for the possibility of back mutations at individual sites. This correction is done through the use of a substitution matrix such as that derived from the Jukes-Cantor model of DNA evolution. The distance correction is only necessary in practice when the evolution rates differ among branches.

The least-squares criterion applied to these distances is more accurate but less efficient than the neighbor-joining methods. An additional improvement that corrects for correlations between distances that arise from many closely related sequences in the data set can also be applied at increased computational cost. Finding the optimal least-squares tree with any correction factor is NP-complete, so heuristic search methods like those used in maximum-parsimony analysis are applied to the search through tree space.

Using outgroups

Independent information about the relationship between sequences or groups can be used to help reduce the tree search space and root unrooted trees. Standard usage of distance-matrix methods involves the inclusion of at least one outgroup sequence known to be only distantly related to the sequences of interest in the query set. This usage can be seen as a type of experimental control. If the outgroup has been appropriately chosen, it will have a much greater genetic distance and thus a longer branch length than any other sequence and it will appear near the root of a rooted tree. Choosing an appropriate outgroup requires the selection of a sequence that is moderately related to the sequences of interest; too close a relationship defeats the purpose of the outgroup and too distant adds noise to the analysis. Care should also be taken to avoid situations in which the species from which the sequences were taken are distantly related, but the gene encoded by the sequences is highly conserved across lineages. Horizontal gene transfer, especially between otherwise divergent bacteria, can also confound outgroup usage.

Maximum parsimony

Maximum parsimony (MP) is a method of identifying the potential phylogenetic tree that requires the smallest total number of evolutionary events to explain the observed sequence data. Some ways of scoring trees also include a "cost" associated with particular types of evolutionary events and attempt to locate the tree with the smallest total cost. This is a useful approach in cases where not every possible type of event is equally likely

- for example, when particular nucleotides or amino acids are known to be more mutable than others.

The most naive way of identifying the most parsimonious tree is simple enumeration - considering each possible tree in succession and searching for the tree with the smallest score. However, this is only possible for a relatively small number of sequences or species because the problem of identifying the most parsimonious tree is known to be NP-hard; consequently a number of heuristic search methods for optimization have been developed to locate a highly parsimonious tree, if not the most optimal in the set. Most such methods involve a steepest descent-style minimization mechanism operating on a tree rearrangement criterion.

Branch and bound

The branch and bound algorithm is a general method used to increase the efficiency of searches for near-optimal solutions of NP-hard problems first applied to phylogenetics in the early 1980s. Branch and bound is particularly well suited to phylogenetic tree construction because it inherently requires dividing a problem into a tree structure as it subdivides the problem space into smaller regions. As its name implies, it requires as input both a branching rule (in the case of phylogenetics, the addition of the next species or sequence to the tree) and a bound (a rule that excludes certain regions of the search space from consideration, thereby assuming that the optimal solution cannot occupy that region). Identifying a good bound is the most challenging aspect of the algorithm's application to phylogenetics. A simple way of defining the bound is a maximum number of assumed evolutionary changes allowed per tree. A set of criteria known as Zharkikh's rules severely limit the search space by defining characteristics shared by all candidate "most parsimonious" trees. The two most basic rules require the elimination of all but one redundant sequence (for cases where multiple observations have produced identical data) and the elimination of character sites at which two or more states do not occur in at least two species. Under ideal conditions these rules and their associated algorithm would completely define a tree.

Sankoff-Morel-Cedergren algorithm

The Sankoff-Morel-Cedergren algorithm was among the first published methods to simultaneously produce an MSA and a phylogenetic tree for nucleotide sequences. The method uses a maximum parsimony calculation in conjunction with a scoring function that penalizes gaps and mismatches, thereby favoring the tree that introduces a minimal number of such events. The imputed sequences at the interior nodes of the tree are scored and summed over all the nodes in each possible tree. The lowest-scoring tree sum provides both an optimal tree and an optimal MSA given the scoring function. Because the method is highly computationally intensive, an approximate method in which initial guesses for the interior alignments are refined one node at a time. Both the full and the approximate version are in practice calculated by dynamic programming.

MALIGN and POY

More recent phylogenetic tree/MSA methods use heuristics to isolate high-scoring, but not necessarily optimal, trees. The MALIGN method uses a maximum-parsimony technique to compute a multiple alignment by maximizing a cladogram score and its companion POY uses an iterative method that couples the optimization of the phylogenetic tree with improvements in the corresponding MSA. However, the use of these methods in constructing evolutionary hypotheses has been criticized as biased due to the deliberate construction of trees reflecting minimal evolutionary events.

Maximum likelihood

The maximum likelihood method uses standard statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees. The method requires a substitution model to assess the probability of particular mutations; roughly, a tree that requires more mutations at interior nodes to explain the observed phylogeny will be assessed as having a lower probability. This is broadly similar to the maximum-parsimony method, but maximum likelihood allows additional statistical flexibility by permitting varying rates of evolution across both lineages and sites. In fact, the method requires that evolution at different sites and along different lineages must be statistically independent. Maximum likelihood is thus well suited to the analysis of distantly related sequences, but because it formally requires search of all possible combinations of tree topology and branch length, it is computationally expensive to perform on more than a few sequences.

The "pruning" algorithm, a variant of dynamic programming, is often used to reduce the search space by efficiently calculating the likelihood of subtrees. The method calculates the likelihood for each site in a "linear" manner, starting at a node whose only descendants are leaves (that is, the tips of the tree) and working backwards toward the "bottom" node in nested sets. However, the trees produced by the method are only rooted if the substitution model is irreversible, which is not generally true of biological systems. The search for the maximum-likelihood tree also includes a branch length optimization component that is difficult to improve upon algorithmically; general global optimization tools such as the Newton-Raphson method are often used. Searching tree topologies defined by likelihood has not been shown to be NP-complete, but remains extremely challenging because branch-and-bound search is not yet effective for trees represented in this way.

Bayesian inference

Bayesian inference can be used to produce phylogenetic trees in a manner closely related to the maximum likelihood methods. Bayesian methods assume a prior probability distribution of the possible trees, which may simply be the probability of any one tree among all the possible trees that could be generated from the data, or may be a more sophisticated estimate derived from the assumption that divergence events such as speciation occur as stochastic processes. The choice of prior distribution is a point of contention among users of Bayesian-inference phylogenetics methods.

Implementations of Bayesian methods generally use Markov chain Monte Carlo sampling algorithms, although the choice of move set varies; selections used in Bayesian phylogenetics include circularly permuting leaf nodes of a proposed tree at each step and swapping descendant subtrees of a random internal node between two related trees. The use of Bayesian methods in phylogenetics has been controversial, largely due to incomplete specification of the choice of move set, acceptance criterion and prior distribution in published work.

Model selection

Molecular phylogenetics methods rely on a defined substitution model that encodes a hypothesis about the relative rates of mutation at various sites along the gene or amino acid sequences being studied. At their simplest, substitution models aim to correct for differences in the rates of transitions and transversions in nucleotide sequences. The use of substitution models is necessitated by the fact that the genetic distance between two sequences increases linearly only for a short time after the two sequences diverge from each other (alternatively, the distance is linear only shortly before coalescence). The longer the amount of time after divergence, the more likely it becomes that two mutations occur at the same nucleotide site. Simple genetic distance calculations will thus undercount the number of mutation events that have occurred in evolutionary history. The extent of this undercount increases with increasing time since divergence, which can lead to the phenomenon of long branch attraction, or the misassignment of two distantly related but convergently evolving sequences as closely related. The maximum parsimony method is particularly susceptible to this problem due to its explicit search for a tree representing a minimum number of distinct evolutionary events.

Types of models

All substitution models assign a set of weights to each possible change of state represented in the sequence. The most common model types are implicitly reversible because they assign the same weight to, for example, a G>C nucleotide mutation as to a C>G mutation. The simplest possible model, the Jukes-Cantor model, assigns an equal probability to every possible change of state for a given nucleotide base. The rate of change between any two distinct nucleotides will be one-third of the overall substitution rate. More advanced models distinguish between transitions and transversions. The most general possible time-reversible model, called the GTR model, has six mutation rate parameters. An even more generalized model known as the general 12-parameter model breaks time-reversibility, at the cost of much additional complexity in calculating genetic distances that are consistent among multiple lineages. One possible variation on this theme adjusts the rates so that overall GC content - an important measure of DNA double helix stability - varies over time.

Models may also allow for the variation of rates with positions in the input sequence. The most obvious example of such variation follows from the arrangement of nucleotides in protein-coding genes into three-base codons. If the location of the open reading frame (ORF) is known, rates of mutation can be adjusted for position of a given site within a codon, since it is known that wobble base pairing can allow for higher mutation rates in the third nucleotide of a given codon without affecting the codon's meaning in the genetic code. A less hypothesis-driven example that does not rely on ORF identification simply assigns to each site a rate randomly drawn from a predetermined distribution, often the gamma distribution or log-normal distribution. Finally, a more conservative estimate of rate variations known as the covarion method allows autocorrelated variations in rates, so that the mutation rate of a given site is correlated across sites and lineages.

Choosing the best model

The selection of an appropriate model is critical for the production of good phylogenetic analyses, both because underparameterized or overly restrictive models may produce aberrant behavior when their underlying assumptions are violated and because overly complex or overparameterized models are computationally expensive and the parameters may be overfit. The most common method of model selection is the likelihood ratio test (LRT), which produces a likelihood estimate that can be interpreted as a measure of "goodness of fit" between the model and the input data. However, care must be taken in using these results, since a more complex model with more parameters will always have a higher likelihood than a simplified version of the same model, which can lead to the naive selection of models that are overly complex. For this reason model selection computer programs will choose the simplest model that is not significantly worse than more complex substitution models. A significant disadvantage of the LRT is the necessity of making a series of pairwise comparisons between models; it has been shown that the order in which the models are compared has a major effect on the one that is eventually selected.

An alternative model selection method is the Akaike information criterion (AIC), formally an estimate of the Kullback-Leibler divergence between the true model and the model being tested. It can be interpreted as a likelihood estimate with a correction factor to penalize overparameterized models. The AIC is calculated on an individual model rather than a pair, so it is independent of the order in which models are assessed. A related alternative, the Bayesian information criterion (BIC), has a similar basic interpretation but penalizes complex models more heavily.

Chapter- 4

Cladistics

Cladistics (Ancient Greek: *κλάδος*, *klados*, "branch") is a method of classifying species of organisms into groups called **clades**, which consist only of firstly, all the descendants of an ancestral organism and secondly, the ancestor itself. For example, birds, dinosaurs, crocodiles and all descendants (living or extinct) of their most recent common ancestor form a clade. In the terms of biological systematics, a clade is a single "branch" on the "tree of life", a monophyletic group.

Cladistics can be distinguished from other taxonomic systems, such as phenetics, by its focus on shared derived characters (synapomorphies). Systems developed earlier usually employed overall morphological similarity to group species into genera, families and other higher level groups (taxa); cladistic classifications (usually in the form of trees called cladograms) are intended to reflect the relative recency of common ancestry or the sharing of homologous features. Cladistics is also distinguished by an emphasis on parsimony and hypothesis testing (particularly falsificationism), leading to a claim that cladistics is more objective than systems which rely on subjective judgements of relationship based on similarity.

Cladistics originated in the work of the German entomologist Willi Hennig, who referred to it as "phylogenetic systematics" (also the name of his 1966 book); the use of the terms "cladistics" and "clade" was popularized by other researchers. The technique and sometimes the name have been successfully applied in other disciplines: for example, to determine the relationships between the surviving manuscripts of the *Canterbury Tales*.

Cladists use *cladograms*, diagrams which show ancestral relations between species, to represent the monophyletic relationships of species, termed sister-group relationships. This is interpreted as representing phylogeny, or evolutionary relationships. Although traditionally such cladograms were generated largely on the basis of morphological characters, genetic sequencing data and computational phylogenetics are now very commonly used in the generation of cladograms.

Cladistics, either generally or in specific applications, has been criticized from its beginnings. A decision as to whether a particular character is a synapomorphy or not may be challenged as involving subjective judgements, raising the issue of whether cladistics as actually practised is as objective as has been claimed. Formal classifications based on cladistic reasoning are said to emphasize ancestry at the expense of descriptive characteristics and thus ignore biologically sensible, clearly defined groups which do not fall into clades (e.g. reptiles as traditionally defined or prokaryotes).

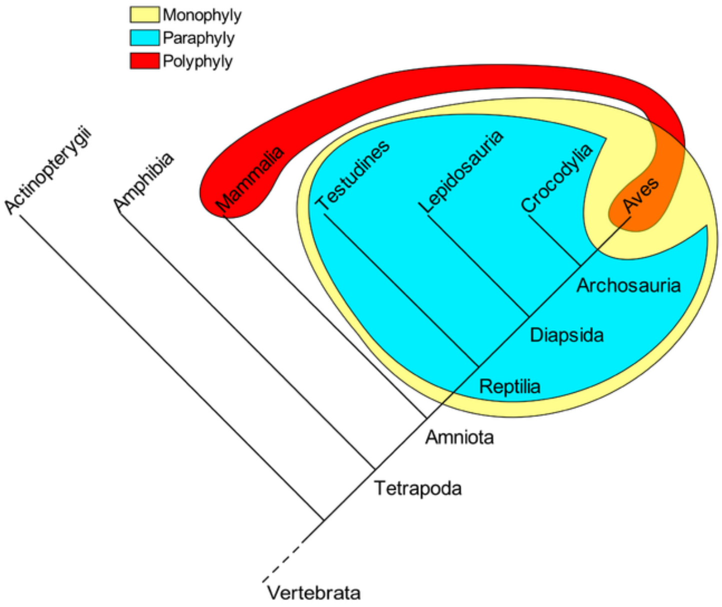
History of cladistics

The term *clade* was introduced in 1958 by Julian Huxley, *cladistic* by Cain and Harrison in 1960 and *cladist* (for an adherent of Hennig's school) by Mayr in 1965. Hennig referred to his own approach as *phylogenetic systematics*. From the time of his original formulation until the end of the 1980s cladistics remained a minority approach to classification. However in the 1990s it rapidly became the dominant method of classification in evolutionary biology. Computers made it possible to process large quantities of data about organisms and their characteristics. At about the same time the development of effective polymerase chain reaction techniques made it possible to apply cladistic methods of analysis to biochemical and molecular genetic features of organisms as well as to anatomical ones.

Cladistics as a successor to phenetics

For some decades in the mid to late twentieth century, a commonly used methodology was phenetics ("numerical taxonomy"). This can be seen as a predecessor to some methods of today's cladistics (namely distance matrix methods such as neighbor-joining), but made no attempt to resolve phylogeny, only similarities.

Clades



Partial Evolutionary Tree of the Vertebrates

The yellow group (sauropsids) is monophyletic, the blue group (traditional reptiles) is paraphyletic and the red group (warm-blooded animals) is polyphyletic.

A clade is a group of taxa consisting only of an ancestor taxon and all of its descendant taxa. In the diagram provided (a **cladogram**), it is hypothesized that all vertebrates, including ray-finned fishes (Actinopterygii), had a common ancestor and so form a clade. Within the vertebrates, all tetrapods, including amphibians, mammals, reptiles (as traditionally defined) and birds, are hypothesized to have had a common ancestor and so also form a clade. The tetrapod ancestor was a descendant of the original vertebrate ancestor, but is not an ancestor of any ray-finned fish living today.

An important caution is that any cladogram is a provisional hypothesis. Although unlikely, future genetic or morphological evidence might suggest that ray-finned fish and amphibians share a common ancestor that was not an ancestor of the other tetrapods. The new information would cause us to define a ray-finned-fish-and-amphibian clade, altering the cladogram.

The relationship between clades can be described in several ways:

- A clade is *basal* to another clade if it contains that other clade as a subset within it. In the example, the vertebrate clade is basal to the tetrapod and ray-finned fish clades. (Some authors have used "basal" differently to mean a clade that is less species-rich than a sister clade, with such a deficit being taken as an indication of 'primitiveness'. Others consider this usage to be incorrect.)
- A clade located within a clade is said to be *nested* within that clade. In the diagram, the tetrapod clade is nested within the vertebrate clade.
- Two clades are *sisters* if they have an immediate common ancestor. In the diagram, crocodiles and birds are sister clades, as are amphibians and amniotes.

Terminology for characters

The following terms are used to identify shared or distinct characters among groups:

- *Plesiomorphy* ("close form") or *ancestral state*, also *symplesiomorphy* ("shared plesiomorphy", i.e. "shared close form"), is a characteristic that is present at the base of a tree (cladogram). Since a plesiomorphy that is inherited from the common ancestor may appear anywhere in a tree, its presence provides no evidence of relationships within the tree. The traditional definition of reptiles (the blue group in the diagram) includes being cold-blooded (i.e. not maintaining a constant high body temperature), whereas birds are warm-blooded. Since cold-bloodedness is a plesiomorphy, inherited from the common ancestor of traditional reptiles and birds, it should not be used to define a group in a system based on cladistics.
- *Apomorphy* ("separate form") or *derived state* is a characteristic believed to have evolved within the tree. It can thus be used to separate one group in the tree from the rest. Within the group which shares the apomorphy it is a *synapomorphy* ("shared apomorphy", i.e. "shared separate form"). For example, within the vertebrates, all tetrapods (and only tetrapods) have four limbs; thus, having four

limbs is an synapomorphy for tetrapods. All the tetrapods can legitimately be grouped together because they have four limbs.
- *Homoplasy* is a characteristic shared by members of a tree but not present in their common ancestor. It arises by convergence or reversion. Both mammals and birds are able to maintain a high constant body temperature (i.e. they are 'warm-blooded'). However, the ancestors of each group did not share this character, so it must have evolved independently. Mammals and birds should not be grouped together on the basis that they are warm-blooded.

The terms (sym)plesiomorphy and (syn)apomorphy are relative and their application depends on the position of a group within a tree. An apomorphy of one clade is a plesiomorphy of another contained within it. For example, when trying to decide whether tetrapods should form a clade, an important question is whether having four limbs is a synapomorphy of all the taxa to be included within Tetrapoda: did all the possible members of the Tetrapoda inherit four limbs from a common ancestor, whereas all other vertebrates did not? By contrast, for a group within the tetrapods, such as birds, having four limbs is a plesiomorphy. The fact that ostriches and rheas both have four limbs does not provide any support for putting them into a separate group of 'flightless birds'. Using these two terms allows a greater precision in the discussion of homology, in particular allowing clear expression of the hierarchical relationships among different homologies.

It can be difficult to decide whether a character is in fact the same and thus can classified as a synapomorphy which may identify a group, or whether it only appears to be the same and is thus a homoplasy which cannot identify a group. There is a danger of circular reasoning: assumptions about the shape of a phylogenetic tree are used to justify decisions about characters, which are then used as evidence for the shape of the tree. It has been argued that this kind of reasoning has been used by proponents of the view that birds are nested within the theropod dinosaur clade.

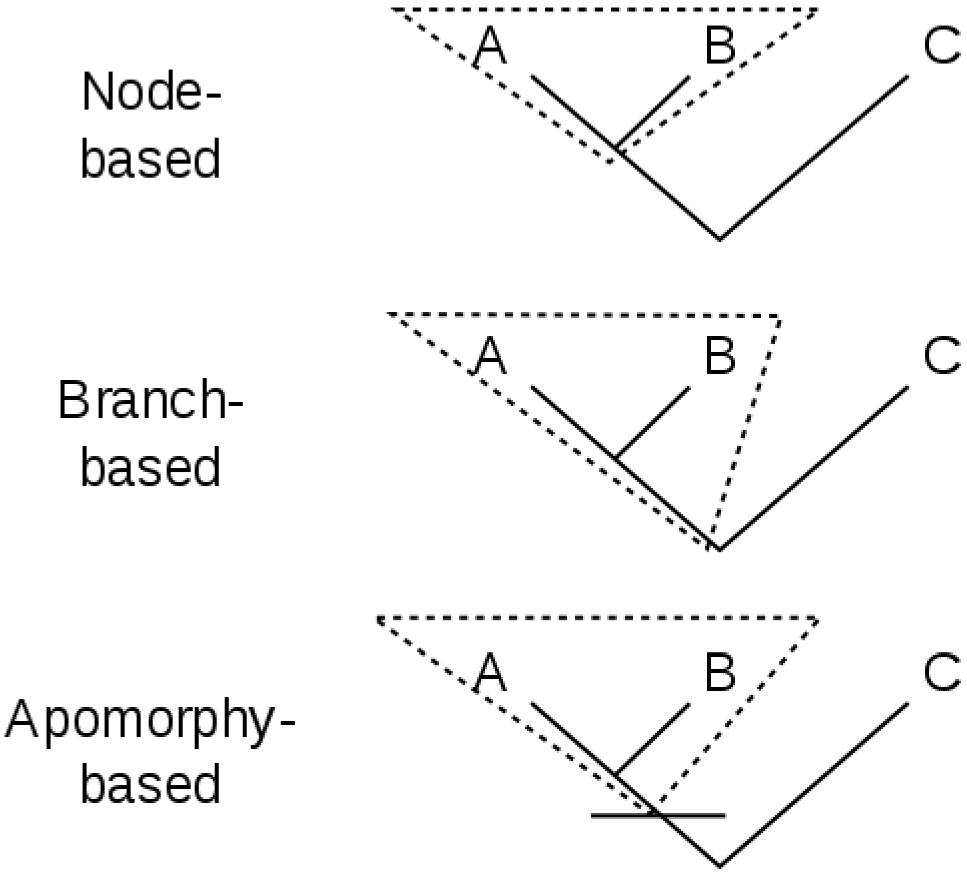
Terminology for groups

Three main types of group can be identified on the basis of their relationships in cladograms. The three can be defined in two different but related ways, as shown in the table below. The first is in terms of the shape of a set of nodes taken from a cladogram. In this approach, an 'ancestor node' is simply a branching point in the diagram; it may or may not correspond to an actual ancestor. The second is in terms of the characters of the taxa being classified and how these characters have been inherited. In this approach, an ancestor is an actual taxon, whether currently known or not.

Term	Node-based definition	Character-based definition
Monophyly	A monophyletic group of nodes in a tree is one which includes all the nodes descended from their most recent common ancestor node, plus the most recent	A monophyletic group of taxa is characterized by one or more **synapomorphies**: derived characters inherited by all members of the group from ancestors and not inherited by any

	common ancestor node, but no other nodes.	other taxa. A monophyletic group is a 'clade'. A 'crown group' is an example of a monophyletic group.
Paraphyly	A paraphyletic group of nodes in a tree is one which is constructed by taking a monophyletic group and removing one or more smaller monophyletic groups. (Removing one group produces a singly paraphyletic group, removing two a doubly paraphylectic group and so on.) A paraphyletic group is necessarily non-monophyletic.	A paraphyletic group of taxa is characterized by one or more **(sym)plesiomorphies**: characters inherited from ancestors but not present in all of their descendants. As a consequence, a paraphyletic group is truncated, in that it excludes one or more monophyletic taxa from an initially monophyletic group. An alternative name is an 'evolutionary grade', refering to the ancestral character state within the group. A 'stem group' is an example of a paraphyletic group.
Polyphyly	A polyphyletic group of nodes in a tree is one which is neither monophyletic nor paraphyletic.	A polyphyletic group of taxa is characterized by by one or more **homoplasies**: characters which have converged or reverted so as to appear to be the same but which have not been inherited from common ancestors. As a consequence, polyphyletic groups of taxa are totally artificial.

Branch-based definitions of clade



Three alternative ways to define a clade

The node-based definition of a monophyletic group (i.e. a clade) given above regards the lines in the cladogram only as a way of showing connections between taxa. This is appropriate when considering only living (extant) taxa; however, when extinct taxa are to be included in a cladogram, lines correspond to sequences of ancestors. There are two alternative ways of defining a clade which explicitly take into account the line below the branching point at the base of a clade. These definitions are most notably set out in the PhyloCode.

Consider how a clade combining A and B in the diagram can be defined.

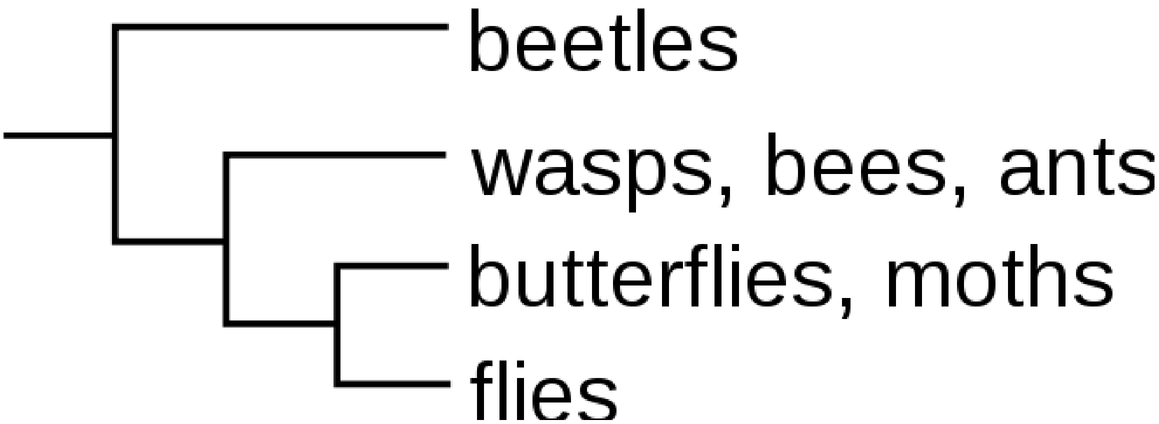
- *Node-based*: The node-based definition specifies A+B as the *last* common ancestor of A and B and all descendants of that ancestor. It thus excludes from the clade the line below the junction of A and B. Crown groups are a type of node-based clade.
- *Branch-based*: A branch-based definition specifies A+B as the *first* ancestor of A which is not also an ancestor of C and all descendants of that ancestor. It thus includes in the clade the line below the junction of A and B. (This type of definition was originally called "stem-based", but this was changed to avoid

confusion with the term "stem group", which is parapyletic.) Total groups are a type of branch-based clade.

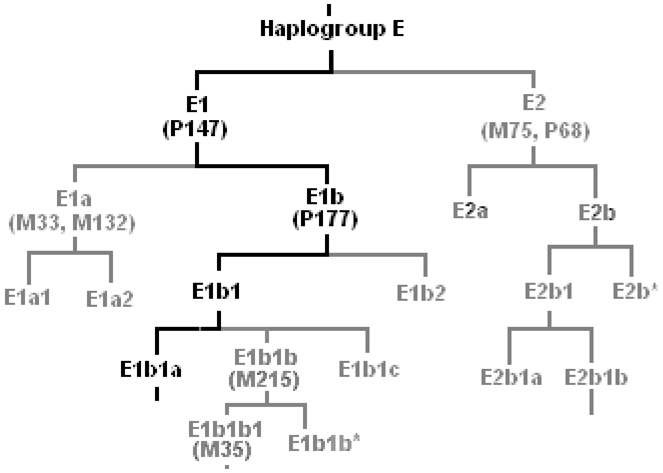
- *Apomorphy-based*: An apomorphy-based definition specifies A+B as the first ancestor of A to possess derived trait M homologously (that is, synapomorphically) with that trait in A and all descendants of that ancestor. It thus includes in the clade only that part of the line below the junction of A and B which corresponds to ancestors possessing the apomorphy. The process of identifying and naming groups based on apomorphies is the method that most resembles classical systematics, with the proviso that cladistic taxa always denote a clade.

Note that these alternative definitions do not alter the classification of the tips of the tree and so are equivalent if only living (extant) taxa are being considered.

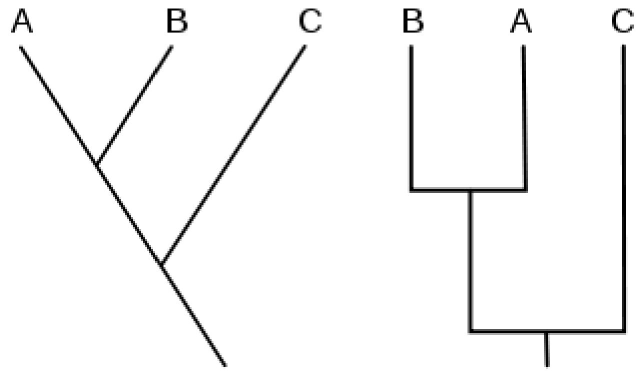
Cladograms



A horizontal cladogram, with the ancestor (not named) to the left



A vertical cladogram, with the ancestor at the top



Two vertical cladograms, the ancestor at the bottom

A cladogram is a diagram used in cladistics which shows ancestral relations between organisms, to represent the evolutionary tree of life. Although traditionally such cladograms were generated largely on the basis of morphological characters, DNA and RNA sequencing data and computational phylogenetics are now very commonly used in the generation of cladograms.

Generating a cladogram

A greatly simplified procedure for generating a cladogram is:

1. Gather and organize data
2. Consider possible cladograms
3. Select best cladogram

Step 1: gather and organize data

A cladistic analysis begins with the following data:

- a list of taxa (for example, species) to be organized
- a list of characteristics to be compared
- for each taxon, the value of each of the listed characteristics or *character states*

For example, if analyzing 20 species of birds, the data might be:

- the list of the 20 species
- characteristics such as genome sequence, skeletal anatomy, biochemical processes and feather coloration
- for each of the 20 species, its particular genome sequence, skeletal anatomy, biochemical processes and feather coloration

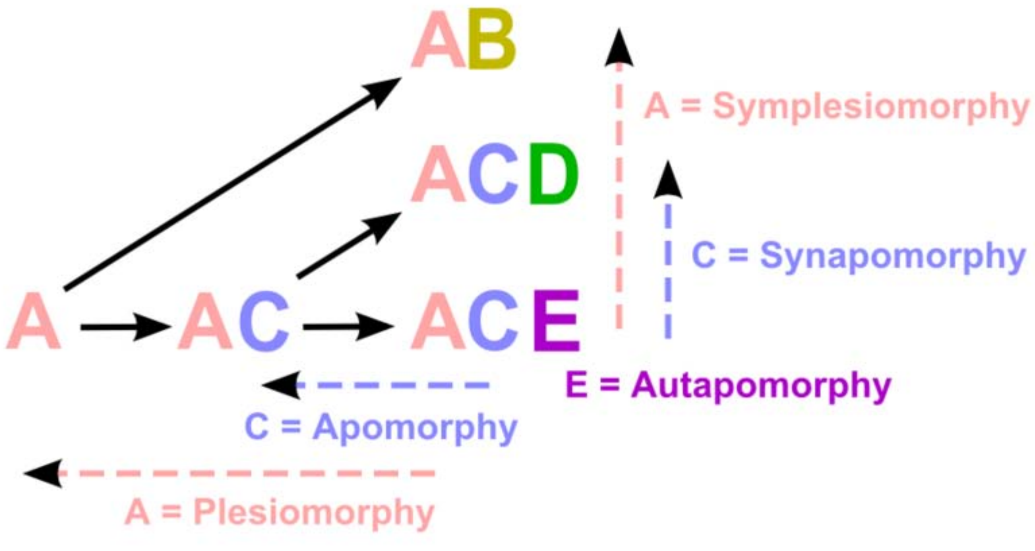
All the data are then organized into a "taxon-character matrix", which is the base to perform phylogenetic analysis.

Molecular versus morphological data

The characteristics used to create a cladogram can be roughly categorized as either morphological (synapsid skull, warm blooded, notochord, unicellular, etc.) or molecular (DNA, RNA, or other genetic information). Prior to the advent of DNA sequencing, all cladistic analysis used morphological data.

As DNA sequencing has become cheaper and easier, molecular systematics has become a more and more popular way to reconstruct phylogenies. Using a parsimony criterion is only one of several methods to infer a phylogeny from molecular data; maximum likelihood and Bayesian inference, which incorporate explicit models of sequence evolution, are non-Hennigian ways to evaluate sequence data. Another powerful method of reconstructing phylogenies is the use of genomic retrotransposon markers, which are thought to be less prone to the problem of reversion that plagues sequence data. They are also generally assumed to have a low incidence of homoplasies because it was once thought that their integration into the genome was entirely random; this seems at least sometimes not to be the case, however.

Ideally, morphological, molecular and possibly other phylogenies should be combined into an analysis of *total evidence*: All have different intrinsic sources of error. For example, character convergence (homoplasy) is much more common in morphological data than in molecular sequence data, but character reversions that are unrecognizable as such are more common in the latter. Morphological homoplasies can usually be recognized as such if character states are defined with enough attention to detail.



Apomorphy in cladistics

Plesiomorphies and synapomorphies

The researcher must decide which character states were present *before* the last common ancestor of the species group (*plesiomorphies*) and which were present *in* the last common ancestor (*synapomorphies*) and does so by comparison to one or more *outgroups*. The choice of an outgroup is a crucial step in cladistic analysis because different outgroups can produce trees with profoundly different topologies. Note that only synapomorphies are of use in characterizing clades.

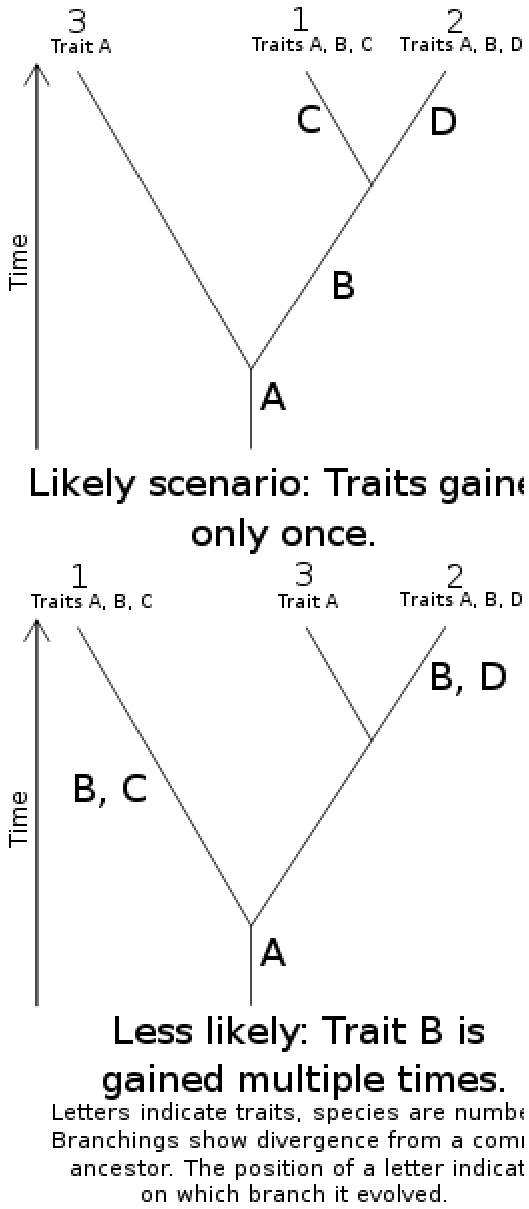
Avoid homoplasies

A homoplasy is a character that is shared by multiple species due to some cause *other* than common ancestry. The two main types of homoplasy are convergence (appearance of the same character in at least two distinct lineages) and reversion (the return to an ancestral character). Use of homoplasies when building a cladogram is sometimes unavoidable but is to be avoided when possible.

A well known example of homoplasy due to convergent evolution would be the character, "presence of wings". Though the wings of birds, bats and insects serve the same function, each evolved independently, as can be seen by their anatomy. If a bird, bat and a winged insect were scored for the character, "presence of wings", a homoplasy would be introduced into the dataset and this would confound the analysis, possibly resulting in a false evolutionary scenario.

Homoplasies can often be avoided outright in morphological datasets by defining characters more precisely and increasing their number. When analyzing "supertrees" (datasets incorporating as many taxa of a suspected clade as possible), it may become unavoidable to introduce character definitions that are imprecise, as otherwise the characters might not apply at all to a large number of taxa; to continue with the "wings" example, the presence of wings would hardly be a useful character if attempting a phylogeny of all Metazoa, as most of these don't have wings at all. Cautious choice and definition of characters thus is another important element in cladistic analyses. With a faulty outgroup or character set, no method of evaluation is likely to produce a phylogeny representing the evolutionary reality.

Step 2: consider possible cladograms


3
Trait A
1
Traits A, B, C
2
Traits A, B, D
C
D
B
A
Time
Likely scenario: Traits gaine
only once.
1
Traits A, B, C
3
Trait A
2
Traits A, B, D
B, D
B, C
Time
A
Less likely: Trait B is
gained multiple times.
Letters indicate traits, species are numbe
Branchings show divergence from a comi
ancestor. The position of a letter indicat
on which branch it evolved.

Simple cladistics

When there are just a few species being organized, it is possible to do this step manually, but most cases require a computer program. There are scores of computer programs available to support cladistics.

Because the total number of possible cladograms grows exponentially with the number of species, it is impractical for a computer program to evaluate every individual cladogram. A typical cladistic program begins by using heuristic techniques to identify a small number of candidate cladograms. Many cladistic programs then continue the search with the following repetitive steps:

1. Evaluate the candidate cladograms by comparing them to the characteristic data
2. Identify the best candidates that are most consistent with the characteristic data
3. Create additional candidates by creating several variants of each of the best candidates from the prior step
4. Use heuristics to create several new candidate cladograms unrelated to the prior candidates
5. Repeat these steps until the cladograms stop getting better

Computer programs that generate cladograms use algorithms that are very computationally intensive, because the cladogram problem is NP-hard.

Step 3: select best cladogram

There are several algorithms available to identify the "best" cladogram. Most algorithms use a metric to measure how consistent a candidate cladogram is with the data. Most cladogram algorithms use the mathematical techniques of optimization and minimization.

In general, cladogram generation algorithms must be implemented as computer programs, although some algorithms can be performed manually when the data sets are trivial (for example, just a few species and a couple of characteristics).

Some algorithms are useful only when the characteristic data are molecular (DNA, RNA); other algorithms are useful only when the characteristic data are morphological. Other algorithms can be used when the characteristic data includes both molecular and morphological data.

Algorithms for cladograms include least squares, neighbor-joining, parsimony, maximum likelihood and Bayesian inference.

Biologists sometimes use the term parsimony for a specific kind of cladogram generation algorithm and sometimes as an umbrella term for all cladogram algorithms.

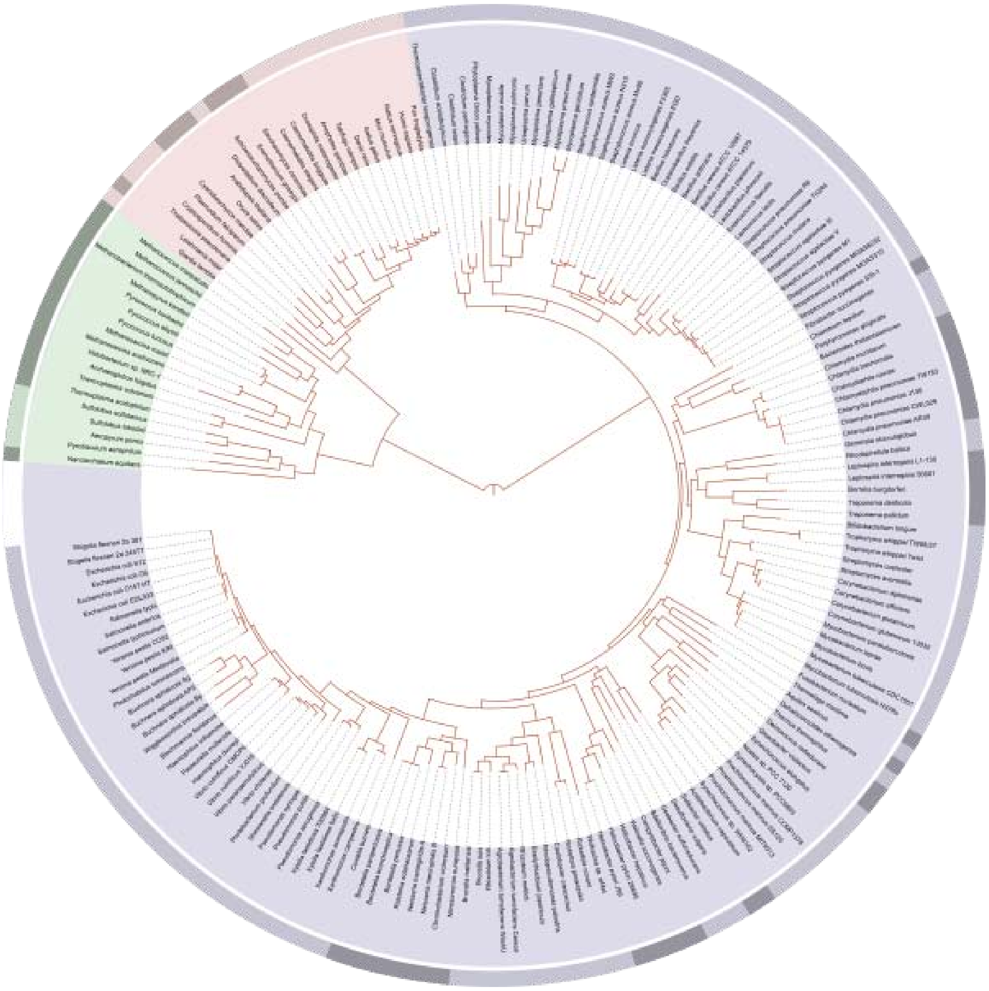
Algorithms that perform optimization tasks (such as building cladograms) can be sensitive to the order in which the input data (the list of species and their characteristics) is presented. Inputting the data in various orders can cause the same algorithm to produce different "best" cladograms. In these situations, the user should input the data in various orders and compare the results.

Using different algorithms on a single data set can sometimes yield different "best" cladograms, because each algorithm may have a unique definition of what is "best".

Because of the astronomical number of possible cladograms, algorithms cannot guarantee that the solution is the overall best solution. A nonoptimal cladogram will be selected if the program settles on a local minimum rather than the desired global minimum. To help solve this problem, many cladogram algorithms use a simulated annealing approach to increase the likelihood that the selected cladogram is the optimal one.

Cladistics in taxonomy

Phylogenetic nomenclature contrasted with traditional taxonomy



A highly resolved, automatically generated tree of life based on completely sequenced genomes

Most taxonomists have used the traditional approaches of Linnaean taxonomy and later Evolutionary taxonomy to organize life forms. These approaches use several fixed levels of a hierarchy, such as kingdom, phylum, class, order and family. Phylogenetic nomenclature does not feature those terms, because the evolutionary tree is so deep and so complex that it is inadvisable to set a fixed number of levels.

Evolutionary taxonomy insists that groups reflect phylogenies. In contrast, Linnaean taxonomy allows both monophyletic and paraphyletic groups as taxa. Since the early 20th century, Linnaean taxonomists have generally attempted to make at least family- and lower-level taxa (i.e. those regulated by the codes of nomenclature) monophyletic. Ernst Mayr in 1985 drew a distinction between the terms cladistics and phylogeny: "It would seem to me to be quite evident that the two concepts of phylogeny (and their role in the construction of classifications) are sufficiently different to require terminological distinction. The term *phylogeny* should be retained for the broad concept of phylogeny, promoted by Darwin and adopted by most students of phylogeny in the ensuing 90 years. The concept of phylogeny as mere genealogy should be terminologically distinguished as *cladistics*. To lump the two concepts together terminologically could not help but produce harmful equivocation."

Willi Hennig's pioneering work provoked a spirited debate about the relative merits of phylogenetic nomenclature versus Linnaean or evolutionary taxonomy, which has continued down to the present; however Hennig did not advocate abandoning the Linnaean nomenclatural system. Some of the debates in which the cladists were engaged had been running since the 19th century, but they were renewed fervor, as can be seen from the *Foreword* to Hennig (1979) by Rosen, Nelson and Patterson:

"Encumbered with vague and slippery ideas about adaptation, fitness, biological species and natural selection, neo-Darwinism (summed up in the "evolutionary" systematics of Mayr and Simpson) not only lacked a definable investigatory method, but came to depend, both for evolutionary interpretation and classification, on consensus or authority."

Phylogenetic nomenclature strictly and exclusively follows phylogeny and has arbitrarily deep trees with binary branching: each taxon corresponds to a clade. Linnaean taxonomy, while since the advent of evolutionary theory following phylogeny, also may subjectively consider similarity and has a fixed hierarchy of taxonomic ranks and its taxa are not required to correspond to clades.

Paraphyletic groups discouraged

Many cladists discourage the use of paraphyletic groups in classification of organisms, because they detract from cladistics' emphasis on clades (monophyletic groups). In contrast, proponents of the use of paraphyletic groups argue that any dividing line in a cladogram creates both a monophyletic section above and a paraphyletic section below. They also contend that paraphyletic taxa are necessary for classifying earlier sections of the tree – for instance, the early vertebrates that would someday evolve into the family Hominidae cannot be placed in any other monophyletic family. They also argue that paraphyletic taxa provide information about significant changes in organisms' morphology, ecology, or life history – in short, that both paraphyletic groups and clades are valuable notions with separate purposes.

Complexity of the Tree of Life

The cladistic tree of life is a fractal:

"The tree of life is inherently fractal-like in its complexity, ... Look closely at the 'lineage' of a phylogeny ... and it dissolves into many smaller lineages and so on, down to a very fine scale."

The overall shape of a dichotomous (bifurcating) tree is recursive; as a viewpoint zooms into the tree of life, the same type of tree appears no matter what the scale. When extinct species are considered (both known and unknown), the complexity and depth of the tree can be very large. Moreover the tree continues to recreate itself by bifurcation, a series of events called fractal evolution. Every single speciation event, including all the species that are now extinct, represents an additional fork on the hypothetical, complete cladogram of the tree of life.

The tree of life is a quasi-self-similar fractal; that is, the deep reconstruction is not as regular as the shallow reconstruction. By shallow Mishler means the most recent branching toward and at the tips and by deep the more ancient branches further back, which are harder to reconstruct and are missing unknown extinct lines. In the shallow part of the tree, branching events are relatively regular; it is often possible to estimate the times between them. In the deep part of the tree, "homology assessments" are "difficult" and the times vary widely. At this level Eldredge's and Gould's punctuated equilibrium applies, which hypothesizes long periods of stability followed by punctuations of rapid speciation, based on the fossil record.

PhyloCode approach to naming species

A formal code of phylogenetic nomenclature, the PhyloCode, is currently under development. It is intended for use by both those who would like to abandon Linnaean taxonomy and those who would like to use taxa and clades side by side. In several instances it has been employed to clarify uncertainties in Linnaean systematics so that in combination they yield a taxonomy that unambiguously places problematic groups in the evolutionary tree in a way that is consistent with current knowledge.

Example

For example, Linnaean taxonomy contains the taxon Tetrapoda, defined morphologically as vertebrates with four limbs (as well as animals with four-limbed ancestors, such as snakes), which is often given the rank of superclass and divides into the classes Amphibia, Reptilia, Aves, Mammalia.

Phylogenetic nomenclature also contains the taxon Tetrapoda, whose living members can be classified phylogenically as "the clade defined by the common ancestor of amphibians and mammals", or more precisely the clade defined by the common ancestor of a specific amphibian and mammal (or bird or snake). This definition gives us the Crown group tetrapods (or Crown-Tetrapoda). A few primitive four legged ancestors (the Ichthyostegalia) fall outside Crown-Tetrapoda. An alternative is to define tetrapoda as all animals more closely related to mammals than to lungfish (our nearest living non-tetrapod relatives). In this definition, the ichthyostegalians are included, together with a host of fossil animals usually classed as crossopterygian fish. This wider definition is termed Pan-Tetrapoda. A third option is to define Tetrapoda according to their apomorphy (their unique trait, i.e. having legs rather than fins), a definition that yield the same group as the Linnaean taxon.

Non of the phylogenetic taxa as described above have a rank and neither do its subtaxa. All the subclades are contained within one another. The clades are not divided into several non-overlapping taxa (as in traditional taxonomy), rather the clade is split into two clades at the first branching, a process repeated throughout. With regards to the traditional classes, Aves and Mammalia are subclades, contained in the subclade Amniota, while Reptilia and Amphibia are paraphyletic taxa, not clades. Instead of classifying non-mammalian, non-avian amniotes as reptiles, Amniota is divided into the two clades Sauropsida (which contains birds and all living amniotes other than mammals, including all living traditional reptiles) and Theropsida (mammals and the extinct mammal-like reptiles). Similarly, Amphibia can be split into the Batrachomorpha (fossil amphibians more closely related to modern amphibians) and Reptiliomorpha, the latter of which the amiotes is a sub-clade. Ichthyostegalians and other Stem-tetrapods represent sister groups from splits predating the Batrachomorpha/Reptilopmorpha split.

Summary of advantages of phylogenetic nomenclature

Proponents of phylogenetic nomenclature enumerate key distinctions between phylogenetic nomenclature and Linnaean taxonomy as follows:

Phylogenetic Nomenclature	**Linnaean Taxonomy**
Handles arbitrarily deep trees.	Often must invent new level names (such as superorder, suborder, infraorder, parvorder, magnorder) to accommodate new discoveries. Biased towards trees about 4 to 12 levels deep.
Discourages naming or use of groups that are not monophyletic	Acceptable to name and use paraphyletic groups
Primary goal is to reflect actual process of evolution	Primary goal is to group species based on morphological similarities
Assumes that the shape of the tree will change frequently with new discoveries	New discoveries often require renaming or releveling of Classes, Orders and Kingdoms

Summary of criticisms of phylogenetic nomenclature

Critics of phylogenetic nomenclature include Ashlock, Mayr and Williams. Some of their criticisms include:

Phylogenetic Nomenclature	**Linnaean Taxonomy**
Limited to entities related by evolution or ancestry	Supports groupings without reference to evolution or ancestry
Does not include a process for naming species	Includes a process for giving unique names to species

Clade definitions emphasize ancestry at the expense of descriptive characteristics	Taxa definitions based on tangible characteristics
Ignores sensible, clearly defined paraphyletic groups such as reptiles	Permits clearly defined groups such as reptiles
Difficult to determine if a given species is in a clade or not (e.g. if clade X is defined as "most recent common ancestor of A and B along with its descendants", then the only way to determine if species Y is in the clade is to perform a complex evolutionary analysis)	Straightforward process to determine if a given species is in a taxon or not
Limited to organisms that evolved by inherited traits; not applicable to organisms that evolved via complex gene sharing or lateral transfer	Applicable to all organisms, regardless of evolutionary mechanism

Application to other disciplines

The comparisons used to acquire data on which cladograms can be based are not limited to the field of biology. Any group of individuals or classes, hypothesized to have a common ancestor and to which a set of common characteristics may or may not apply, can be compared pairwise. Cladograms can be used to depict the hypothetical descent relationships within groups of items in many different academic realms. The only requirement is that the items have characteristics that can be identified and measured.

Recent attempts to use cladistic methods outside of biology address the reconstruction of lineages in:

- Anthropology and archeology. Compares cultures or artifacts using groups of cultural traits or artifact features.
- Linguistics. Compares languages using groups of linguistic features.
- Textual criticism or Stemmatics. Compares manuscripts of the same work (original lost) using groups of distinctive copying errors.
- Ethology. Compares animal species using behavioral traits presumed hereditary.

Chapter- 5

Phylogenetic Tree

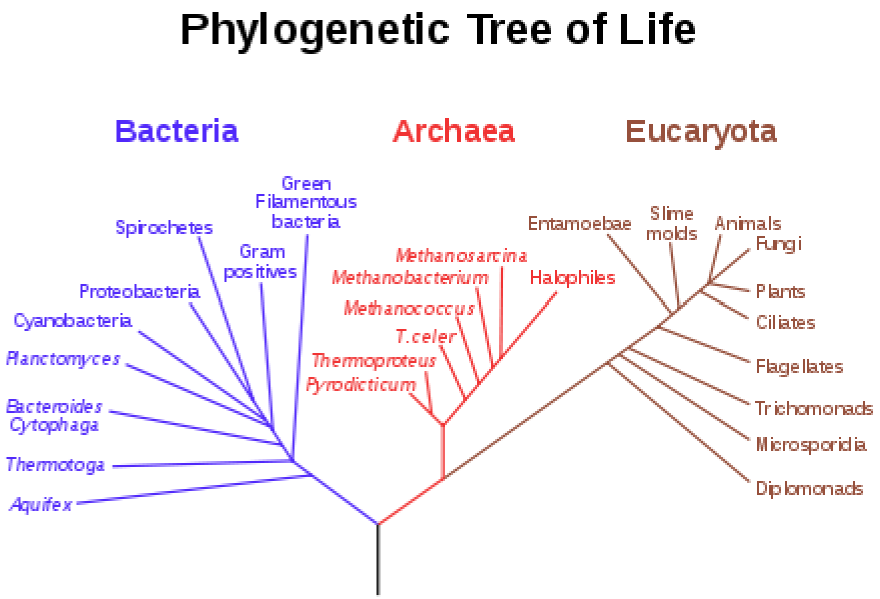


Fig. 1: A speculatively rooted tree for rRNA genes

A **phylogenetic tree** or **evolutionary tree** is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics. The taxa joined together in the tree are implied to have descended from a common ancestor. In a **rooted** phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of the descendants and the edge lengths in some trees may be interpreted as time estimates. Each node is called a taxonomic unit. Internal nodes are generally called hypothetical taxonomic units (HTUs) as they cannot be directly observed. Trees are useful in fields of biology such as systematics and comparative phylogenetics.

History

The idea of a "tree of life" arose from ancient notions of a ladder-like progression from lower to higher forms of life (such as in the Great Chain of Being). Early representations of *branching* phylogenetic trees include a "Paleontological chart" showing the geological relationships among plants and animals in the book Elementary Geology, by Edward Hitchcock (first edition: 1840).

Charles Darwin (1859) also produced one of the first illustrations and crucially popularized the notion of an evolutionary "tree" in his seminal book *The Origin of Species*. Over a century later, evolutionary biologists still use tree diagrams to depict evolution because such diagrams effectively convey the concept that speciation occurs through the adaptive and random splitting of lineages. Over time, species classification has become less static and more dynamic.

Types

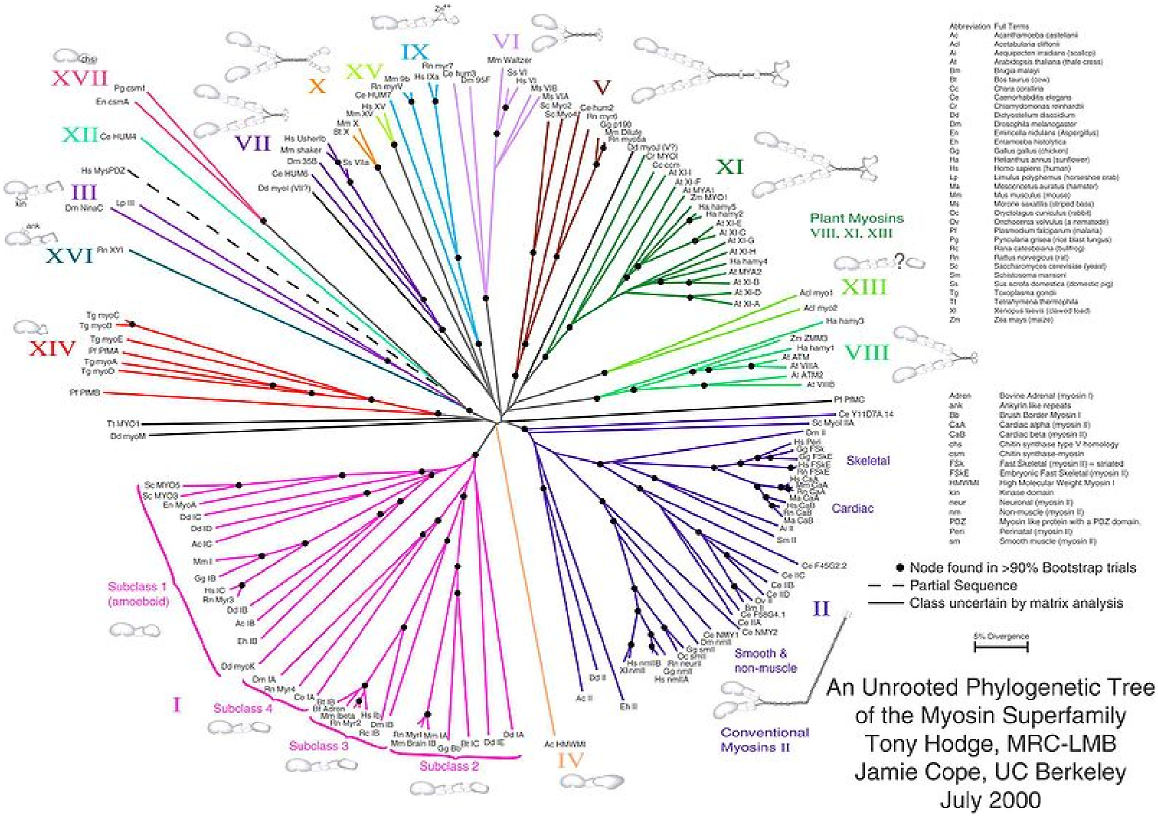


Fig. 1: Unrooted tree of the myosin supergene family

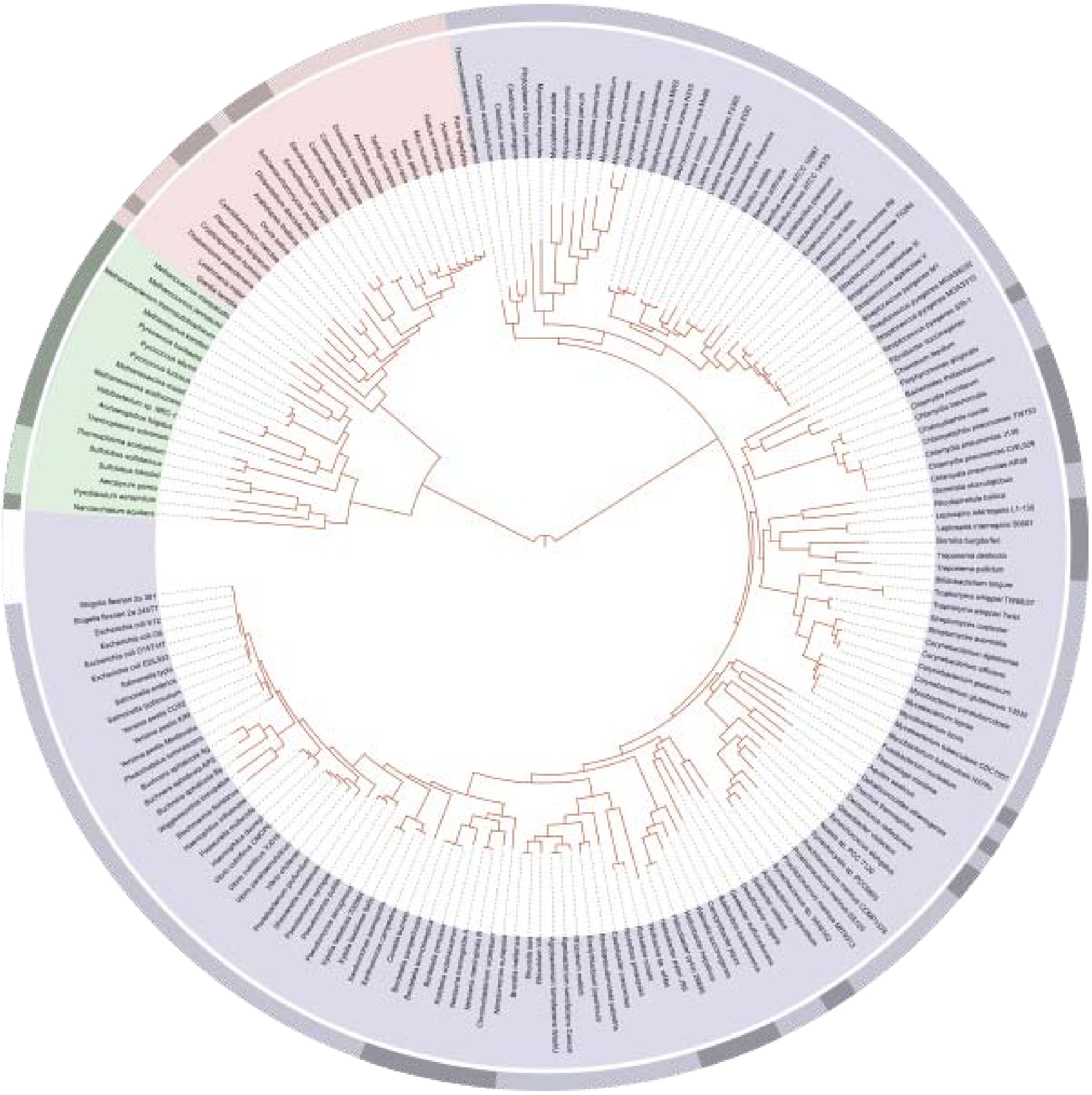
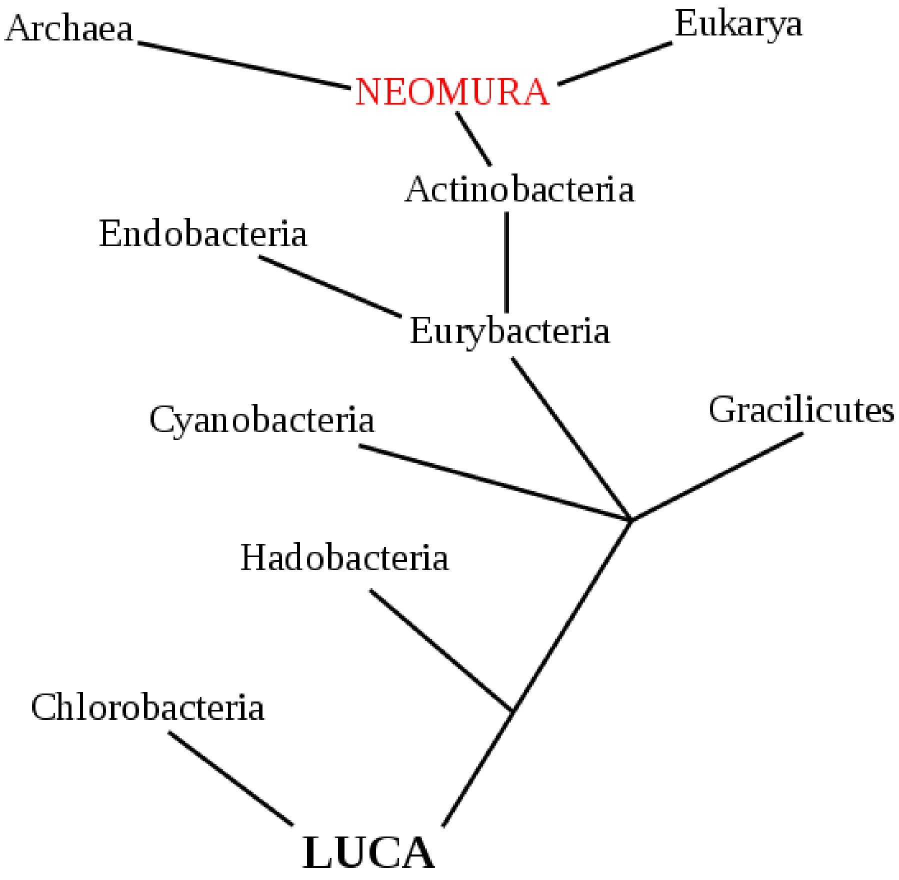

Chlamydophila pneumoniae TW183
Chlamydophila pneumoniae J138
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae AR39
Leptospira interrogans L1-130
Leptospira interrogans 56601
Borrelia burgdorferi
Treponema denticola
Treponema pallidum
Shigella flexneri 2a 301
Shigella flexneri 2a 2457T

Fig. 2: A highly resolved, automatically generated Tree Of Life, based on completely sequenced genomes.



A phylogenetic tree, showing how Eukaryota and Archaea are more closely related to each other than to Bacteria, based on Cavalier-Smith's theory of bacterial evolution. (Cf. LUCA, Neomura.)

A **rooted** phylogenetic tree is a directed tree with a unique node corresponding to the (usually imputed) most recent common ancestor of all the entities at the leaves of the tree. The most common method for rooting trees is the use of an uncontroversial outgroup — close enough to allow inference from sequence or trait data, but far enough to be a clear outgroup.

Unrooted trees illustrate the relatedness of the leaf nodes without making assumptions about ancestry at all. While unrooted trees can always be generated from rooted ones by simply omitting the root, a root cannot be inferred from an unrooted tree without some means of identifying ancestry; this is normally done by including an outgroup in the input data or introducing additional assumptions about the relative rates of evolution on each branch, such as an application of the molecular clock hypothesis. Figure 1 depicts an unrooted phylogenetic tree for myosin, a superfamily of proteins.

Both rooted and unrooted phylogenetic trees can be either **bifurcating** or **multifurcating** and either **labeled** or **unlabeled**. A rooted bifurcating tree has exactly two descendants arising from each interior node (that is, it forms a binary tree) and an unrooted bifurcating tree takes the form of an unrooted binary tree, a free tree with exactly three neighbors at each internal node. In contrast, a rooted multifurcating tree may have more than two children at some nodes and an unrooted multifurcating tree may have more than three neighbors at some nodes. A labeled tree has specific values assigned to its leaves, while an unlabeled tree, sometimes called a **tree shape**, defines a topology only. The number of possible trees for a given number of leaf nodes depends on the specific type of tree, but there are always more multifurcating than bifurcating trees, more labeled than unlabeled trees and more rooted than unrooted trees. The last distinction is the most biologically relevant; it arises because there are many places on an unrooted tree to put the root. For labeled bifurcating trees, there are

$$(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!}\text{ , for} n \geq 2$$

total rooted trees and

$$(2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!}\text{ , for} n \geq 3$$

total unrooted trees, where *n* represents the number of leaf nodes. Among labeled bifurcating trees, the number of unrooted trees with *n* leaves is equal to the number of rooted trees with *n* − 1 leaves.

A **dendrogram** is a broad term for the diagrammatic representation of a phylogenetic tree.

A **cladogram** is a tree formed using cladistic methods. This type of tree only represents a branching pattern, i.e., its branch lengths do not represent time.

A **phylogram** is a phylogenetic tree that explicitly represents number of character changes through its branch lengths.

A **chronogram** is a phylogenetic tree that explicitly represents evolutionary time through its branch lengths.

Construction

Phylogenetic trees among a nontrivial number of input sequences are constructed using computational phylogenetics methods. Distance-matrix methods such as neighbor-joining or UPGMA, which calculate genetic distance from multiple sequence alignments, are simplest to implement, but do not invoke an evolutionary model. Many sequence alignment methods such as ClustalW also create trees by using the simpler algorithms (i.e. those based on distance) of tree construction. Maximum parsimony is another simple method of estimating phylogenetic trees, but implies an implicit model of evolution (i.e. parsimony). More advanced methods use the optimality criterion of maximum likelihood, often within a Bayesian Framework and apply an explicit model of evolution to phylogenetic tree estimation. Identifying the optimal tree using many of these techniques is NP-hard, so heuristic search and optimization methods are used in combination with tree-scoring functions to identify a reasonably good tree that fits the data.

Tree-building methods can be assessed on the basis of several criteria:

- efficiency (how long does it take to compute the answer, how much memory does it need?)
- power (does it make good use of the data, or is information being wasted?)
- consistency (will it converge on the same answer repeatedly, if each time given different data for the same model problem?)
- robustness (does it cope well with violations of the assumptions of the underlying model?)
- falsifiability (does it alert us when it is not good to use, i.e. when assumptions are violated?)

Tree-building techniques have also gained the attention of mathematicians. Trees can also be built using T-theory.

Limitations

Although phylogenetic trees produced on the basis of sequenced genes or genomic data in different species can provide evolutionary insight, they have important limitations. They do not necessarily accurately represent the species evolutionary history. The data on which they are based is noisy; the analysis can be confounded by horizontal gene transfer, hybridisation between species that were not nearest neighbors on the tree before hybridisation takes place, convergent evolution and conserved sequences.

Also, there are problems in basing the analysis on a single type of character, such as a single gene or protein or only on morphological analysis, because such trees constructed from another unrelated data source often differ from the first and therefore great care is needed in inferring phylogenetic relationships among species. This is most true of genetic material that is subject to lateral gene transfer and recombination, where different haplotype blocks can have different histories. In general, the output tree of a phylogenetic analysis is an estimate of the *character'*s phylogeny (i.e. a gene tree) and not the phylogeny of the taxa (i.e. species tree) from which these characters were sampled, though ideally, both should be very close. For this reason, serious phylogenetic studies generally use a combination of genes that come from different genomic sources (e.g., from mitochondrial or plastid vs. nuclear genomes), or genes that would be expected to evolve under different selective regimes, so that homoplasy (false homology) would be unlikely to result from natural selection.

When extinct species are included in a tree, they are terminal nodes, as it is unlikely that they are direct ancestors of any extant species. Scepticism must apply when extinct species are included in trees that are wholly or partly based on DNA sequence data, due to the fact that little useful "ancient DNA" is preserved for longer than 100,000 years and except in the most unusual circumstances no DNA sequences long enough for use in phylogenetic analyses have yet been recovered from material over 1 million years old.

In some organisms, endosymbionts have an independent genetic history from the host.

Phylogenetic networks are used when bifurcating trees are not suitable, due to these complications which suggest a more reticulate evolutionary history of the organisms sampled.

Chapter- 6

Maximum Parsimony

Parsimony is a non-parametric statistical method commonly used in computational phylogenetics for estimating phylogenies. Under parsimony, the preferred phylogenetic tree is the tree that requires the least evolutionary change to explain some observed data.

In detail

Parsimony is part of a class of character-based tree estimation methods which use a matrix of discrete phylogenetic characters to infer one or more optimal phylogenetic trees for a set of taxa, commonly a set of species or reproductively-isolated populations of a single species. These methods operate by evaluating candidate phylogenetic trees according to an explicit optimality criterion; the tree with the most favorable score is taken as the best estimate of the phylogenetic relationships of the included taxa. Maximum parsimony is used with most kinds of phylogenetic data; until recently, it was the only widely-used character-based tree estimation method used for morphological data.

Estimating phylogenies is not a trivial problem. A huge number of possible phylogenetic trees exist for any reasonably sized set of taxa; for example, a mere ten species gives over two million possible unrooted trees. These possibilities must be searched to find a tree that best fits the data according to the optimality criterion. However, the data themselves do not lead to a simple, arithmetic solution to the problem. Ideally, we would expect the distribution of whatever evolutionary characters (such as phenotypic traits or alleles) to directly follow the branching pattern of evolution. Thus we could say that if two organisms possess a shared character, they should be more closely related to each other than to a third organism that lacks this character (provided that character was not present in the last common ancestor of all three, in which case it would be a symplesiomorphy). We would predict that bats and monkeys are more closely related to each other than either is to a fish, because they both possess hair—a synapomorphy. However, we cannot say that bats and monkeys are more closely related to one another than they are to whales because they share hair, because we believe the last common ancestor of the three had hair.

However, the well-understood phenomena of convergent evolution, parallel evolution and evolutionary reversals (collectively termed *homoplasy*) add an unpleasant wrinkle to the problem of estimating phylogeny. For a number of reasons, two organisms can possess a trait not present in their last common ancestor: If we naively took the presence of this trait as evidence of a relationship, we would reconstruct an incorrect tree. Real phylogenetic data include substantial homoplasy, with different parts of the data suggesting sometimes very different relationships. Methods used to estimate phylogenetic trees are explicitly intended to resolve the conflict within the data by picking the phylogenetic tree that is the best fit to all the data overall, accepting that some data simply will not fit. It is often mistakenly believed that parsimony assumes that convergence is rare; in fact, even convergently-derived characters have some value in maximum-parsimony-based phylogenetic analyses and the prevalence of convergence does not systematically affect the outcome of parsimony-based methods.

Data that do not fit a tree perfectly are not simply "noise", they can contain relevant phylogenetic signal in some parts of a tree, even if they conflict with the tree overall. In the whale example given above, the lack of hair in whales is homoplastic: It reflects a return to the condition present in ancient ancestors of mammals, who lacked hair. This similarity between whales and ancient mammal ancestors is in conflict with the tree we accept, since it implies that the mammals with hair should form a group excluding whales. However, among the whales, the reversal to hairlessness actually correctly associates the various types of whales (including dolphins and porpoises) into the group Cetacea. Still, the determination of the best-fitting tree—and thus which data do not fit the tree—is a complex process. Maximum parsimony is one method developed to do this.

Character data

The input data used in a maximum parsimony analysis is in the form of "characters" for a range of taxa. There is no generally agreed-upon definition of a phylogenetic character, but operationally a character can be thought of as an attribute, an axis along which taxa are observed to vary. These attributes can be physical (morphological), molecular, genetic, physiological, or behavioral. The only widespread agreement on characters seems to be that variation used for character analysis should reflect heritable variation. Whether it must be directly heritable, or whether indirect inheritance (e.g., learned behaviors) is acceptable, is not entirely resolved.

Each character is divided into discrete character states, into which the variations observed are classified. Character states are often formulated as descriptors, describing the condition of the character substrate. For example, the character "eye color" might have the states "blue" and "brown." Characters can have two or more states (they can have only one, but these characters lend nothing to a maximum parsimony analysis and are often excluded).

Coding characters for phylogenetic analysis is not an exact science and there are numerous complicating issues. Typically, taxa are scored with the same state if they are more similar to one another in that particular attribute than each is to taxa scored with a different state. This is not straightforward when character states are not clearly delineated or when they fail to capture all of the possible variation in a character. How would one score the previously mentioned character for a taxon (or individual) with hazel eyes? Or green? As noted above, character coding is generally based on similarity: Hazel and green eyes might be lumped with blue because they are more similar to that color (being light) and the character could be then recoded as "eye color: light; dark." Alternately, there can be multi-state characters, such as "eye color: brown; hazel, blue; green."

Ambiguities in character state delineation and scoring can be a major source of confusion, dispute and error in phylogenetic analysis using character data. Note that, in the above example, "eyes: present; absent" is also a possible character, which creates issues because "eye color" is not applicable if eyes are not present. For such situations, a "?" ("unknown") is scored, although sometimes "X" or "-" (the latter usually in sequence data) are used to distinguish cases where a character cannot be scored from a case where the state is simply unknown. Current implementations of maximum parsimony generally treat unknown values in the same manner: the reasons the data are unknown have no particular effect on analysis. Effectively, the program treats a ? as if it held the state that would involve the fewest extra steps in the tree (see below), although this is not an explicit step in the algorithm.

Genetic data are particularly amenable to character-based phylogenetic methods such as maximum parsimony because protein and nucleotide sequences are naturally discrete: A particular position in a nucleotide sequence can be either adenine, cytosine, guanine, or thymine / uracil, or a sequence gap; a position (residue) in a protein sequence will be one of the basic amino acids or a sequence gap. Thus, character scoring is rarely ambiguous, except in cases where sequencing methods fail to produce a definitive assignment for a particular sequence position. Sequence gaps are sometimes treated as characters, although there is no consensus on how they should be coded.

Characters can be treated as unordered or ordered. For a binary (two-state) character, this makes little difference. For a multi-state character, unordered characters can be thought of as having an equal "cost" (in terms of number of "evolutionary events") to change from any one state to any other; complementarily, they do not require passing through intermediate states. Ordered characters have a particular sequence in which the states must occur through evolution, such that going between some states requires passing through an intermediate. This can be thought of complementarily as having different costs to pass between different pairs of states. In the eye-color example above, it is possible to leave it unordered, which imposes the same evolutionary "cost" to go from brown-blue, green-blue, green-hazel, etc. Alternately, it could be ordered brown-hazel-green-blue; this would normally imply that it would cost two evolutionary events to go from brown-green, three from brown-blue, but only one from brown-hazel. This can also be thought of as requiring eyes to evolve through a "hazel stage" to get from brown to green and a "green stage" to get from hazel to blue, etc.

There is a lively debate on the utility and appropriateness of character ordering, but no consensus. Some authorities order characters when there is a clear logical, ontogenetic, or evolutionary transition among the states (for example, "legs: short; medium; long"). Some accept only some of these criteria. Some run an unordered analysis and order characters that show a clear order of transition in the resulting tree (which practice might be accused of circular reasoning). Some authorities refuse to order characters at all, suggesting that it biases an analysis to require evolutionary transitions to follow a particular path.

It is also possible to apply differential weighting to individual characters. This is usually done relative to a "cost" of 1. Thus, some characters might be seen as more likely to reflect the true evolutionary relationships among taxa and thus they might be weighted at a value 2 or more; changes in these characters would then count as two evolutionary "steps" rather than one when calculating tree scores (see below). There has been much discussion in the past about character weighting. Most authorities now weight all characters equally, although exceptions are common. For example, allele frequency data is sometimes pooled in bins and scored as an ordered character. In these cases, the character itself is often downweighted so that small changes in allele frequencies count less than major changes in other characters. Also, the third codon position in a coding nucleotide sequence is particularly labile and is sometimes downweighted, or given a weight of 0, on the assumption that it is more likely to exhibit homoplasy. In some cases, repeated analyses are run, with characters reweighted in inverse proportion to the degree of homoplasy discovered in the previous analysis (termed successive weighting); this is another technique that might be considered circular reasoning.

Character state changes can also be weighted individually. This is often done for nucleotide sequence data; it has been empirically determined that certain base changes (A-C, A-T, G-C, G-T and the reverse changes) occur much less often than others. These changes are therefore often weighted more. As shown above in the discussion of character ordering, ordered characters can be thought of as a form of character state weighting.

Some systematists prefer to exclude characters known to be, or suspected to be, highly homoplastic or that have a large number of unknown entries ("?"). As noted below, theoretical and simulation work has demonstrated that this is likely to sacrifice accuracy rather than improve it. This is also the case with characters that are variable in the terminal taxa: theoretical, congruence and simulation studies have all demonstrated that such polymorphic characters contain significant phylogenetic information.

Taxon sampling

The time required for a parsimony analysis (or any phylogenetic analysis) is proportional to the number of taxa (and characters) included in the analysis. Also, because more taxa require more branches to be estimated, more uncertainty may be expected in large analyses. Because data collection costs in time and money often scale directly with the number of taxa included, most analyses include only a fraction of the taxa that could have been sampled. Indeed, some authors have contended that four taxa (the minimum required to produce a meaningful unrooted tree) are all that is necessary for accurate phylogenetic analysis and that more characters are more valuable than more taxa in phylogenetics. This has led to a raging controversy about taxon sampling.

Empirical, theoretical and simulation studies have led to a number of dramatic demonstrations of the importance of adequate taxon sampling. Most of these can be summarized by a simple observation: a phylogenetic data matrix has dimensions of characters *times* taxa. Doubling the number of taxa doubles the amount of information in a matrix just as surely as doubling the number of characters. Each taxon represents a new sample for every character, but, more importantly, it (usually) represents a new *combination* of character states. These character states can not only determine where that taxon is placed on the tree, they can inform the entire analysis, possibly causing different relationships among the remaining taxa to be favored by changing estimates of the pattern of character changes.

The most disturbing weakness of parsimony analysis, that of long-branch attraction (see below) is particularly pronounced with poor taxon sampling, especially in the four-taxon case. This is a well-understood case in which additional character sampling may not improve the quality of the estimate. As taxa are added, they often break up long branches (especially in the case of fossils), effectively improving the estimation of character state changes along them. Because of the richness of information added by taxon sampling, it is even possible to produce highly accurate estimates of phylogenies with hundreds of taxa using only a few thousand characters.

Although many studies have been performed, there is still much work to be done on taxon sampling strategies. Because of advances in computer performance and the reduced cost and increased automation of molecular sequencing, sample sizes overall are on the rise and studies addressing the relationships of hundreds of taxa (or other terminal entities, such as genes) are becoming common. Of course, this is not to say that adding characters is not also useful; the number of characters is increasing as well.

Some systematists prefer to exclude taxa based on the number of unknown character entries ("?") they exhibit, or because they tend to "jump around" the tree in analyses (i.e., they are "wildcards"). As noted below, theoretical and simulation work has demonstrated that this is likely to sacrifice accuracy rather than improve it. Although these taxa may generate more most-parsimonious trees (see below), methods such as agreement subtrees and reduced consensus can still extract information on the relationships of interest.

It has been observed that inclusion of more taxa tends to lower overall support values (bootstrap percentages or decay indices, see below). The cause of this is clear: as additional taxa are added to a tree, they subdivide the branches to which they attach and thus dilute the information that supports that branch. While support for individual branches is reduced, support for the overall relationships is actually increased. Consider analysis that produces the following tree: (fish, (lizard,(whale, (cat, monkey)))). Adding a rat and a walrus will probably reduce the support for the (whale, (cat, monkey)) clade, because the rat and the walrus may fall within this clade, or outside of the clade and since these five animals are all relatively closely related, there should be more uncertainty about their relationships. Within error, it may be impossible to determine any of these animals' relationships relative to one another. However, the rat and the walrus will probably add character data that cements the grouping any two of these mammals exclusive of the fish or the lizard; where the initial analysis might have been mislead, say, by the presence of fins in the fish and the whale, the presence of the walrus, with blubber and fins like a whale but whiskers like a cat and a rat, firmly ties the whale to the mammals.

To cope with this problem, agreement subtrees, reduced consensus and double-decay analysis seek to identify supported relationships (in the form of "n-taxon statements," such as the four-taxon statement "(fish, (lizard, (cat, whale)))") rather than whole trees. If the goal of an analysis is a resolved tree, as is the case for comparative phylogenetics, these methods cannot solve the problem. However, if the tree estimate is so poorly supported, the results of any analysis derived from the tree will probably be too suspect to use anyway.

Analysis

A maximum parsimony analysis runs in a very straightforward fashion. Trees are scored according to the degree to which they imply a parsimonious distribution of the character data. The most parsimonious tree for the dataset represents the preferred hypothesis of relationships among the taxa in the analysis.

Trees are scored (evaluated) by using a simple algorithm to determine how many "steps" (evolutionary transitions) are required to explain the distribution of each character. A step is, in essence, a change from one character state to another, although with ordered characters some transitions require more than one step. Contrary to popular belief, the algorithm does not explicitly assign particular character states to nodes (branch junctions) on a tree: the least number of steps can involve multiple, equally costly assignments and distributions of evolutionary transitions. What is optimized is the total number of changes.

There are many more possible phylogenetic trees than can be searched exhaustively for more than eight taxa or so. A number of algorithms are therefore used to searching among the possible trees. Many of these involve taking an initial tree (usually the favored tree from the last iteration of the algorithm) and perturbing it to see if the change produces a higher score.

The trees resulting from parsimony search are unrooted: They show all the possible relationships of the included taxa, but they lack any statement on relative times of divergence. A particular branch is chosen to root the tree by the user. This branch is then taken to be outside all the other branches of the tree, which together form a monophyletic group. This imparts a sense of relative time to the tree. Incorrect choice of a root can result in incorrect relationships on the tree, even if the tree is itself correct in its unrooted form.

Parsimony analysis often returns a number of equally most-parsimonious trees (MPTs). A large number of MPTs is often seen as an analytical failure and is widely believed to be related to the number of missing entries ("?") in the dataset, characters showing too much homoplasy, or the presence of topologically labile "wildcard" taxa (which may have many missing entries). Numerous methods have been proposed to reduce the number of MPTs, including removing characters or taxa with large amounts of missing data before analysis, removing or downweighting highly homoplastic characters (successive weighting) or removing wildcard taxa (the phylogenetic trunk method) *a posteriori* and then reanalyzing the data.

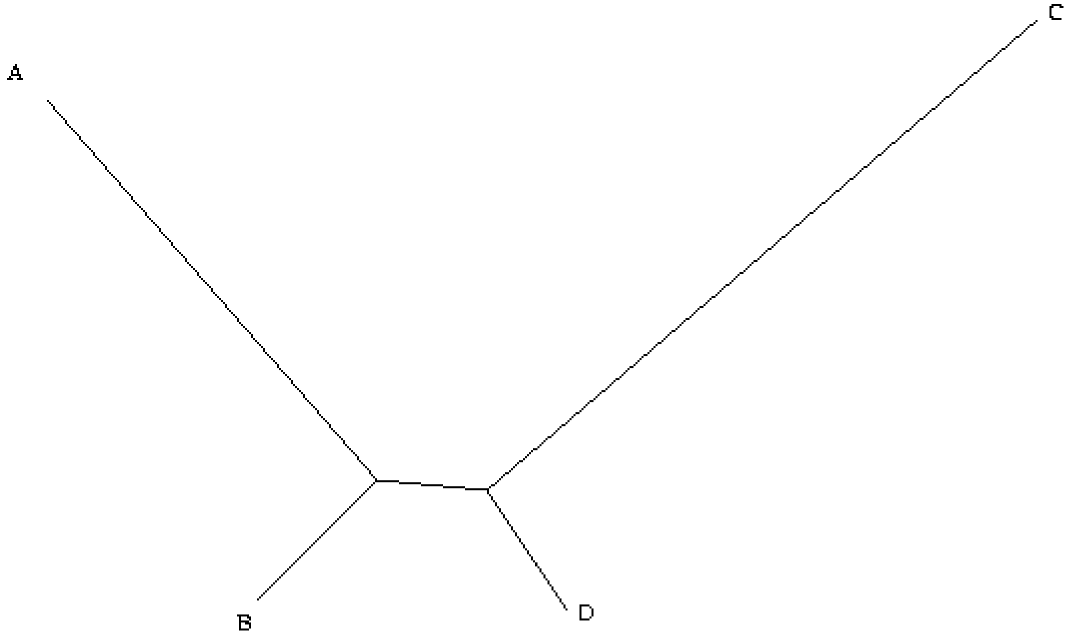
Numerous theoretical and simulation studies have demonstrated that highly homoplastic characters, characters and taxa with abundant missing data and "wildcard" taxa contribute to the analysis. Although excluding characters or taxa may appear to improve resolution, the resulting tree is based on less data and is therefore a less reliable estimate of the phylogeny. Today's general consensus is that having multiple MPTs is a valid analytical result; it simply indicates that there is insufficient data to resolve the tree completely. In many cases, there is substantial common structure in the MPTs and differences are slight and involve uncertainty in the placement of a few taxa. There are a number of methods for summarizing the relationships within this set, including consensus trees, which show common relationships among all the taxa and pruned agreement subtrees, which show common structure by temporarily pruning "wildcard" taxa from every tree until they all agree. Reduced consensus takes this one step further, but showing all subtrees (and therefore all relationships) supported by the input trees.

Even if multiple MPTs are returned, parsimony analysis still basically produces a point-estimate, lacking confidence intervals of any sort. This has often been levelled as a criticism, since there is certainly error in estimating the most-parsimonious tree and the method does not inherently include any means of establishing how sensitive its conclusions are to this error. Several methods have been used to assess support.

Jackknifing and bootstrapping, well-known statistical resampling procedures, have been employed with parsimony analysis. The jackknife, which involves resampling without replacement ("leave-one-out") can be employed on characters or taxa; interpretation may become complicated in the latter case, because the variable of interest is the tree and comparison of trees with different taxa is not straightforward. The bootstrap, resampling with replacement (sample x items randomly out of a sample of size x, but items can be picked multiple times), is only used on characters, because adding duplicate taxa does not change the result of a parsimony analysis. The bootstrap is much more commonly employed in phylogenetics (as elsewhere); both methods involve an arbitrary but large number of repeated iterations involving perturbation of the original data followed by analysis. The resulting MPTs from each analysis are pooled and the results are usually presented on a 50% Majority Rule Consensus tree, with individual branches (or nodes) labelled with the percentage of bootstrap MPTs in which they appear. This "bootstrap percentage" (which is not a P-value, as is sometimes claimed) is used as a measure of support. Technically, it is supposed to be a measure of repeatability, the probability that that branch (node, clade) would be recovered if the taxa were sampled again. Experimental tests with viral phylogenies suggest that the bootstrap percentage is not a good estimator of repeatability for phylogenetics, but it is a reasonable estimator of accuracy. In fact, it has been shown that the bootstrap percentage, as an estimator of accuracy, is biased and that this bias results on average in an underestimate of confidence (such that as little as 70% support might really indicate up to 95% confidence). However, the direction of bias cannot be ascertained in individual cases, so assuming that high values bootstrap support indicate even higher confidence is unwarranted.

Another means of assessing support is Bremer support, or the decay index (which is technically not an index). This is simply the difference in number of steps between the score of the MPT(s) and the score of the most parsimonious tree that does NOT contain a particular clade (node, branch). It can be thought of as the number of steps you have to add to lose that clade; implicitly, it is meant to suggest how great the error in the estimate of the score of the MPT must be for the clade to no longer be supported by the analysis, although this is not necessarily what it does. Decay index values are often fairly low (one or two steps being typical), but they often appear to be proportional to bootstrap percentages. However, interpretation of decay values is not straightforward and they seem to be preferred by authors with philosophical objections to the bootstrap (although many morphological systematists, especially paleontologists, report both). Double-decay analysis is a decay counterpart to reduced consensus that evaluates the decay index for all possible subtree relationships (n-taxon statements) within a tree.

Problems with maximum parsimony phylogeny estimation



An example of long branch attraction. Branches A & C have a high number of substitutions

Maximum parsimony is a very simple approach and is popular for this reason. However, it is not statistically consistent. That is, it is not guaranteed to produce the true tree with high probability, given sufficient data. Consistency, here meaning the monotonic convergence on the correct answer with the addition of more data, is a desirable property of any statistical method. As demonstrated in 1978 by Joe Felsenstein, maximum parsimony can be inconsistent under certain conditions. The category of situations in which this is known to occur is called *long branch attraction* and occurs, for example, where there are long branches (a high level of substitutions) for two characters (A & C), but short branches for another two (B & D). A and B diverged from a common ancestor, as did C and D.

Assume for simplicity that we are considering a single binary character (it can either be + or -). Because the distance from B to D is small, in the vast majority of all cases, B and D will be the same. Here, we will assume that they are both + (+ and - are assigned arbitrarily and swapping them is only a matter of definition). If this is the case, there are four remaining possibilities. A and C can both be +, in which case all taxa are the same and all the trees have the same length. A can be + and C can be -, in which case only one character is different and we cannot learn anything, as all trees have the same length. Similarly, A can be - and C can be +. The only remaining possibility is that A and C are both -. In this case, however, we group A and C together and B and D together. As a consequence, when we have a tree of this type, the more data we collect (i.e. the more characters we study), the more we tend towards the wrong tree.

A simple and effective method for determining whether or not long branch attraction is affecting tree topology is the SAW method, named for Siddal and Whiting. If long branch attraction is suspected in a pair of taxa (A and B), simply remove taxon A ("saw" off the branch) and re-run the analysis. Then remove A and replace B, running the analysis again. If either of the taxa appear at different branch points in the absence of the other, there is evidence of long branch attraction. Since long branches can't possibly attract one another when only one is in the analysis, consistent taxon placement between treatments would indicate long branch attraction is not a problem (Siddal & Whiting, 1999).

Several other methods of phylogeny estimation are available, including maximum likelihood, Bayesian phylogeny inference, neighbour joining and quartet methods. Of these, the first two both use a likelihood function and, if used properly, are theoretically immune to long-branch attraction. These methods are both parametric, meaning that they rely on an explicit model of character evolution. It has been shown that, for some suboptimal models, these methods can also be inconsistent.

Another complication with maximum parsimony is that finding the most parsimonious tree is an NP-Hard problem. The only currently available, efficient way of obtaining a solution, given an arbitrarily large set of taxa, is by using heuristic methods which do not guarantee that the most parsimonious tree will be recovered. These methods employ hill-climbing algorithms to progressively approach the best tree. However, it has been shown that there can be "tree islands" of suboptimal solutions and the analysis can become trapped in these local optima. Thus, complex, flexible heuristics are required to ensure that tree space has been adequately explored. Several heuristics are available, including nearest neighbor interchange (NNI), tree bisection / reconnection (TBR) and the phylogenetic ratchet. This problem is certainly not unique to MP; any method that uses an optimality criterion faces the same problem and none offer easy solutions.

Criticism

It has been asserted that a major problem, especially for paleontology, is that maximum parsimony assumes that the only way two species can share the same character is if they are genetically related. Although this statement is confusingly worded, it appears to assert that phylogenetic applications of parsimony assume that all similarity is homologous (other interpretations, such as the assertion that two organisms might NOT be related at all, are nonsensical). This is emphatically not the case: as with any form of character-based phylogeny estimation, parsimony is used to test the homologous nature of similarities by finding the phylogenetic tree which best accounts for all of the similarities.

For example, birds and bats have wings, while crocodiles and humans do not. If these were the only data available, maximum parsimony would tend to group crocodiles with humans and birds with bats (as would any other method of phylogenetic inference). We believe that humans are actually more closely related to bats than to crocodiles or birds. Our belief is founded on additional data that was not considered in the one-character example (using wings). If even a tiny fraction of these additional data, including information on skeletal structure, soft-tissue morphology, integument, behaviour, genetics, etc., were included in the analysis, the faint phylogenetic signal produced by the presence of wings in birds and bats would be overwhelmed by the preponderance of data supporting the (human, bat)(bird, crocodile) tree.

It is often stated that parsimony is not relevant to phylogenetic inference because "evolution is not parsimonious." In most cases, there is no explicit alternative proposed; if no alternative is available, any statistical method is preferable to none at all. Additionally, it is not clear what would be meant if the statement "evolution is parsimonious" were in fact true. This could be taken to mean that more character changes may have occurred historically than are predicted using the parsimony criterion. Because parsimony phylogeny estimation reconstructs the minimum number of changes necessary to explain a tree, this is quite possible. However, it has been shown through simulation studies, testing with known in vitro viral phylogenies and congruence with other methods, that the accuracy of parsimony is in most cases not compromised by this. Parsimony analysis uses the number of character changes on trees to choose the best tree, but it does not require that exactly that many changes and no more, produced the tree. As long as the changes that have not been accounted for are randomly distributed over the tree (a reasonable null expectation), the result should not be biased. In practice, the technique is robust: maximum parsimony exhibits minimal bias as a result of choosing the tree with the fewest changes.

An analogy can be drawn with choosing among contractors based on their initial (nonbinding) estimate of the cost of a job. The actual finished cost is very likely to be higher than the estimate. Despite this, choosing the contractor who furnished the lowest estimate should theoretically result in the lowest final project cost. This is because, in the absence of other data, we would assume that all of the relevant contractors have the same risk of cost overruns. In practice, of course, unscrupulous business practices may bias this result; in phylogenetics, too, some particular phylogenetic problems (for example, long branch attraction, described above) may potentially bias results. In both cases, however, there is no way to tell if the result is going to be biased, or the degree to which it will be biased, based on the estimate itself. With parsimony too, there is no way to tell that the data are positively misleading, without comparison to other evidence.

Along the same lines, parsimony is often characterized as implicitly adopting the philosophical position that evolutionary change is rare, or that homoplasy (convergence and reversal) is minimal in evolution. This is not entirely true: parsimony minimizes the number of convergences and reversals that are assumed by the preferred tree, but this may result in a relatively large number of such homoplastic events. It would be more appropriate to say that parsimony assumes only the minimum amount of change implied by the data. As above, this does not require that these were the only changes that occurred; it simply does not infer changes for which there is no evidence. The shorthand for describing this is that "parsimony minimizes assumed homoplasies, it does not assume that homoplasy is minimal."

Parsimony is also sometimes associated with the notion that "the simplest possible explanation is the best," a generalisation of Occam's Razor. Parsimony does prefer the solution that requires the least number of unsubstantiated assumptions and unsupportable conclusions, the solution that goes the least theoretical distance beyond the data. This is a very common approach to science, especially when dealing with systems that are so complex as to defy simple models. Parsimony does not by any means necessarily produce a "simple" assumption. Indeed, as a general rule, most character datasets are so "noisy" that no truly "simple" solution is possible.

Alternatives

There are several other methods for inferring phylogenies based on discrete character data. Each offers potential advantages and disadvantages. Most of these methods have particularly avid proponents and detractors; parsimony especially has been advocated as philosophically superior (most notably by ardent cladists).

Maximum likelihood

Among the most popular alternative phylogenetic methods is maximum likelihood phylogenetic inference, sometimes simply called "likelihood" or "ML." Maximum likelihood is an optimality criterion, as is parsimony. Mechanically, maximum likelihood analysis functions much like parsimony analysis, in that trees are scored based on a character dataset and the tree with the best score is selected. Maximum likelihood is a parametric statistical method, in that it employs an explicit model of character evolution. Such methods are potentially much more powerful than non-parametric statistical methods like parsimony, but only if the model used is a reasonable approximation of the processes that produced the data. Maximum likelihood has probably surpassed parsimony in popularity with nucleotide sequence data and Bayesian phylogenetic inference, which uses the likelihood function, is becoming almost as prevalent.

Likelihood is the relative counterpart to absolute probability. If we know the number of possible outcomes of a test (N) and we know the number of those outcomes that fit a particular criterion (n), we can say that the probability of that criterion being met by an execution of that test is n/N. Thus, the probability of heads in the toss of a fair coin is 50% (1/2). What if we don't know the number of possible outcomes? Obviously, we cannot then calculate probabilities. However, if we observe that one outcome happens twice as often as the other over an arbitrarily large number of tests, we can say that that outcome is twice as likely. Likelihoods are proportional to the true probabilities: if an outcome is twice as likely, we can say that it is twice as probable, even though we cannot say how probable it is.

Practically, the probability of a tree cannot be calculated directly. The probability of the data given a tree can be calculated if you assume a specific set of probabilities of character change (a model). The critical part of likelihood analysis is that the probability of the data given the tree is the likelihood of the tree given the data. Thus, the tree that has the highest probability of producing the observed data is the most likely tree.

Maximum likelihood, as implemented in phylogenetics, uses a stochastic model that gives the probability of a particular character changing at any given point on a tree. This model can have a potentially large number of parameters, which can account for differences in the probabilities of particular states, the probabilities of particular changes and differences in the probabilities of change among characters.

A likelihood tree has meaningful branch lengths (i.e. it is a phylogram); these lengths are usually interpreted as being proportional to the average probability of change for characters on that branch (thus, on a branch of length 1, we would expect an average of one change per character, which is a lot). The state of each character is plotted on the tree and the probability of that distribution of character states is calculated using the model and the branch lengths (which can be altered to maximize the probability of the data). This is the probability of that character, given the tree. The probabilities of all of the characters is multiplied together; they are usually negative log-transformed and added (producing the same effect), because the numbers become very small very quickly. This sum is the probability of the data, given the tree, or the likelihood of the tree. The tree with the highest likelihood (lowest negative log-transformed likelihood) given the data is preferred.

In the above analogy regarding choosing a contractor, maximum likelihood would be analogous to gathering data on the final cost of broadly comparable jobs performed by each contractor over the past year and selecting the contractor with the lowest average cost for those comparable jobs. This method would be highly dependent on how comparable the jobs are, but, if they are properly chosen, it will produce a better estimate of the actual cost of the job. Further, it would not be mislead by bias in contractor estimates, because it is based on the final cost, not on the (potentially biased) estimates.

In practice, maximum likelihood tends to favor trees that are very similar to the most parsimonious tree(s) for the same dataset. It has been shown to outperform parsimony in certain situations where the latter is known to be biased, including long-branch attraction. Note, however, that the performance of likelihood is dependent on the quality of the model employed; an incorrect model can produce a biased result. Studies have shown that incorporating a parameter to account for differences in rate of evolution among characters is often critical to accurate estimation of phylogenies; failure to model this or other crucial parameters may produce incorrect or biased results. Model parameters are usually estimated from the data and the number (and type) of parameters is often determined using the hierarchical likelihood ratio test. The consequences of mis-specified models are just beginning to be explored in detail.

Likelihood is generally regarded as a more desirable method than parsimony, in that it is statistically consistent and has a better statistical foundation and because it allows complex modeling of evolutionary processes. A major drawback is that ML is still quite slow relative to parsimony methods, sometimes requiring days to run large datasets. Maximum likelihood phylogenetic inference was proposed in the mid-Twentieth Century, but it has only been a popular method for phylogenetic inference since the 1990s, when computational power caught up with tremendous demands of ML analysis. Newer algorithms and implementations are bringing analysis times for large datasets into acceptable ranges. Until these methods gain widespread acceptance, parsimony will probably be preferred for extremely large datasets, especially when bootstrapping is used to assess confidence in the results.

One area where parsimony still holds much sway is in the analysis of morphological data. Until recently, stochastic models of character change were not available for non-molecular data. New methods, proposed by Paul Lewis, make essentially the same assumptions that parsimony analysis does, but do so within a likelihood framework. These models are not, however, widely implemented and, unless modified, they require the modification of existing datasets (to deal with ordered characters and the tendency to not record autapomorphies in morphological datasets.

Maximum likelihood has been criticised as assuming neutral evolution implicitly in its adoption of a stochastic model of evolution. This is not necessarily the case: as with parsimony, assuming a stochastic model does not presume that all evolution is stochastic. In practice, likelihood is robust to deviations from stochasticity. It performs well even on coding sequences that include sites believed to be under selection.

A related objection (often brought up by parsimony-only advocates) is the idea that evolution is too complex or too poorly understood to be modeled. This objection probably rests on a misunderstanding of the term "model." While it is customary to think of models as representing the mechanics of a process, this is not necessarily literally the case. In fact, a model is often selected not so much for its faithful reproduction of the phenomenon as its ability to make predictions. In practice, it is best not to try and exactly fit a model to a process, because there is a trade-off between number of parameters in a model and its statistical power. Stochasticity may be a reasonably good fit to evolutionary data at a broad level, even if it does not accurately mirror the process at finer scales.

By analogy, no one claims that the human foot varies only in length and width, but differing combinations of length and width values can be combined to fit a wide variety of feet. In some cases, a slightly wider overall foot may be better fitted by increasing overall size rather than instep width, while a foot with a narrower heel might be better fit by a wider instep and a smaller shoe. Adding several more measurements would probably improve shoe fit somewhat, but would be impractical from a business standpoint. With increasingly precise fitting, differences between feet would make selling matched pairs of shoes impossible and differences through time would mean that a proper fit at purchase might not be a proper fit when worn.

Parsimony has recently been shown to be more likely to recover the true tree in the face of profound changes in evolutionary ("model") parameters (e.g., the rate of evolutionary change) within a tree (Kolaczkowski & Thornton 2004). This is particularly troublesome, since it is generally agreed that such changes may be a significant feature of deep divergences. Likelihood has had substantial success recovering known in vitro viral phylogenies, simulated phylogenies and phylogenies confirmed by other method. It seems likely therefore that this potential complication does not strongly bias results for more shallow divergences. Several research groups are currently exploring ways to incorporate profound shifts in evolutionary parameters into likelihood analysis.

Bayesian phylogenetic inference

Bayesian phylogenetics uses the likelihood function and is normally implemented using the same models of evolutionary change used in Maximum Likelihood. It is very different, however, in both theory and application. Bayesian phylogenetic analysis uses Bayes' theorem, which relates the posterior probability of a tree to the likelihood of data and the prior probability of the tree and model of evolution. However, unlike parsimony and likelihood methods, Bayesian analysis does not produce a single tree or set of equally optimal trees. Bayesian analysis uses the likelihood of trees in a Markov chain Monte Carlo (MCMC) simulation to sample trees in proportion to their likelihood, thereby producing a credible sample of trees.

One commonly cited drawback of Bayesian analysis is the need to explicitly set out a set of prior probabilities for the range of potential outcomes. The idea of incorporating prior probabilities into an analysis has been suggested as a potential source of bias. Bayesian methods involve other potential issues, such as the evaluation of "convergence," the point at which the MCMC process stops searching for the "space" of credible solutions and begins to build the credible sample.

Distance matrix methods

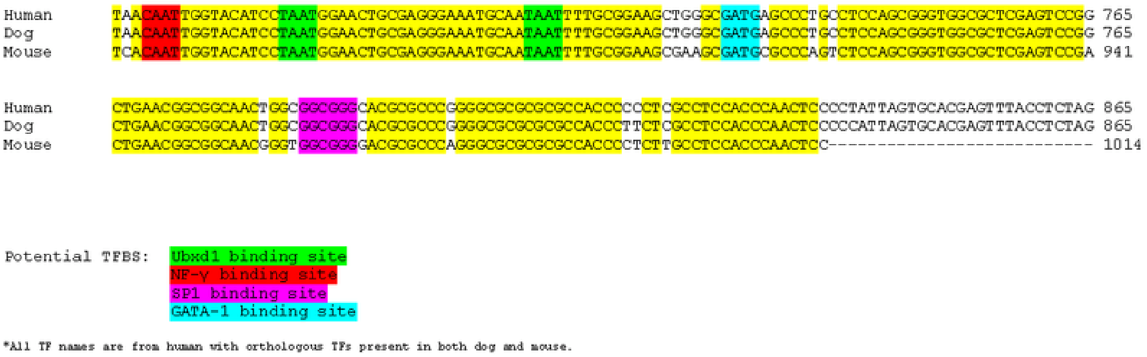
Non-parametric distance methods were originally applied to phenetic data using a matrix of pairwise distances. These distances are then reconciled to produce a tree (a phylogram, with informative branch lengths). The distance matrix can come from a number of different sources, including measured distance (for example from immunological studies) or morphometric analysis, various pairwise distance formulae (such as euclidean distance) applied to discrete morphological characters, or genetic distance from sequence, restriction fragment, or allozyme data. For phylogenetic character data, raw distance values can be calculated by simply counting the number of pairwise differences in character states (Manhattan distance).

Several simple algorithms exist to construct a tree directly from pairwise distances, including UPGMA and neighbor joining (NJ), but these will not necessarily produce the best tree for the data. UPGMA assumes an ultrametric tree (a tree where all the path-lengths from the root to the tips are equal). Neighbor-joining is a form of star decomposition and can very quickly produce reasonable trees. It is very often used on its own and in fact quite frequently produces reasonable trees.

Phylogeny estimation using distance methods has produced a number of controversies. The relationship between individual characters and the tree is lost in the process of reducing characters to distances. Since these methods do not use character data directly and information locked in the distribution of character states can be lost in the pairwise comparisons. Also, some complex phylogenetic relationships may produce biased distances. Despite these potential problems, distance methods are extremely fast and they often produce a reasonable estimate of phylogeny. They also have certain benefits over the methods that use characters directly. Notably, distance methods allow use of data that may not be easily converted to character data, such as DNA-DNA hybridization assays.

Chapter- 7

Phylogenetic Footprinting



```
Human      TAACAATTGGTACATCCTAATGGAACTGCGAGGGAAATGCAATAATTTTGCGGAAGCTGGGCGATGAGCCCTGCCTCCAGCGGGTGGCGCTCGAGTCCGG 765
Dog        TAACAATTGGTACATCCTAATGGAACTGCGAGGGAAATGCAATAATTTTGCGGAAGCTGGGCGATGAGCCCTGCCTCCAGCGGGTGGCGCTCGAGTCCGG 765
Mouse      TCACAATTGGTACATCCTAATGGAACTGCGAGGGAAATGCAATAATTTTGCGGAAGCGAAGCGATGCGCCCAGTCTCCAGCGGGTGGCGCTCGAGTCCGA 941

Human      CTGAACGGCGGCAACTGGCGGCGGGCACGCGCCCGGGGCGCGCGCGCCACCCCCCTCGCCTCCACCCAACTCCCCTATTAGTGCACGAGTTTACCTCTAG 865
Dog        CTGAACGGCGGCAACTGGCGGCGGGCACGCGCCCGGGGCGCGCGCGCCACCCTTCTCGCCTCCACCCAACTCCCCCATTAGTGCACGAGTTTACCTCTAG 865
Mouse      CTGAACGGCGGCAACGGGTGGCGGGGACGCGCCCAGGGCGCGCGCGCCACCCCTCTTGCCTCCACCCAACTCC--------------------------- 1014


Potential TFBS:  Ubxd1 binding site
                 NF-y binding site
                 SP1 binding site
                 GATA-1 binding site

*All TF names are from human with orthologous TFs present in both dog and mouse.
```

Phylogenetic Footprinting of HOXA5 gene

Phylogenetic footprinting is a technique used to identify transcription factor binding sites (TFBS) within a non-coding region of DNA of interest by comparing it to the orthologous sequence in different species. When this technique is used with a large number of closely related species, this is called **phylogenetic shadowing**.

Researchers have found that non-coding pieces of DNA contain binding sites for regulatory proteins that govern the spatiotemporal expression of genes. These transcription factor binding sites (TFBS), or regulatory motifs, have proven hard to identify, primarily because they are short in length and can show sequence variation. The importance of understanding transcriptional regulation to many fields of biology has led researchers to develop strategies for predicting the presence of TFBS, many of which have led to publicly available databases. One such technique is Phylogenetic Footprinting.

Phylogenetic footprinting relies upon two major concepts:

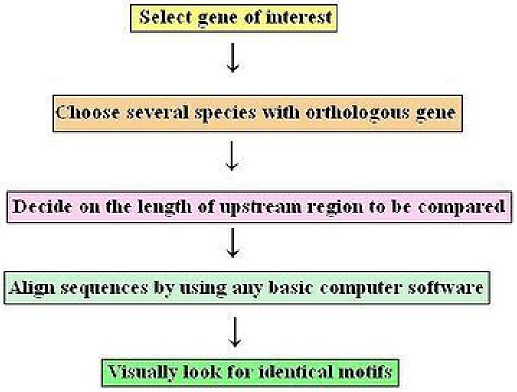
1. The function and DNA binding preferences of transcription factors are well-conserved between diverse species.
2. Important non-coding DNA sequences that are essential for regulating gene expression will show differential selective pressure. A slower rate of change occurs in TFBS than in other, less critical, parts of the non-coding genome.

History

Phylogenetic footprinting was first used and published by Tagle et al. in 1988, which allowed researchers to predict evolutionary conserved cis-regulatory elements responsible for embryonic ε and γ globulin gene expression in primates.

Before phylogenetic footprinting, DNase footprinting was used, where protein would be bound to DNA transcription factor binding sites (TFBS) protecting it from DNase digestion. One of the problems with this technique was the amount of time and labor it would take. Unlike DNase footprinting, phylogenetic footprinting relies on evolutionary constraints within the genome, with the “important” parts of the sequence being conserved among the different species.

Protocol


Select gene of interest
Choose several species with orthologous gene
Decide on the length of upstream region to be compared
Align sequences by using any basic computer software
Visually look for identical motifs

Phylogenetic Footprinting technique protocol

It is important when using this technique to decide which genome your sequence should be aligned to. More divergent species will have less sequence homology between orthologous genes. Therefore, the key is to pick species that are related enough to detect homology, but divergent enough to maximize non-alignment “noise”. Step wise approach to Phylogenetic footprinting consists of:

1. One should decide on the gene of interest.
2. Carefully choose species with orthologous genes.
3. Decide on the length of the upstream or maybe downstream region to be looked at.
4. Align the sequences.
5. Look for conserved regions and analyse them.

Not all TFBS are found

Not all transcription binding sites can be found using phylogenetic footprinting due to the statistical nature this technique. Here are several reasons why some TFBS are not found:

Species specific binding sites

Some binding sites seem to have no significant matches in most other species. Therefore, detecting these sites by phylogenetic footprinting is likely impossible unless a large number of closely related species are available.

Very short binding sites

Some binding sites show excellent conservation, but just in a shorter region than the ones we looked for. Such short motifs (e.g., GC-box) are often happened by chance in nonfunctional sequences and detecting these motifs would be challenging.

Less specific binding factors

Some binding sites show some conservation but have had insertions or deletions. It is not obvious if these sequences with insertions or deletions are still functional. Though they may still be functional if the finding factor is less specific (or less 'picky' if you will). Because deletions and insertions are rare in binding sites, considering insertions and deletions in the sequence would detect a few more true TFBSs, but it could likely include many more false positives.

Not enough data

Some motifs are quite well conserved, but they are statistically insignificant a specific dataset. The motif might have appeared in different species by chance. These motifs could be detected if sequences from more organisms are available. So this will be less of a problem in the future.

Compound binding regions

Some transcription factors bind as dimers. Therefore, their binding sites may consist of two conserved regions, separated by a few variable nucleotides. Because of the variable internal sequence, the motif cannot be detected. However, if we could use a program to search for motifs containing a variable sequence in the middle, without counting mutations, these motifs could be discovered.

Accuracy

It is important to keep in mind that not all conserved sequences are under selection pressure. To eliminate false positives statistical analysis must be performed that will show that the motifs reported have a mutation rate meaningfully less than that of the surrounding nonfunctional sequence.

Moreover, results could be more accurate if the prior knowledge about the sequence is considered. For example, some regulatory elements are repeated 15 times in a promoter region (e.g., some metallothionein promoters have up to 15 metal response elements (MREs)). Thus, to eliminate false motifs with inconsistent order across species, the orientation and order of regulatory elements in a promoter region should be the same in all species. This type of information could help us to identify regulatory elements that are not adequately conserved but occur in several copies in the input sequence.

Chapter- 8

Most Recent Common Ancestor

In genetics, the **most recent common ancestor** (**MRCA**) of any set of organisms is the most recent individual from which all organisms in the group are directly descended. The term is often applied to human genealogy.

The MRCA of a set of individuals can sometimes be determined by referring to an established pedigree. However, in general, it is impossible to identify the specific MRCA of a set of individuals, but an estimate of the time at which the MRCA lived can often be given; such estimates can be given based on DNA test results and established mutation rates, or by reference to a non-genetic genealogical model. This time estimate is referred to as **TMRCA** in scientific papers.

The term *MRCA* is usually used to describe a common ancestor of individuals within a species. It can also be used to describe a common ancestor between species. To avoid confusion, **last common ancestor** (**LCA**) or the equivalent term **concestor** is sometimes used in place of MRCA when discussing ancestry between species.

The term *MRCA* may also be used to identify a common ancestor between a set of organisms via specific gene pathways. Mitochondrial Eve and Y-chromosomal Adam are examples of such MRCAs.

MRCA of all living humans

Tracing one person's lineage back in time forms a binary tree of parents, grandparents, great-grandparents and so on. However, the number of individuals in such an ancestor tree grows exponentially and will eventually become impossibly high. For example, an individual human alive today would, over 30 generations, going back to about the High Middle Ages, have 2^{30} or about 1.07 billion ancestors, more than the total world population at the time.

In reality, an ancestor tree is not a binary tree. Rather, pedigree collapse changes the binary tree to a directed acyclic graph.

Consider the formation, one generation at a time, of the ancestor graph of all currently living humans with no descendants. Start with living people with no descendants at the bottom of the graph. Adding the parents of all those individuals at the top of the graph will connect (half-) siblings via one or two common ancestors, their parent(s). Adding the next generation will connect all first cousins. As each of the following generations of ancestors is added to the top of the graph, the relationship between more and more people is mapped (second cousins, third cousins and so on). Eventually a generation is reached where one or more of the many top-level ancestors is an MRCA from whom it is possible to trace a path of direct descendants all the way down to every living person at the bottom generations of the graph.

The MRCA of everyone alive today could thus have co-existed with a large human population, most of whom either have no living descendants today or else are ancestors of a subset of people alive today. The existence of an MRCA does therefore not imply the existence of a population bottleneck or first couple.

It is incorrect to assume that the MRCA passed all of his or her genes (or indeed any single gene) down to every person alive today. Because of sexual reproduction, at every generation, an ancestor only passes half of his or her genes to each particular descendant in the next generation. Save for inbreeding, the percentage of genes inherited from the MRCA becomes smaller and smaller in individuals at every successive generation, sometimes even decreasing to zero (at which point the Ship of Theseus situation arises), as genes inherited from contemporaries of MRCA are interchanged via sexual reproduction.

Ways to find the MRCA

There are a number of ways to estimate the MRCA such as genetics, archaeology, mathematical models, computer simulations and History. DNA studies have a problem in telling us about the MRCA. As Chang notes, the MRCA will be much more recent than any MRCA that could ever be found in DNA studies, even if one were to study the ancestry of every single gene. The reason being that we are considering people who are simply ancestors, through any route, whether or not any of their genes actually survived the journey. As the human genome consists of roughly 2^{32} base pairs, the genetic contribution of a single ancestor may be flushed out of an individual's genome completely after 32 generations, or roughly 1,000 years.

Time estimates

Depending on the survival of isolated lineages without admixture from modern migrations and taking into account long-isolated peoples, such as historical tribes in central Africa, Australia and remote islands in the South Pacific, the human MRCA was generally assumed to have lived in the Upper Paleolithic period.

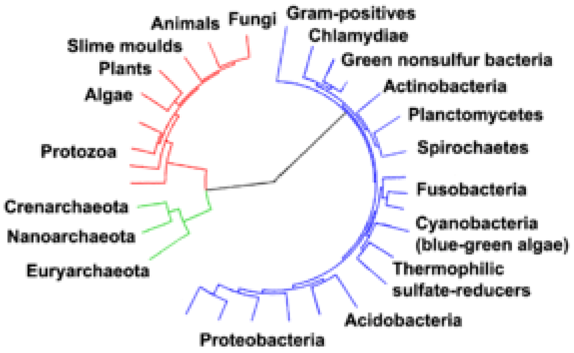
According to Rohde and his colleagues, if we consider not just our all-female and all-male lines, but our ancestors along all parental lines, it turns out that everyone on earth may share a common ancestor who is remarkably recent. Rohde, Olson and Chang (2004), using a non-genetic model, estimated that the MRCA of all living humans may have lived within historical times (3rd millennium BC to 1st millennium AD). The paper suggests, "No matter the languages we speak or the color of our skin, we share ancestors who planted rice on the banks of the Yangtze, who first domesticated horses on the steppes of the Ukraine, who hunted giant sloths in the forests of North and South America and who labored to build the Great Pyramid of Khufu". Rohde (2005) refined the simulation with parameters from estimated historical human migrations and of population densities.

For conservative parameters, he pushes back the date for the MRCA to the 6th millennium BC (p. 20), but still concludes with a "surprisingly recent" estimate of a MRCA living in the second or first millennium BC (p. 27). An explanation of this result is that, while humanity's MRCA was indeed a Paleolithic individual up to early modern times, the European explorers of the 16th and 17th centuries would have fathered enough offspring so that some "mainland" ancestry by today pervades even remote habitats. Besides dating our most recent common ancestor, Rohde's team also calculates that in 5,400 BC everyone alive was either an ancestor of all of humanity, or of nobody alive today. The researchers call this the 'identical ancestors' point: the time before which all the family trees of people today are composed of exactly the same individuals.

Other models reported in Rohde, Olson and Chang (2004) suggest that the MRCA of Western Europeans and people of Western European ancestry lived as recently as AD 1000. The same article provides surprisingly recent estimates for the identical ancestors point, the most recent time when each person then living was either an ancestor of all the persons alive today or an ancestor of none of them. The estimates for this are similarly uncertain, but date to considerably earlier than the MRCA, according to Rohde (2005) roughly to between 15,000 and 5,000 years ago.

A 2008 study found only six women contributed mitochondrial DNA for 95% of all surviving Native Americans and examination of mutation rates has been used to date waves of migration.

MRCA of different species



Evolutionary tree showing the divergence of modern species from the last universal ancestor in the center. The three domains are colored, with bacteria blue, archaea green and eukaryotes red.

It is also possible to use the term MRCA to describe the common ancestor of two or more different species. In the past, the term MRCA was used interchangeably with **last common ancestor** (**LCA**) to denote both the common ancestor within a species and that between species. But MRCA is now more frequently used to describe common ancestors within a species. On the other hand, LCA now describes the common ancestor between two species.

The concept of the last common ancestor is described in Richard Dawkins' book, *The Ancestor's Tale*, in which he imagines a 'pilgrimage' backwards in time, during which we humans travel back through our own evolutionary history and as we do so are joined at each successive stage by all the other species of organism with which we share each respective common ancestor. Dawkins uses the word "**concestor**" (coined by Nicky Warren) as an alternative to LCA.

In *The Ancestor's Tale*, following the human evolutionary tree backwards, we first meet the concestor which we share with the species that are our closest relatives, the chimpanzee and bonobo. Dawkins estimates this to have occurred between 5 and 7 million years ago. Another way of looking at this is to say that our (approximately) 250,000-greats-grandparent was a creature from which all humans, chimpanzees and bonobos are directly descended. Further on in Dawkins' imaginary journey, we meet the concestor we share with the Gorilla, our next nearest relative, then the Orangutan and so on, until we finally meet the concestor of all living organisms, known as the last universal ancestor.

A common mistake is to refer to a proposed *last* common ancestor as an *earliest* ancestor (as in the book *The Link: Uncovering Our Earliest Ancestor* by Colin Tudge and the documentary *Uncovering Our Earliest Ancestor: The Link* screened on the History Channel (US) and BBC One (UK), both referring to the primate fossil dubbed Ida).

MRCA of a population identified by a single genetic marker

It is also possible to consider the ancestry of individual genes (coalescent theory), instead of a person (an organism) as a whole. Unlike organisms, a gene is passed down from a generation of organisms to the next generation either as perfect replicas of itself or as slightly mutated *descendant genes*. While organisms have ancestry graphs and progeny graphs via sexual reproduction, a gene has a single chain of ancestors and a tree of descendants. An organism produced by non-autogamous sexual reproduction has at least two ancestors (immediate parents), but a gene always has one single ancestor.

Given any gene in the body of a person, we can trace a single chain of *human ancestors* back in time, following the lineage of this one gene. Because a typical organism is built from tens of thousands of genes, there are numerous ways to trace the ancestry of organisms using this mechanism. But all these inheritance pathways share one common feature. If we start with all humans alive in 1995 and trace their ancestry by one particular gene (actually a locus), we find that the farther we move back in time, the smaller the number of ancestors becomes. The pool of ancestors continues to shrink until we find the *most recent common ancestor* (MRCA) of all humans who were alive in 1995, *via this particular gene pathway*.

Patrilineal and matrilineal MRCA

It is not possible to trace human ancestry via autosomal chromosomes. Although a chromosome contains genes that are passed down from parents to children via independent assortment from only one of the two parents, genetic recombination (chromosomal crossover) mixes genes from non-sister chromatids from both parents during meiosis, thus changing the nature of the chromosome. In addition, each parent will pass on only one of their autosomal chromosomes to their offspring.

However, the mitochondrial DNA (mtDNA) is nearly immune to sexual mixing. Mitochondrial DNA, therefore, can be used to trace matrilineal inheritance of a population of related individuals. Similarly Y-DNA is present as a single chromosome in the male individual and is passed on to sons and grandsons without recombination.

Time estimates or TMRCA

Both Y-DNA and mtDNA are thus used to trace ancestry. Populations are defined by the accumulation of mutations on the gene and special trees are created for the mutations and the order in which they occurred in a population. The tree is formed through the testing of a large number of individuals all over the world for the presence or lack of a certain set of mutations. Once this is done it is possible to determine how many mutations separate one population from another in the case of mtDNA. The number of mutations in turn allow scientists to determine the approximate time passed, or the TMRCA, since the populations separated. This estimate is based on the estimated mutation rate of the mtDNA in the regions tested. The TMRCA would be the time when both populations still shared the same set of mutations or belonged to the same haplogroup.

In the case of Y-DNA TMRCA is arrived at in a different way. Haplogroups are defined by single-nucleotide polymorphism in various regions of the Y-DNA. The TMRCA within the haplogroup is defined by the accumulation of mutations in STR sequences of the Y-Chromosome of that haplogroup only. Y-DNA network analysis of Y-STR haplotypes yielding a star cluster can be regarded as representing a population descended from a single ancestor. In this case the variability of the DYS sequence, also called the microsatellite variation, can be regarded as a measure of the time passed since the ancestor founded this particular population. Variability due to multiple founding individuals will not display as a star cluster and overall variability in the Y-DNA is thus due in part to variability in the original founding population. The descendants of Genghis Khan or one of his ancestors represents a famous star cluster than can be dated back to the time of Genghis Khan.

Mitochondrial Eve and the most recent common patrilineal ancestor of all living male humans, known as Y-chromosomal Adam, have been established by researchers using tests of the same kinds of DNA as for two individuals. Mitochondrial Eve is estimated to have lived about 140,000 years ago. Y-chromosomal Adam is estimated to have lived around 60,000 years ago. The MRCA of humans alive today would therefore need to have lived more recently than either.

TMRCA calculations are considered critical evidence when attempting to determine migration dates of various populations as they spread around the world. For example, if a mutation is deemed to have occurred 30,000 years ago, then this mutation should be found amongst all populations that diverged after this date. If archeological evidence indicates cultural spread and formation of regionally isolated populations then this must be reflected in the isolation of subsequent genetic mutations in this region. If genetic divergence and regional divergence coincide it can be concluded that the observed divergence is due to migration as evidenced by the archaeological record. However, if the date of genetic divergence occurs at a different time than the archaeological record, then scientists will have to look at alternate archaeological evidence to explain the genetic divergence. The issue is best illustrated in the debate surrounding the demic diffusion versus cultural diffusion during the European Neolithic.

Identical ancestors point

In genetic genealogy, the **identical ancestors point** (**IAP**) is that point in a given population's past where each individual then alive turned out to be either the ancestor of every individual alive now, or to have no living descendants at all. This point lies further in the past than the population's most recent common ancestor (MRCA).

The MRCA had many contemporary companions of both sexes. Many of these contemporaries left direct descendants, but not all of them left an unbroken link of descendants all the way down to today's population. That is, some contemporaries are ancestors of no one in current population. The rest of contemporaries of MRCA may claim ancestry over a subset of current population, but not the entirety of current population.

Because ancestors of MRCA are by definition also common ancestors, we can continue to find (less recent) common ancestors by pushing further back in time to the MRCA's ancestors. Eventually we will reach a point in the past where all humans can be divided into two groups: those who left no descendants today and those who are common ancestors of all living humans today. This point in time is termed the *identical ancestors point*. Even though each living person receives genes in dramatically different proportions from these ancestors from the identical ancestors point, all living people share exactly the same set of ancestors from this point back, all the way to the very first single-celled organism.

The identical ancestors point for *Homo sapiens* has been estimated to between 15,000 and 5,000 years ago, with an estimate of the human MRCA living about 2,000 to 5,000 years ago, that is, estimating the IAP to be about three times as distant as the MRCA. Note that both the matrilineal and the patrilineal human MRCAs are far more remote still, dating to some 150,000 and 90,000 years ago, respectively.

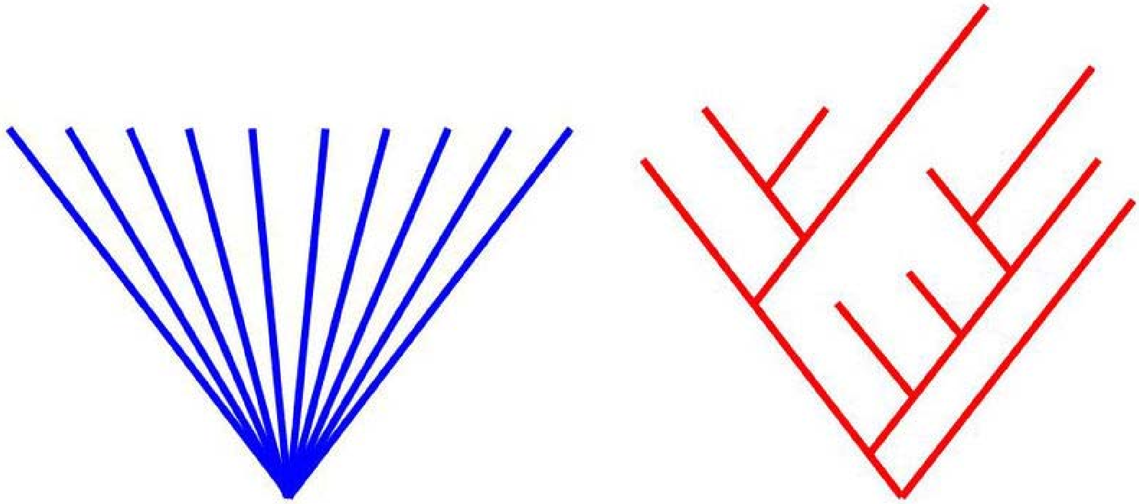
It is incorrect to assume that the MRCA and his/her ancestors passed all their genes down to every person alive today. Because of sexual reproduction, at every generation, an ancestor only passes half of his or her genes to the next generation. The percentage of genes inherited from the MRCA becomes smaller and smaller at every successive generation, as genes inherited from contemporaries of MRCA are interchanged via sexual reproduction. As the human genome consists of roughly 2^{32} base pairs, the genetic contribution of a single ancestor may be flushed out of an individual's genome completely after 32 generations, or roughly 1,000 years.

The MRCA had many contemporary companions of both sexes. Many of these contemporaries left direct descendants, but not all of them left an unbroken link of descendants all the way down to today's population. That is, some contemporaries of the MRCA are ancestors of no one in the current population. The rest of the contemporaries of the MRCA may claim ancestry over a subset of current population, but not the entirety of current population.

Because ancestors of MRCA are by definition also common ancestors, we can continue to find (less recent) common ancestors by pushing further back in time to the MRCA's ancestors. Eventually we reach a point in the past where all humans can be divided into two groups: those who left no descendants today and those who are common ancestors of all living humans today. This point in time is termed the *identical ancestors point*. Even though each living person receives genes (in original or mutated forms) in dramatically different proportions from these ancestors from the identical ancestors point, from this point back, all living people share exactly the same set of ancestors, all the way to the very first single-celled organism.

Chapter- 9

Phylogenetic Comparative Methods



When applied to comparative data, conventional statistical methods assume, in effect, that all species are completely unrelated, as if they descended from a "big bang" of special creation. Such a scenario can be depicted as a "star" phylogeny (left). Most comparative studies will involve species that have descended in a hierarchical fashion from common ancestors, as shown on the right. Comparative data sets may include species that have gone extinct and rates of evolution may vary among branches. Therefore, trees used in phylogenetic comparative methods may not be "ultrametric" (have tips that are contemporaneous).

Phylogenetic comparative methods (PCMs) use information on the evolutionary relationships of organisms (phylogenetic trees) to compare species (Harvey and Pagel, 1991). The most common applications are to test for correlated evolutionary changes in two or more traits, or to determine whether a trait contains a phylogenetic signal (the tendency for related species to resemble each other). However, several methods are available to relate particular phenotypic traits to variation in rates of speciation and/or extinction, including attempts to identify evolutionary key innovations. Although most studies that employ PCMs focus on extant organisms, the methods can also be applied to extinct taxa and can incorporate information from the fossil record.

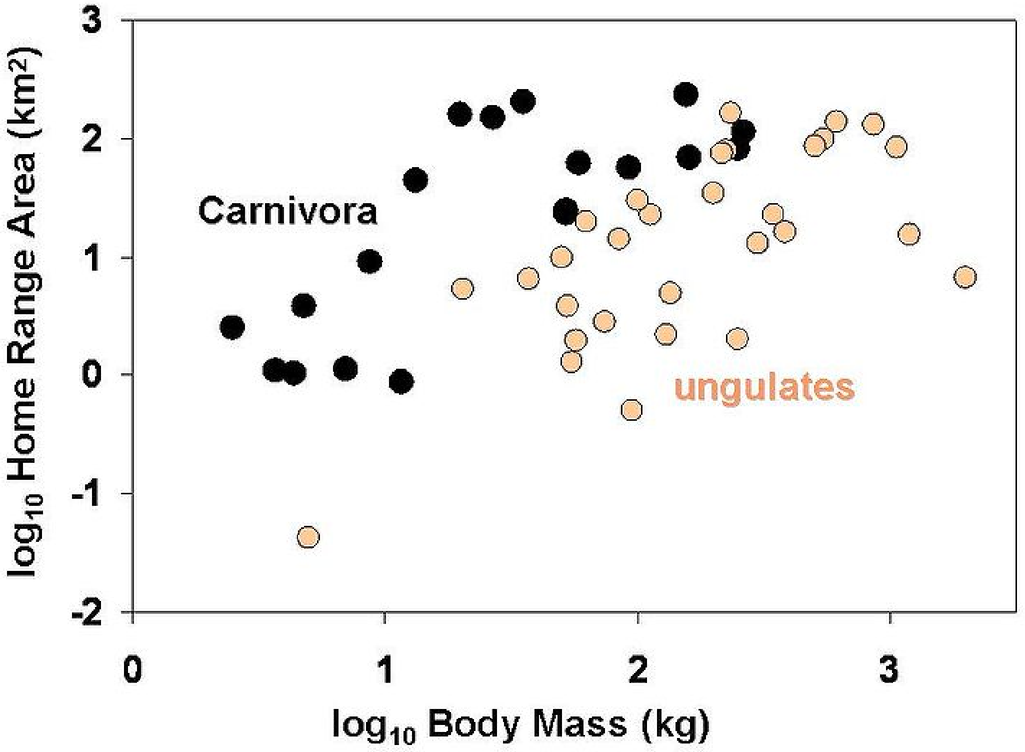
Owing to their computational requirements, they are usually implemented by computer programs. PCMs can be viewed as part of evolutionary biology, systematics, phylogenetics, bioinformatics or even statistics, as most methods involve statistical procedures and principles for estimation of various parameters and drawing inferences about evolutionary processes.

What distinguishes PCMs from most traditional approaches in systematics and phylogenetics is that they typically do not attempt to infer the phylogenetic relationships of the species under study. Rather, they use an independent estimate of the phylogenetic tree (topology plus branch lengths) that is derived from a separate phylogenetic analysis, such as comparative DNA sequences that have been analyzed by maximum parsimony or maximum likelihood methods. PCMs are consumers of phylogenetic trees, not primary producers of them. Accordingly, the list of phylogenetics software shows little overlap with the programs for PCMs (see below).

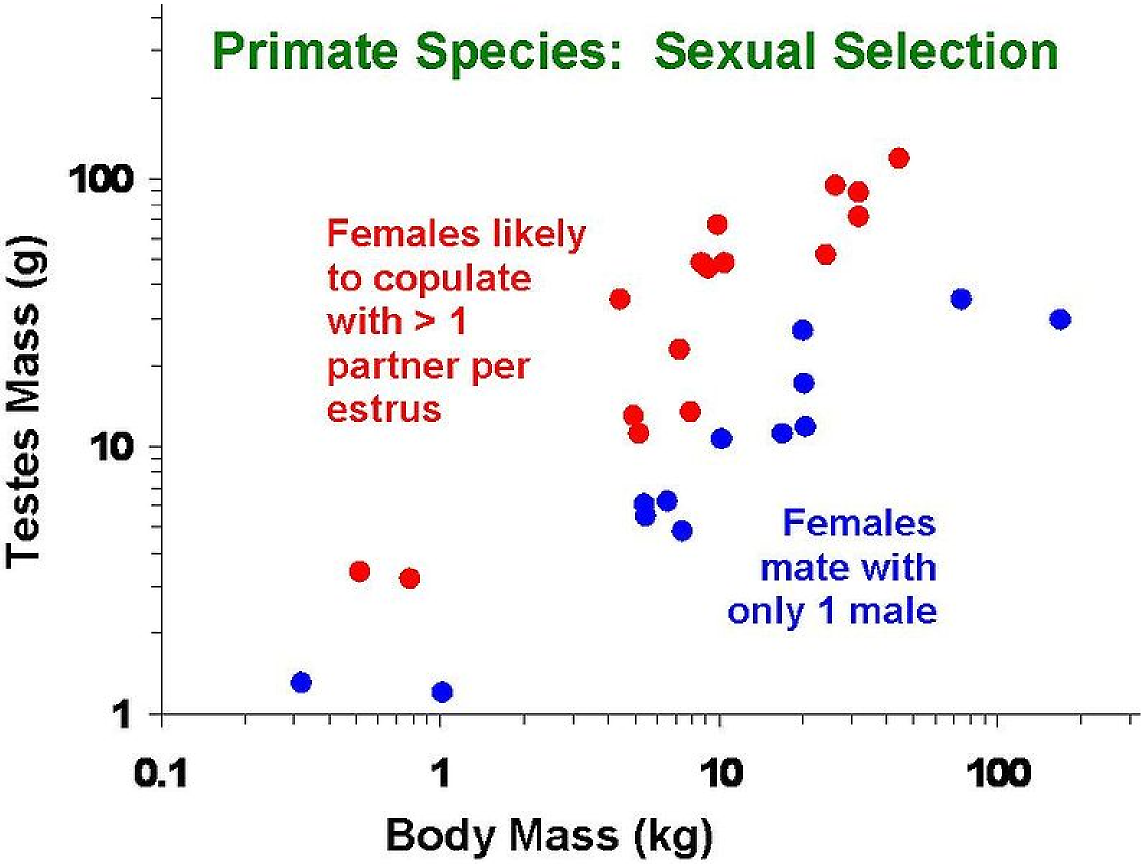
Comparison of species to elucidate aspects of biology has a long history. Charles Darwin relied on such comparisons as a major source of evidence when writing The Origin of Species. Many other fields of biology use interspecific comparison as well, including behavioral ecology, ethology, ecophysiology, comparative physiology, evolutionary physiology, functional morphology, comparative biomechanics and the study of sexual selection.

Applications

PCMs can be used to analyze the origin and maintenance of biodiversity. Biodiversity is most commonly discussed in terms of the number of species, but it can also be phrased in terms of the amount of phenotypic (e.g., physiological, morphological) space that a given set of species occupies.



Home range areas of 49 species of mammals in relation to their body size. Larger-bodied species tend to have larger home ranges, but at any given body size members of the order Carnivora (carnivores and omnivores) tend to have larger horme ranges than ungulates (all of which are herbivores). Whether this difference is considered statistically significant depends on what type of analysis is applied (Garland et al., 1993).



Testes mass of various species of Primates in relation to their body size and mating system. Larger-bodied species tend to have larger testes, but at any given body size species in which females tend to mate with multiple males have males with larger testes.

Phylogenetic comparative methods are commonly applied to such questions as:

- What is the slope of an allometric scaling relationship?

→ *Example: how does brain mass vary in relation to body mass?*

- Do different clades of organisms differ with respect to some phenotypic trait?

→ *Example: do canids have larger hearts than felids?*

- Do groups of species that share a behavioral or ecological feature (e.g., social system, diet) differ in average phenotype?

→ *Example: do carnivores have larger home ranges than herbivores?*

- What was the ancestral state of a trait?

→ *Example: where did endothermy evolve in the lineage that led to mammals?*

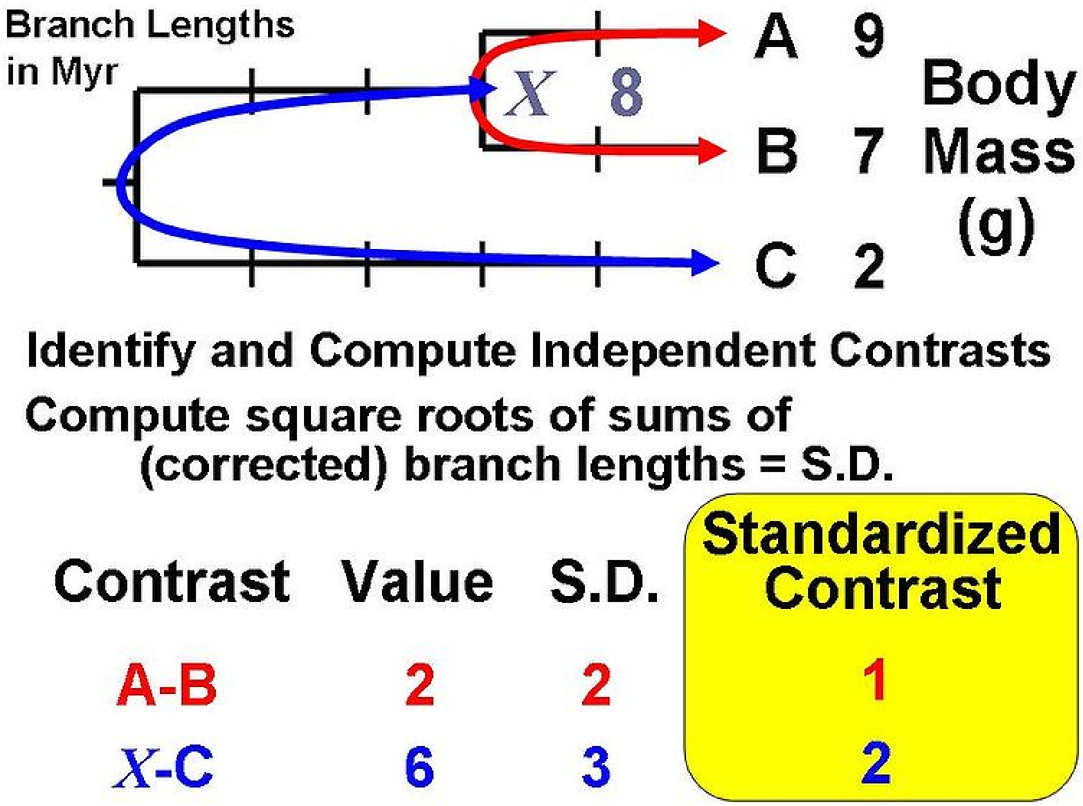
- Does a trait exhibit significant phylogenetic signal in a particular group of organisms? Do certain types of traits tend to "follow phylogeny" more than others?

→ *Example: are behavioral traits more labile during evolution?*

- Do species differences in life history traits trade-off, as in the so-called fast-slow continuum?

→ *Example: why do small-bodied species have shorter life spans than their larger relatives?*

Phylogenetically independent contrasts



Identify and Compute Independent Contrasts

Compute square roots of sums of (corrected) branch lengths = S.D.

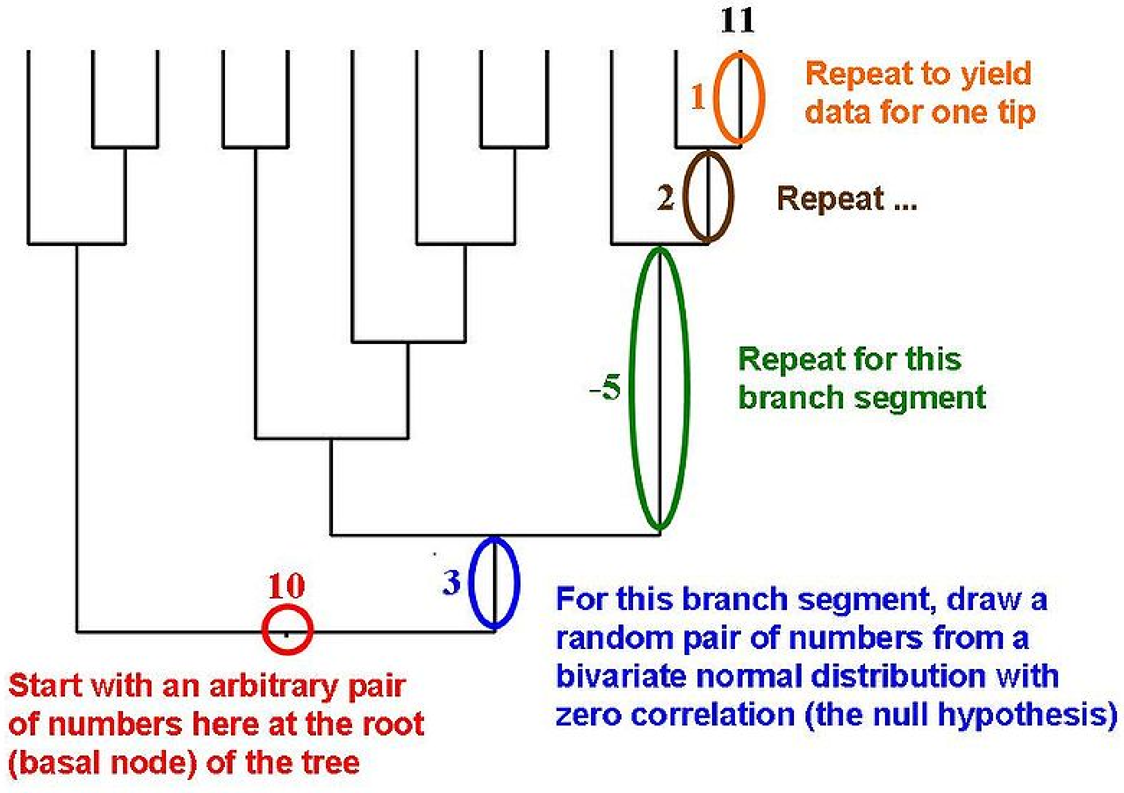
Contrast	Value	S.D.	Standardized Contrast
A-B	2	2	1
X-C	6	3	2

The standardized contrasts are used in conventional statistical procedures, with the constraint that all regressions, correlations, analysis of covariance, etc., must pass through the origin.

Felsenstein (1985) proposed the first general statistical method for incorporating phylogenetic information, i.e., the first that could use any arbitrary topology (branching order) and a specified set of branch lengths. The method is now recognized as an algorithm that implements a special case of what are termed phylogenetic generalized least-squares models (Grafen, 1989). The logic of the method is to use phylogenetic information (and an assumed Brownian motion like model of trait evolution) to transform the original tip data (mean values for a set of species) into values that are statistically independent and identically distributed.

The algorithm involves computing values at internal nodes as an intermediate step, but they are generally not used for inferences by themselves. An exception occurs for the basal (root) node, which can be interpreted as an estimate of the ancestral value for the entire tree (assuming that no directional evolutionary trends [e.g., Cope's rule] have occurred) or as a phylogenetically weighted estimate of the mean for the entire set of tip species (terminal taxa). The value at the root is equivalent to that obtained from the "squared-change parsimony" algorithm and is also the maximum likelihood estimate under Brownian motion. The independent contrasts algebra can also be used to compute a standard error or confidence interval.

Phylogenetically informed Monte Carlo computer simulations


11
1
Repeat to yield data for one tip
2
Repeat ...
-5
Repeat for this branch segment
10
3
For this branch segment, draw a random pair of numbers from a bivariate normal distribution with zero correlation (the null hypothesis)
Start with an arbitrary pair of numbers here at the root (basal node) of the tree

Data for a continuous-valued trait can be simulated in such a way that taxa at the tips of a hypothetical phylogenetic tree will exhibit phylogenetic signal, i.e., closely related species will tend to resemble each other.

Martins and Garland (1991) proposed that one way to account for phylogenetic relations when conducting statistical analyses was to use computer simulations to create many data sets that are consistent with the null hypothesis under test (e.g., no correlation between two traits, no difference between two ecologically defined groups of species) but that mimic evolution along the relevant phylogenetic tree. If such data sets (typically 1,000 or more) are analyzed with the same statistical procedure that is used to analyze a real data set, then results for the simulated data sets can be used to create phylogenetically correct (or "PC" [Garland et al., 1993]) null distributions of the test statistic (e.g., a correlation coefficient, t, F). Such simulation approaches can also be combined with methods like phylogenetically independent contrasts (see above).

Related quotations

"Ought we, for instance, to begin by discussing each separate species - man, lion, ox and the like - taking each kind in hand independently of the rest, or ought we rather to deal first with the attributes which they have in common in virtue of some common element of their nature and proceed from this as a basis for the consideration of the separately?"

"In parallel with laboratory experimental methods, the comparative method increases in value with its sample size, i.e., the number of species being compared. When species are few and their phylogenetic relationships are clouded in the distant past of unfossilized ancestors, the comparative method reaps fewer conclusions of trust."

"... there is no easy way, except by comparison, to test most questions about the long-term history of life, or to generate predictions from evolutionary considerations."

"... we must learn to treat comparative data with the same respect as we would treat experimental results ..."

"In the past, however, cooperation between systematists and other comparative biologists has been sporadic at best. Most experimental biologists have ignored taxonomy and systematics, some even to the extent of not bothering to provide their animals with proper identifications or scientific names."

"If we assume that the ... cladogram .. is correct, we can then hypothesize what the particular common ancestor must have been like."

"... biology will never secure fossils of all species in the past because fossilization is such a rare process, requiring just the right physical and chemical conditions. Therefore, in order to trace probable phylogenetic lineages one must reason from the evidence at hand: the characteristics of contemporary animals themselves which are the end-points of phylogeny (evolutionary history)."

"... the comparative method of 1950 was indistinguishable from the comparative method of 350 BC ..."

"Focusing only on highly adapted species may, of course, bring valuable information on extreme situations but might also obscure the basic mechanisms."

"Most of what we know is based upon comparison. When asked to describe a food not previously tasted or a new kind of music, one often responds that the taste is "like" some other food, or that the sensation "differs" in a particular way from something that is familiar. Indeed, comparison and the similarities and differences it discloses is ingrained in our approach to description of objects, events and processes. Hence the questions "what can we compare?" and its ancillary "how shall we compare?" prove to be the key to any study of natural phenomena."

"Some reviewers of this paper felt that the message was "rather nihilistic," and suggested that it would be much improved if I could present a simple and robust method that obviated the need to have an accurate knowledge of phylogeny. I entirely sympathize, but do not have a method that solves the problem. The best we can do is perhaps to use pairs of close relatives as suggested above, although this discards at least half of the data. Comparative biologists may understandably feel frustrated upon being told that they need to know the phylogenies of their groups in great detail, when this is not something they had much interest in knowing. Nevertheless, efforts to cope with the effects of the phylogeny will have to be made. Phylogenies are fundamental to comparative biology; there is no doing it without taking them into account."

"Comparative biologists tend to suspect comparisons of distantly related species; they hope to base their comparisons on recent evolutionary events that have not been overlaid by much subsequent change."

"Yet, one of the most embarrassing things that could be done to a group of respected biologists would be to ask them to spend a few minutes to write down what is meant by the comparative method and what are the basic goals and principles of biological comparison."

"As welcome as it is to see the lung successfully employed as a systematic index, it is, on the other hand, unfortunate. Thereby, lungs lose their innocence in the sense that phylogenetic trees and cladograms can no longer be used to help resolve the sequence in the development of lung structure without the danger of circular argumentation."

"To be maximally informative, such studies should be undertaken on closely related groups of organisms, so that factors extraneous to the comparison can be minimized ..."

"The comparative approach is not new. Indeed, it was Darwin's favoured technique. ... In short, comparative studies have taught us most of what we know about adaptation."

"In short, all useful comparative methods are based on explicit models of evolutionary change."

"Adaptation is an inherently comparative idea ..."

"The usual symptom of non-independence is that closely related species tend to be more alike than more distantly related species."

"... to use species as independent data points in a comparative analysis requires that one ignores phylogenetic relationships."

"Because life-history traits are likely to be correlated with a species' phylogenetic history, unequivocal evidence for adaptation to local environmental conditions may be recognized only after the variation in a trait attributable to phylogeny is removed."

"Broad-scale comparative evidence from across a large number of taxa may often help set limits to the applicability of hypotheses that have been generated from a particular phenomenon in a particular species."

"Any comparative study lacking a phylogenetic perspective now would be incomplete."

"However, in general, the evolutionary process involves descent with modification and in the absence of modification, one must conclude that similarities among closely related taxa reflect shared ancestry. Phylogeny, then, is an important explanatory principle for understanding shared characteristics and should be the null hypothesis in all tests of similarity or differentiation among taxa."

"Moreover, comparative studies supply only correlational data. Correlation does not demonstrate causation ..."

"In addition to his theory of natural selection, the comparative method is what made Darwin great. If you don't believe this claim, look at any of his major works. They are packed full with interspecific comparisons based on detailed studies and anecdotal observations."

"Population comparisons can provide particularly powerful means of evaluating adaptive hypotheses for two reasons. The first is that there tend to be fewer differences between populations than between species. Consequently, there are fewer covarying traits to confound analyses ... Second, divergent populations are often relatively young and may be more likely to reside in the habitats in which their derived character states evolved than is the case for divergent species with potentially longer intervening histories ..."

"Naive, prephylogenetic comparative tests should be kept at the other end of a barge pole."

"It is the study of the bizarre, the outliers, the freaks, that gives us some of our clearest insights into the hows and whys of evolution."