When substitutions are common, a particular site may have undergone multiple changes. Thus, alignments between sequences with many differences will underestimate the true number of substitutions that has occurred. The true number of substitutions can be estimated by

\[ K = -\frac{3}{4} \ln \left[ 1 - \left(\frac{4}{3}\right)(p) \right], \]

where \( p \) is the fraction of nucleotides that differ between the two sequences.
Constraints on Mutations

Transitions vs. Transversions
Transitions (exchanging one purine (A or G) for another purine (G or A), or one pyrimidine (C,T or U) for another pyrimidine (U,T or C)) are three times as common as transversions (purine for pyrimidine or vice versa).

Functional Constraints
Functional constraints in coding or regulatory regions also impact the rate of change.

Divergence Among Human, Mouse, Rabbit and Cow Globin Genes by Region

<table>
<thead>
<tr>
<th>Region</th>
<th>Substitutions/site/$10^9$ years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noncoding</td>
<td>3.33</td>
</tr>
<tr>
<td>Coding</td>
<td>1.58</td>
</tr>
<tr>
<td>5' untranslated</td>
<td>1.86</td>
</tr>
<tr>
<td>3' untranslated</td>
<td>3.00</td>
</tr>
</tbody>
</table>
Synonymous vs. Nonsynonymous Substitutions

**Nondegenerate Sites**
Codon positions where any nucleotide mutation would cause a change in the amino acid (a nonsynonymous substitution).
**Example:** Phenylalanine (UUU)

**Twofold Degenerate Sites**
Codon positions where one nucleotide mutation would not cause a change in the amino acid (but the two other possible mutations would).
**Example:** Aspartic acid third codon position (GAU, GAC)

**Fourfold Degenerate Sites**
Codon positions where any nucleotide mutation would not cause a change in the amino acid (a synonymous substitution).
**Example:** Glycine third codon position (GGG, GGA, GGU, GGC).

**Divergence Among Human and Rabbit Globin Genes by Sites**
- Nondegenerate: 0.56 substitutions/site/10^9 years
- Twofold Degenerate: 1.67 substitutions/site/10^9 years
- Fourfold Degenerate: 2.35 substitutions/site/10^9 years

**vs.**
- Noncoding: 3.33 substitutions/site/10^9 years
- Coding: 1.58 substitutions/site/10^9 years
Variations in Evolutionary Rates

Variations in Substitution Rates
Substitution rates do not appear to be constant within even genomes of closely related species, and also vary from species to species.

Relative Rate Test
It is possible to estimate the overall rate of substitution in different lineages without knowing the exact divergence time by determining a relative rate of substitution. To compare the relative rate of substitution in species A and species B, one designates a less related species, C, as an outgroup. This allows you to estimate the amount of divergence that has taken place in species A and B since they last shared a common ancestor.

Relative Rates of Synonymous Substitutions in Genes

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammalian Fibrinopeptide</td>
<td>4</td>
</tr>
<tr>
<td>Mammalian Hemoglobin</td>
<td>1</td>
</tr>
<tr>
<td>Mammalian Cytochrome c</td>
<td>0.2</td>
</tr>
<tr>
<td>Influenza NS</td>
<td>1,000,000</td>
</tr>
</tbody>
</table>

Relative Rates of Synonymous Substitution in Genomes

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human genome</td>
<td>1</td>
</tr>
<tr>
<td>Plant genome</td>
<td>1</td>
</tr>
<tr>
<td>Mouse genome</td>
<td>2</td>
</tr>
<tr>
<td>Rat genome</td>
<td>2</td>
</tr>
<tr>
<td>Human mitochondria</td>
<td>10</td>
</tr>
<tr>
<td>Plant chloroplast</td>
<td>0.3</td>
</tr>
</tbody>
</table>
A Brief History of Phylogeny

1735  Taxonomy         Karl von Linné
1750  Phenetic Taxonomy  Michel Adanson
1859  Evolution          Charles Darwin
1866  Phylogeny          Ernst Haeckel
1950  Cladistic Taxonomy  Willi Hennig

“[in time] we shall have very fairly true genealogical trees of each great kingdom of nature”
– Charles Darwin, 1857, letter to T. H. Huxley

“The History of the Germ in an epitome of the History of the Descent, or, in other words: that Ontogeny is a recapitulation of Phylogeny”
– Ernst Haeckel, 1897, The Evolution of Man

“The universal phylogenetic tree not only spans all extant life, but its root and earliest branchings represent stages in the evolutionary process before modern cell types had come into being. The evolution of the cell is an interplay between vertically derived and horizontally acquired variation. Primitive cellular entities were necessarily simpler and more modular in design than are modern cells. Consequently, horizontal gene transfer early on was pervasive, dominating the evolutionary dynamic. The root of the universal phylogenetic tree represents the first stage in cellular evolution when the evolving cell became sufficiently integrated and stable to the erosive effects of horizontal gene transfer that true organismal lineages could exist.”
– Carl Woese, 2000, Interpreting the Universal Phylogenetic Tree
Taxonomy is the classification of organisms into an ordered system that indicates natural relationships.

Karl von Linné, a.k.a. Caroli Linnaei or Linnaeus (1707-1778), invented modern taxonomy by developing a hierarchy of taxa (kingdom, class, order, genus, and species) and a system of binomial nomenclature based on genus and a characteristic feature of a species.
**Phylogeny** is the sequence of events involved in the evolutionary development of a species or taxonomic group.

**Ernest Haeckel (1834-1919)**, was originally trained as a physician, but devoted himself to the study of evolution after reading Darwin's *Origin of Species*. He was made famous by his own phrase “ontogeny recapitulates phylogeny.” He coined the term “phylogeny” and created the first phylogenetic trees.
Phenetic vs. Cladistic Approaches

Phylogenetic Reconstruction
Phylogenetic reconstruction attempts to estimate the phylogeny for some data. Any collection of sequences will share some ancestral relationship, and the data within the sequences contains information that can be used to reconstruct or infer these ancestral relationships. A phylogenetic tree is a branching structure which illustrates the relationships between the sequences.

Nearly any approach currently used in phylogenetic reconstruction has adherents and detractors. At the moment, there is a distinct lack of consensus as to best practices and principles for phylogenetic analysis.

Phenetic Approach
Phenetic taxonomy was invented in 1750 by Michel Adanson.

In the phenetic approach, a tree is constructed by considering the phenotypic similarities of the species without trying to understand the evolutionary pathways of the species, and thus may or not be the correct phylogeny. Trees constructed by this method are called phenograms or dendrograms.

Cladistic Approach
Cladistic taxonomy was invented by the German entomologist Willi Hennig in 1950. It involves the rigorous application of the concept of evolution to taxonomy. Taxa are defined by what distinctive features their members have, not what features they share with others.

In the cladistic approach, a phylogenetic tree is reconstructed by considering the various possible pathways of evolution and choosing from amongst these the best possible tree, that is, the tree that involves the fewest changes, and thus the least amount of convergent evolution. Trees reconstructed by this method are called cladograms.
**Phylogenetic Trees**

**Rooted Trees**
In a rooted tree, a single node is designated as a common ancestor, and a unique path leads from it through evolutionary time to all other nodes. It thus provides information about the common ancestry of sequences and the direction of evolution, and is the most common type of tree used to study evolutionary relationships.

**Unrooted Trees**
Unrooted trees specify only the relationship between nodes, and nothing about the direction in which evolution occurred. A root can be assigned to an unrooted tree through the use of an outgroup, for example a species that unambiguously previously separated from the other species being compared (e.g., baboon, when comparing humans and gorillas).

**Source:** Krane & Raymer, Fundamental Concepts of Bioinformatics, NCBI
Tree Topology

**Operational Taxonomic Unit (OTU)**
This corresponds to the terminal nodes of a phylogenetic tree (also known as leaves, tips or external nodes). They represent the genes, organisms, families, species or populations, as appropriate, for which you have data.

**Internal Node**
This corresponds to points within a phylogenetic tree where interior branches meet (also known as vertices). These represent inferred ancestors.

**Outgroup**
An OTU or taxa included for the purpose of rooting a tree.
Rooted Tree Reconstruction

The possible number of unrooted trees is one step less (i.e. 5 species or OTUs → 15 trees, still an enormous number with many species or OTUs). The number of possible trees for n OTUs can be estimated by $(2n-3)!/(2n-2(n-2)!)$ for bifurcating rooted trees and $(2n-5)!/(2n-3(n-3)!)$ for bifurcating unrooted trees.

Source: Brian Golding, Reconstructing Phylogenies
Phylogenetic Terminology

**Orthologs**
Homologous genes in different species with analogous functions. They tend to have similar structure and function.

**Paralogs**
Similar genes that are the result of a gene duplication. They tend to have different but related functions.

**Xenologs**
Genes acquired by horizontal transfer between species, typically mediated by a plasmid or virus.

**Symplesiomorphy**
Having characters are both derived from a common ancestor and shared. This is essential to clearly establishing a phylogeny. Having only derived (non-ancestral) or shared characters is not sufficient to establish a phylogeny.

**Homoplasies**
Convergences of a particular character at a particular site. These typically pose the most difficulty in attempting to reconstruct the ancestral phylogenetic tree.
**Monophyletic**
A group descended from a single common ancestor that contains only and all descendants from that ancestor.

**Paraphyletic**
A group descended from a single common ancestor that does not contain all the descendants from that ancestor.

**Polyphyletic**
A group whose members are not descended from a single common ancestor.

**Gene Tree**
A phylogenetic tree based on divergence observed within a single homologous gene in different species. It may accurately represent the evolutionary history of that gene, but not necessarily of the species. Species trees are best based on the comparison of numerous genes.

**Bootstrapping**
A method for checking the robustness of a given phylogenetic tree by checking whether every portion of the alignment equally supports the structure of the tree.

**Newick Tree Format**
A common text file format for representing phylogenetic trees in a set of nested parenthesis, i.e. i.e. (B, (A, C, E), D); or, with branch lengths included, (B: 6.0, (A: 5.0, C: 3.0, E: 4.0): 5.0, D: 11.0);
Distance Matrix Methods

**Distance Method**
Distance based methods attempt to construct trees based on measures of distance between OTUs (i.e. genes or species). In contrast, character based methods evaluate particular features (i.e. DNA sequence, aa sequence, # of legs, etc.).

**Unweighted-Pair-Group Method with Arithmetic Mean (UPGMA)**
A clustering algorithm which constructs a distance matrix, then clusters together the least distant pair of Operational Taxonomic Units (OTUs), followed by successively more distant OTUs. At each step of the algorithm, the number of OTUs declines by one, replaced by a joint OTU, from which subsequent distances from other OTUs are calculated, until the algorithm finishes by clustering the last pair of OTUs.

This method assumes that the rate of evolutionary change between all branches of the tree is the same, which is generally not a valid assumption. In nature, examples of rates of evolution varying between taxa are common. As a result, corrections to this assumption are often used with this approach.

**Neighbor Joining Method**
This attempts to correct for the assumption made by UPGMA that the same rate of of evolutionary change applies to all branches of the tree. It is otherwise similar to UPGMA, but generally gives better results. It yields an unrooted tree.

**Fitch and Margoliash**
This method attempts to find an optimal tree of minimal distance. It yields an unrooted tree.
Maximum Parsimony
The maximum parsimony method involves evaluating as many trees as possible, giving each a score that is used to choose between different trees. The highest scoring, or most parsimonious tree is the one with the minimum number of evolutionary changes. A number of different methods can be used to calculate scoring.

Fitch Parsimony
For a particular tree, traverse from the leaves toward the root of the tree. At each internal node, determine the set of possible states (i.e. nucleotides). Then, traverse the tree from the root towards the leaves, picking ancestral states for each internal node to minimize the number of changes required.

The Fitch algorithm assumes position independence, and that any state is equally likely to change to any other state. Variations which weight the costs of changes differently exist.

Dollo Parsimony
Assumes that derived states are irreversible, that is, a derived character state cannot be lost and then regained. Hence, the state can evolve and be lost many times throughout evolution, but cannot be inferred to have evolved twice. The tree with maximum parsimony is the one in which derived characters have been lost the fewest number of times.

This method has been used with restriction fragment length polymorphism (RFLP) data, since restriction sites are difficult to gain, but easy to lose. It may be more useful when dealing with non-sequence data, for example, complex phenotypes, which are unlikely to have evolved more than once.

Source: Brian Golding, Reconstructing Phylogenies
The Principle of Parsimony

Occam’s Razor

“Pluralitas non est ponenda sine necessitas” (Do not increase the number of entities required to explain anything beyond what is strictly necessary)
– William of Occam (or Ockham) (1284-1347)

Requires less changes than its neighbor

These two trees are equally parsimonious
Other Methods

Maximum Likelihood
The method of maximum likelihood attempts to reconstruct a phylogeny using an explicit model of evolution. It specifies values for the likelihood of a given trait evolving within a lineage, and chooses the most likely tree, given these values. It attempts to predict the most likely interior nodes given the OTUs, then the most likely tree.

Theoretically, this may be the most powerful method available. For a given model of evolution, no other method will perform as well nor provide you with as much information about the tree. Unfortunately, this is computationally difficult to do and hence, the model of evolution must be a simple one. Even with simple models of evolutionary change the computational task is enormous and this is the slowest of all methods.

Compatibility
Compatibility methods recode data involving multi-state characters to include knowledge of the ancestral states of characters, and from this determine what changes are compatible.

Compatibility methods are more accurate when there are slow rates of evolutionary change. Both compatibility and parsimony assume that homoplasies will be rare.

Source: Brian Golding, Reconstructing Phylogenies
Rules of Thumb for Phylogeny

• Use more than one method. Each one will provide a phylogenetic history biased by that method’s assumptions.

• Bootstrap or jackknife your data to test the quality of your tree. When bootstrapping, use at least several hundred iterations of resampling and tree generation.

• Run your analysis with different subsets of taxa to see if the trees thus generated are congruent. Dropping a single OTU should not dramatically change your tree.

• Treat long branches with caution. They tend to attract each other.

• Beware of non-orthologous genes, horizontal gene transfers, or recombinant sequences. Standard phylogenetic methods do not handle them well.

• When using outgroups, consider including more than one outgroup taxa, and choose outgroup species that are evenly spaced on the tree.

• Including intermediate taxa can help resolve even the relationship of a few taxa.

• When the number of substitutions per site is unusually high or low, distance methods may perform better than parsimony methods.

• If you expect homoplasies to be scattered at random throughout the sequence data, then a parsimony method will perform best. If homoplasies are expected to be concentrated in a few characters, whose identities are known in advance, then compatibility will perform better than parsimony.
Phylogenetic Software Packages

**PHYLIP**
The leading free package for phylogeny. It includes programs to carry out parsimony, distance matrix methods, maximum likelihood, and other methods on a variety of types of data, including DNA and RNA sequences, protein sequences, restriction sites, 0/1 discrete characters data, gene frequencies, continuous characters and distance matrices. Although it is free, it can be complex to use.


**Phylodendron**
A free web-based tree drawing program with a simple user interface and many output options.

http://iubio.bio.indiana.edu/treeapp/treeprint-form.html

**ClustalX**
A free multiple sequence alignment program that includes the ability to create phylogenetic trees based on the Neighbor Joining Method.

http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html

**PAUP**
The leading commercial package for phylogeny. It includes parsimony, distance matrix, invariants, and maximum likelihood methods and many indices and statistical tests.

http://paup.csit.fsu.edu/
http://www.sinauer.com/

**MacClade**
A commercial package for interactive analysis of evolution of a variety of character types, including discrete characters and molecular sequence. It works well with PAUP.

http://www.sinauer.com/