Molecular Phylogenetics
Core course: ZOOL3014
Semester: VI

Prof. Pranveer Singh

Molecular Phylogenetics

Combination of molecular and statistical techniques to infer evolutionary relationships among organisms or genes

Common computational methods used to infer phylogenetic information from molecular data

Use the structure and function of molecules and how they change over time to infer these evolutionary relationship

Protein sequencing, PCR, electrophoresis, and other molecular biology techniques coupled with computer algorithms and complicated stochastic and probabilistic problems aid defining evolution at the molecular level more effectively

Phylogenetic data sets can consist of hundreds of different species, each of which may have varying mutation rates and patterns that influence evolutionary change

The primary objective of molecular phylogenetic studies is to recover the order of evolutionary events and represent them in evolutionary trees that graphically depict relationships among species or genes over time

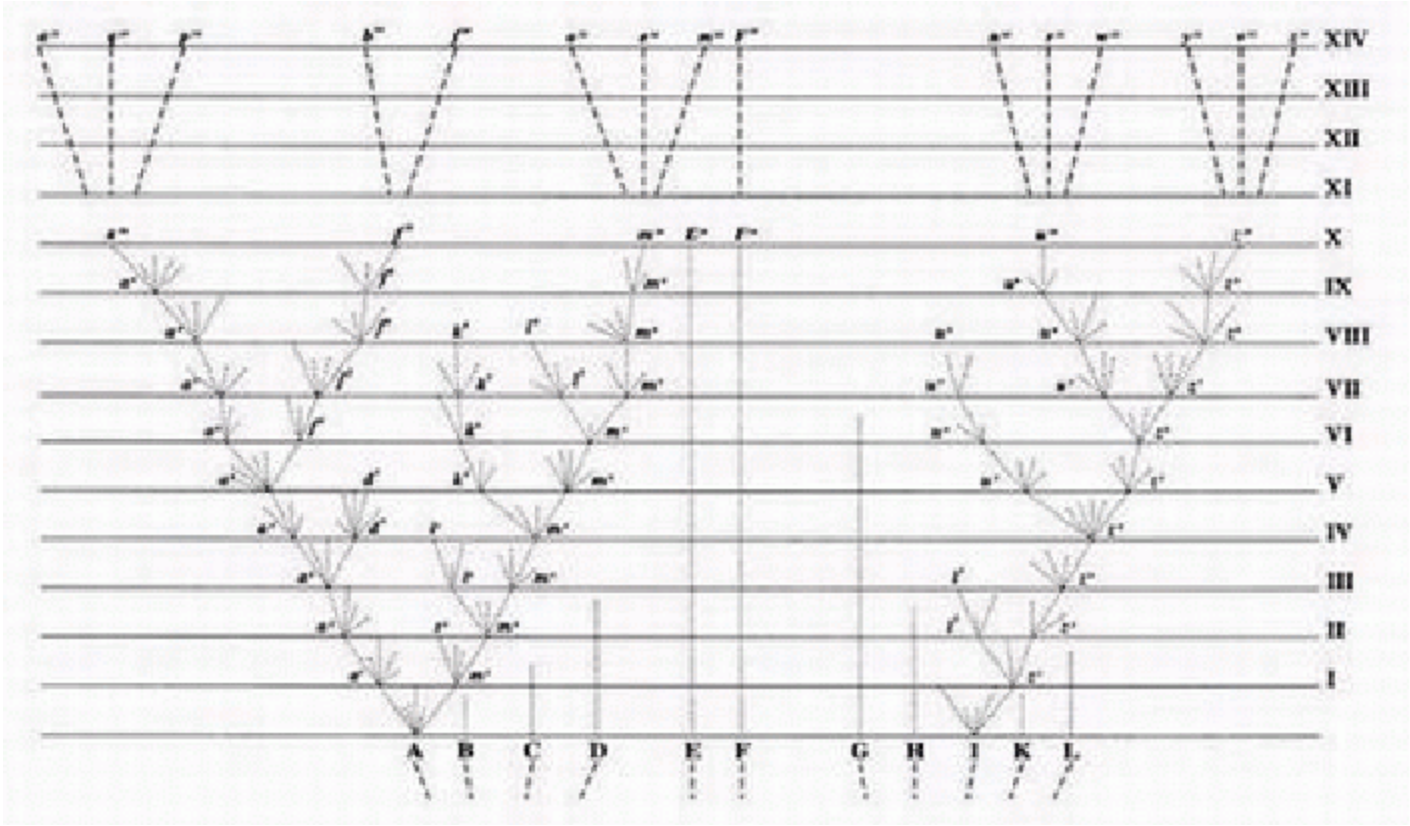
Consequently, there are numerous different evolutionary models and stochastic methods available

Post-Darwin it has been determined that evolution is actually a molecular process based on genetic information, encoded in DNA, RNA, and proteins

One molecule undergoes *diversification* into many variations. One or more of those variants can be *selected* to be reproduced or *amplified* throughout a population over many generations

Such variations at the molecular level can be caused by mutations, such as deletions, insertions, inversions, or substitutions at the nucleotide level, which in turn affect protein structure and biological function

Evolution Defined Graphically



The sole illustration in Darwin's *Origin of the Species* uses a tree-like structure to describe evolution. This drawing shows ancestors at the limbs and branches of the tree, more recent ancestors at its twigs, and contemporary organisms at its buds

Phylogeny

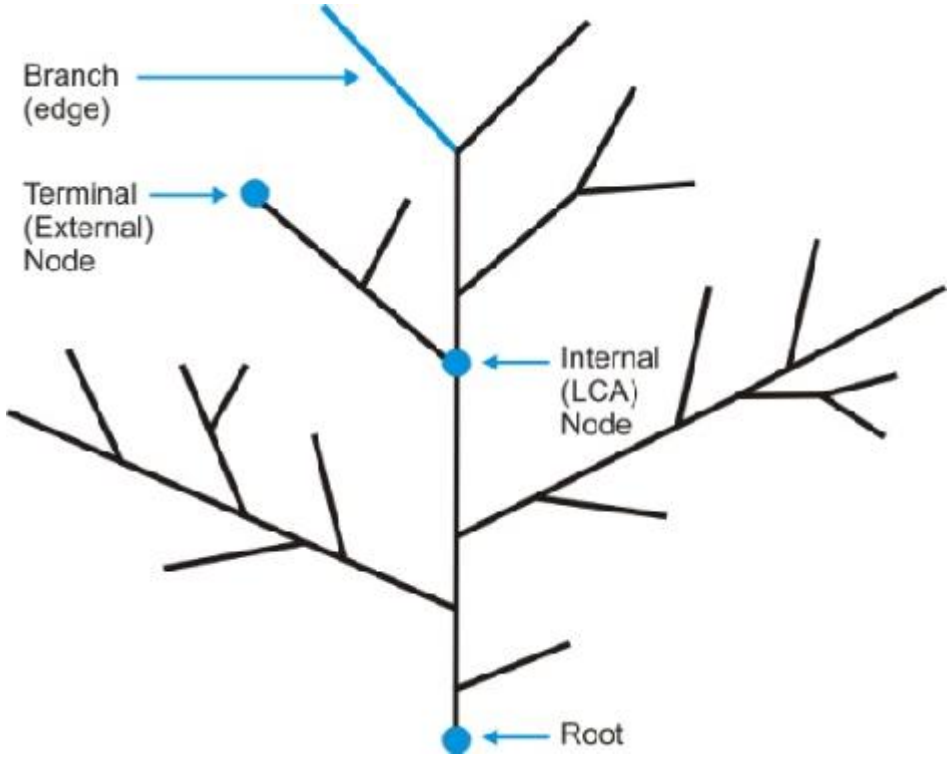
All organisms on earth have descended from a common ancestor, which means that any set of species, extant or extinct, is related

This relationship is called a *phylogeny*, and is represented by *phylogenetic trees*, which graphically represent the evolutionary history related to the species of interest

Phylogenetics infers trees from observations about existing organisms using morphological, physiological, and molecular characteristics

This phylogenetic tree shows the evolutionary relationships among six orders of Mammalian species (taxa). Taxa listed in grey are extinct

Phylogenetic Tree



Phylogenetic trees are composed of *branches*, also known as *edges*, that connect and terminate at *nodes*

Branches and nodes can be *internal* or external (*terminal*). The terminal nodes at the tips of trees represent operational taxonomic units (*OTUs*)

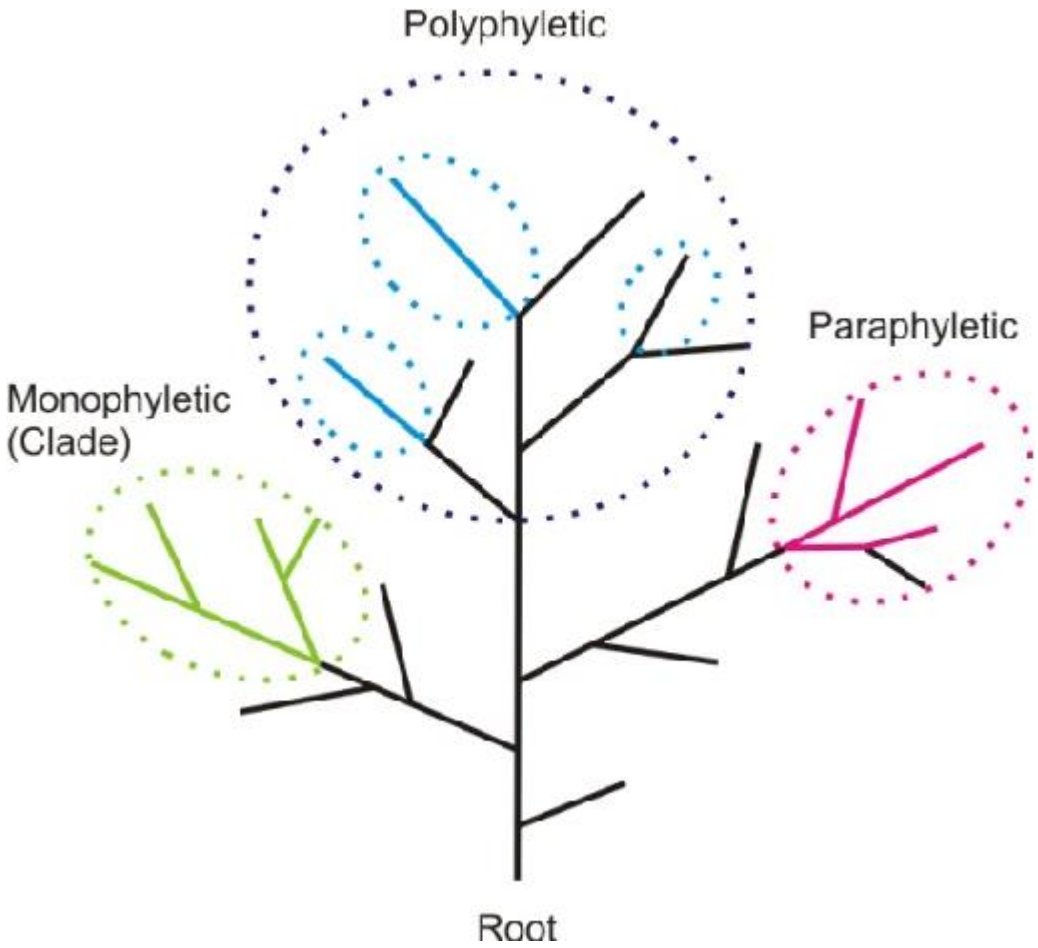
OTUs correspond to the molecular sequences or taxa (species) from which the tree was inferred. Internal nodes represent the last common ancestor (*LCA*) to all nodes that arise from that point Trees can be made of a single gene from many taxa (a species tree) or multi-gene families (gene trees)

A tree is considered to be "rooted" if there is a particular node or outgroup (an external point of reference) from which all OTUs in the tree arises

The root is the oldest point in the tree and the common ancestor of all taxa in the analysis

In the absence of a known outgroup, the root can be placed in the middle of the tree or a rootless tree may be generated

Branches of a tree can be grouped together in different ways



A *monophyletic* group consists of an internal LCA node and all OTUs arising from it

All members within the group are derived from a common ancestor and have inherited a set of unique common traits

A *paraphyletic* group excludes some of its descendents (for examples all mammals, except the marsupialia taxa)

And a *polyphyletic* group can be a collection of distantly related OTUs that are associated by a similar characteristic or phenotype, but are not directly descended from a common ancestor

Trees and Homology

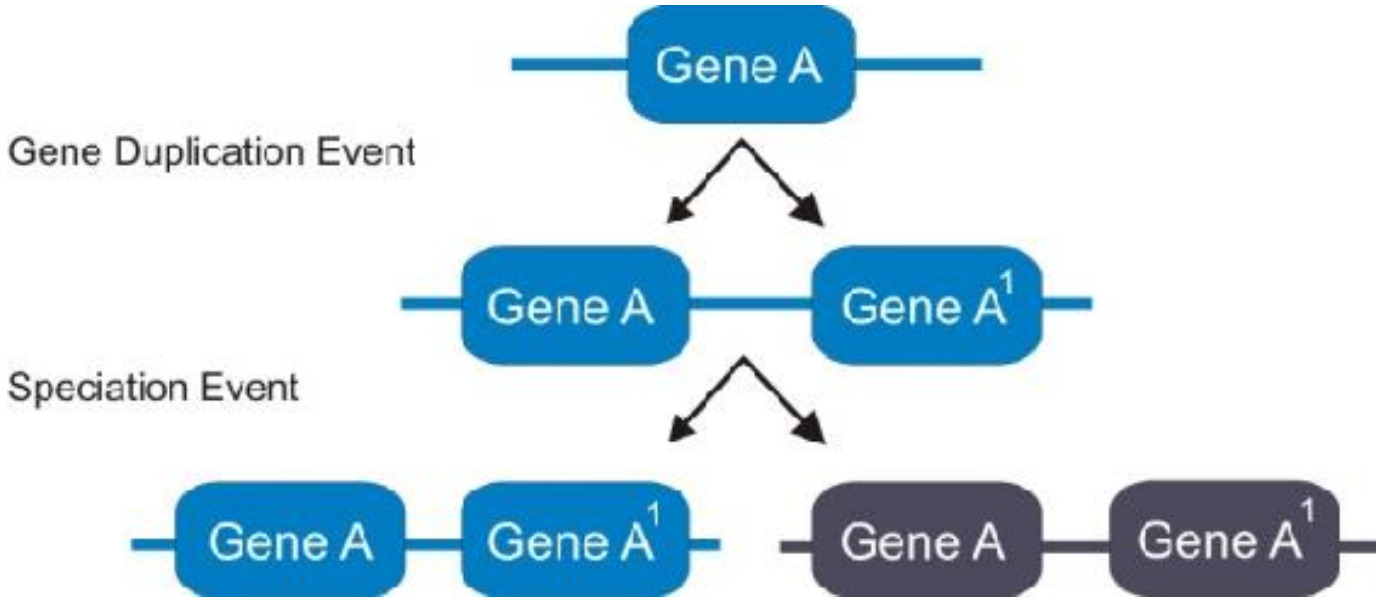
Evolution is shaped by homology, which refers to any similarity due to common ancestry. Similarly, phylogenetic trees are defined by homologous relationships.

Paralogs **are homologous sequences separated by a gene duplication event**

Orthologs **are homologous sequences separated by a speciation event (when one species diverges into two)**

Homologs **can be either paralogs or orthologs**

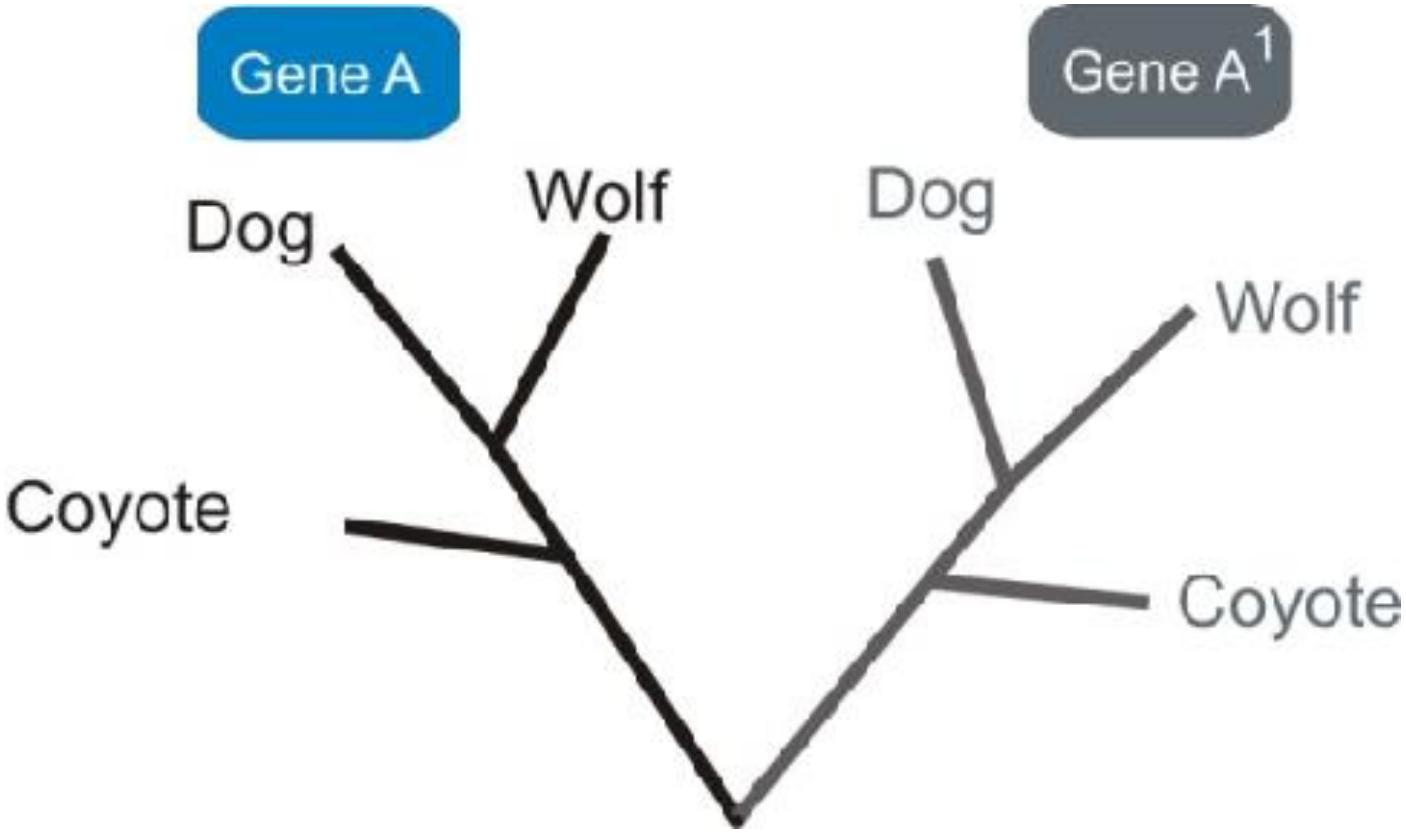
Molecular phylogenetic trees are drawn so that branch length corresponds to amount of evolution (the percent difference in molecular sequences) between nodes



Paralogs are created by gene duplication events. Once a gene has been duplicated, all subsequent species in the phylogeny will inherit both copies of the gene, creating *orthologs*

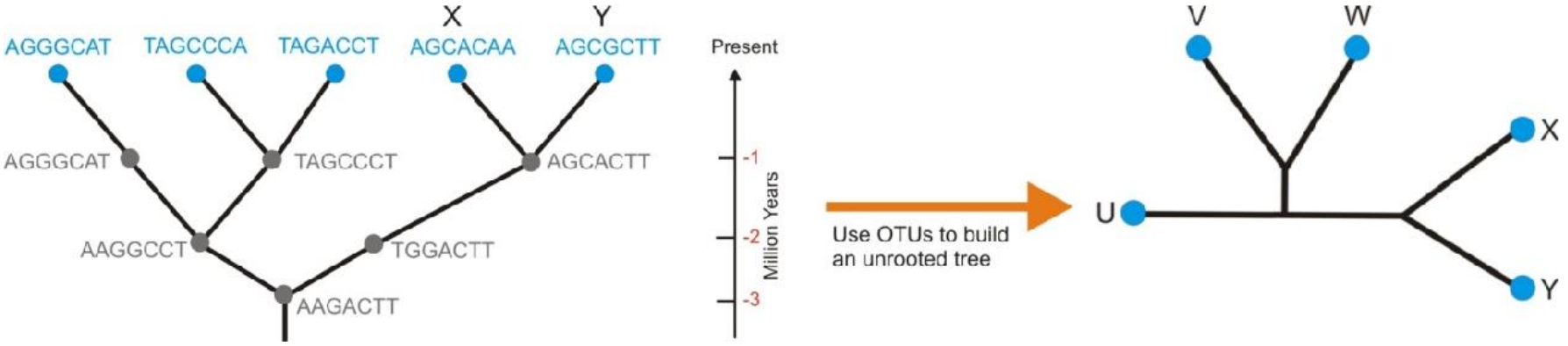
Evolutionary divergence of different species may result in many variations of a protein, all with similar structures and functions, but with very different amino acid sequences

Phylogenetic studies can trace the origin of such proteins to an ancestral protein family or gene



Mirror Phylogenies: Gene A and Gene A^1 are paralogs, whereas all instances of Gene A are orthologs of each other in different Canid species

Estimating Molecular Phylogenetic Trees



How a molecular sequence might evolve over time as a result of multiple mutations that results small, but evolutionarily important changes in a nucleotide sequence

At the protein level, these changes may not initially affect protein structure or function, but over time, they may eventually shape a new purpose for a protein within divergent species

OTUs can be used to build an unrooted phylogenetic tree that clearly depicts a path of evolutionary change

Molecular phylogenetic trees are generated from character datasets that provides evolutionary content and context

Character data may consist of biomolecular sequence alignments of DNA, RNA, or amino acids, molecular markers, such as single nucleotide polymorphisms (SNPs) or restriction fragment length polymorphisms (RFLPs), morphology data, or information on gene order and content

Evolution is modeled as a process that changes the *state* of a character, such as the type of nucleotide (AGTC) at a specific location in a DNA sequence; each character is a *function* that maps a set of taxa to distinct states

Steps in Phylogenetic Analysis

Assemble and align a dataset

Build (estimate) phylogenetic trees from sequences using computational methods

Stochastic models, and statistically test and assess the estimated trees

Assemble and Align Datasets

Identify a protein or DNA sequence of interest

DNA sequences of interest can be retrieved using NCBI BLAST or similar search tools based on E-values

A high score indicates the subject sequence retrieved with closely related to the sequence used to initiate the query

The smaller the E-value, the higher the probability that the homology reflects a true evolutionary relationship, as opposed to sequence similarity due to chance

As a general rule, sequences with E-values less than 10^{-5} are homologs of a query sequence

Create multiple sequence alignment

ClustalW, MSA, MAFFT, and T-Coffee, designed to perform multiple sequence on a given set of molecular data

Building Phylogenetic Trees

To build phylogenetic trees, statistical methods are applied to determine the tree topology

Calculation of the branch lengths that best describe the phylogenetic relationships of the aligned sequences in a dataset

The most common computational methods applied include distance-matrix methods, and

discrete data methods, such as maximum parsimony and maximum likelihood

There are several software packages, such as Paup*, PAML, PHYLIP, that apply most popular methods

Paup* is a commercially available program that implements maximum likelihood analysis for DNA data using different models. Paup* also includes a set of exact and heuristic methods for searching optimal trees

PAML (Phylogenetic Analysis by Maximum Likelihood) is open-access set of programs for phylogenetic analysis and evolutionary model comparison

PAML includes many advanced models—DNA- and AAbased models as well as codon-based models that can be used to detect positive selection

Many of the programs in PAML can model heterogeneity of evolutionary rates among sequence sites using distributions, and evolutionary dynamics of different sequence regions (concatenated gene sequences)

PHYLIP is another large suite of open-access programs for phylogenetic inference that estimates trees using pairwise distance, maximum parsimony, and maximum likelihood

The maximum likelihood programs can handle a few simple stochastic models and have good tree searching capabilities

Distance-Matrix Methods

Distance matrix methods compute a matrix of pairwise "distances" between sequences that approximate evolutionary distance

Distance-based methods tend to be in polynomial time and are quite fast in practice

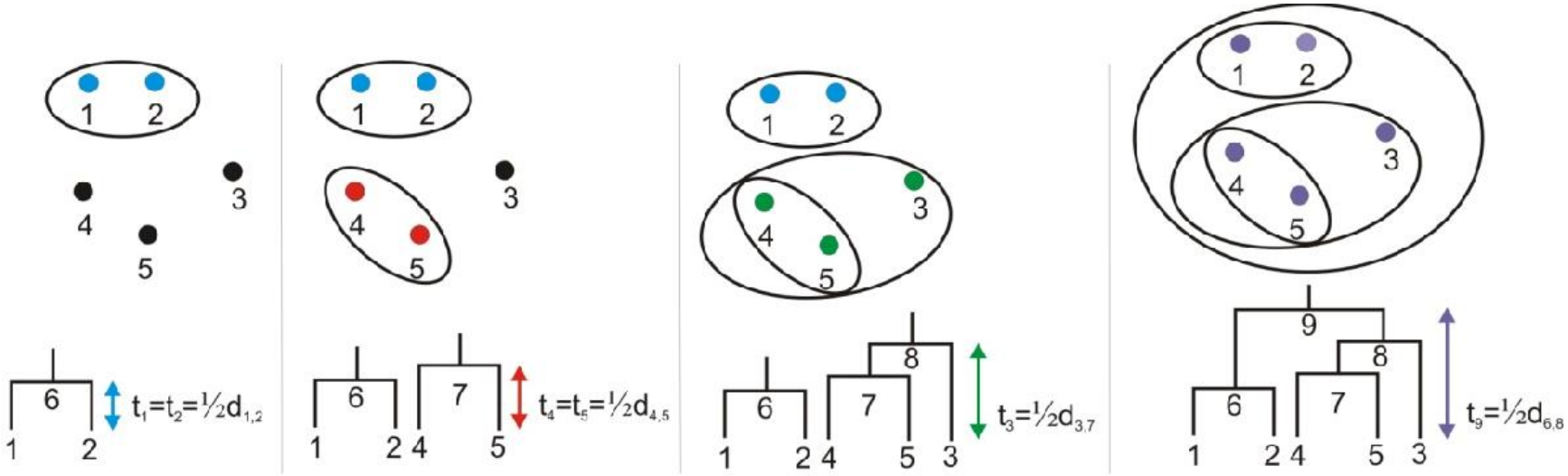
These methods use clustering techniques to compute evolutionary distances, such as the number of nucleotide or amino acid substitutions between sequences, for all pairs of taxa

They then construct phylogenetic trees using algorithms based on functional relationships among distance values

E.g. ***Unweighted Pair-Group Method with Arithmetic Mean (UPGMA)*****, which uses a sequential clustering algorithm**

$$d_{ij} = \frac{\sum d_{pq}}{|C_i||C_j|}$$

Where |Ci| and |Cj| are the number of sequences in clusters i and j



$$D_{6,8} = 1/6\ (d_{1,3} + d_{1,4} + d_{1,5} + d_{2,3} + d_{2,4} + d_{2,5})$$

***Transformed Distance Method*, which uses an outgroup as a reference, then applies UPGMA**

***Neighbor-Relations Method*, which applies 4-point condition to adjust the distance matrix, then applies UPGMA**

***Neighbor-Joining Method*, which arranges OTUs in a star, the finds neighbors sequentially to minimize total length of tree**

Discrete Data Methods

Discrete data methods examine each column of a multiple sequence alignment dataset separately and search for the tree that best represents all this information

Distance-based methods tend to be much faster than discrete data methods, they typically yield little information beyond the basic tree structure

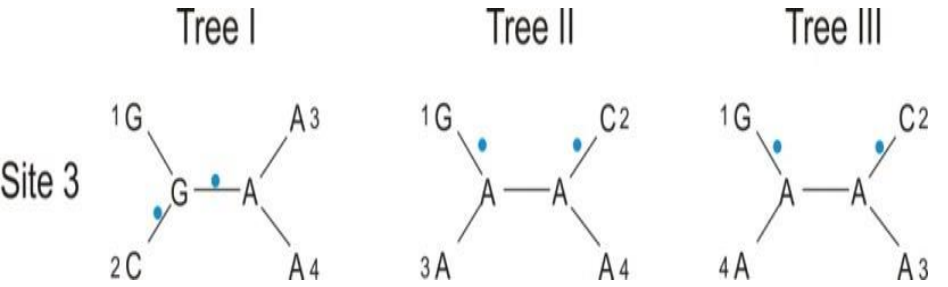
Discrete data analyses methods produce a separate tree for each column in the alignment, so it is possible to trace the evolution for specific elements within a given sequence, such as catalytic sites or regulatory regions, e.g. maximum parsimony, maximum likelihood, Bayesian MCMC

Maximum parsimony

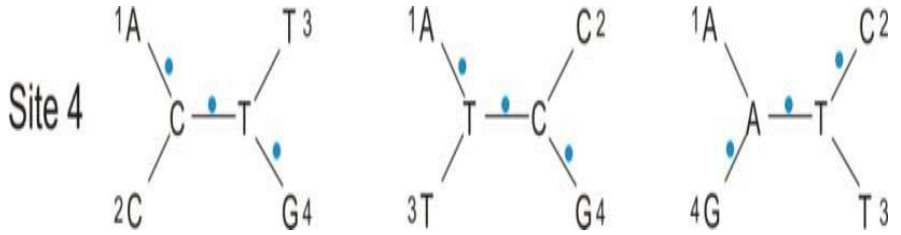
Searches the most parsimonious tree that requires the least number of evolutionary changes to explain differences observed

	Site								
Seq	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

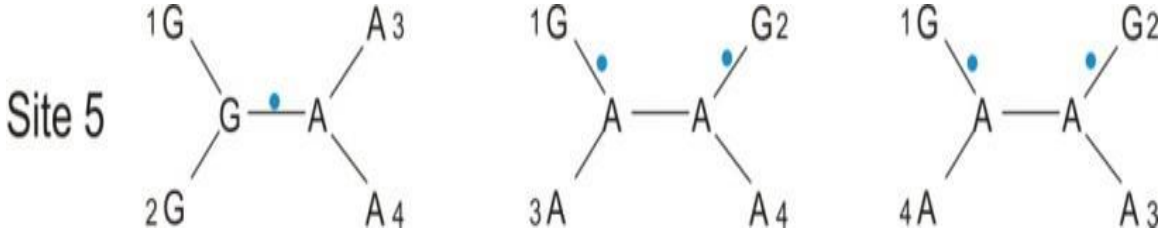
Sample sequences for a maximum parsimony study



Site 3 trees all require one evolutionary change



Site 4 trees all require three evolutionary changes



Site 5 trees vary in the number of evolutionary changes required

Maximum likelihood

Requires a probabilistic model for the process of nucleotide substitution

For a one-parameter model with rate of substitution 1 per site per unit time, the probability that the nucleotide at time *t* is *i* is:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4lt/3}$$

The probability that the nucleotide at time *t* is *j* is:

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4lt/3}$$

To set up a likelihood function, given *x* as the ancestral node and *y* and *z* as internal nodes, the probability of observing nucleotides *i*, *j*, *k*, *l* at the tips of the tree is computed as:

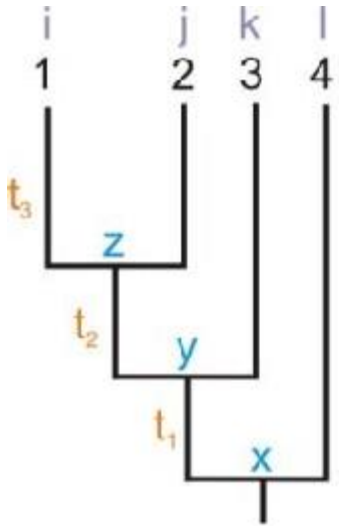
$$P_{xl}(t_1+t_2+t_3)P_{xy}(t_1)P_{yk}(t_2+t_3)P_{yz}(t_2)P_{zi}(t_3)P_{zj}(t_3)$$

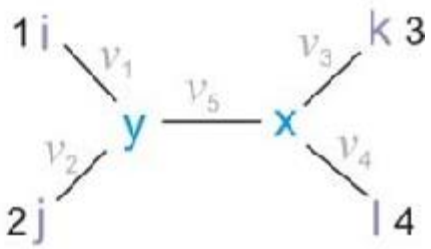
For the ancestral node (root) *x*, the probability of having nucleotide l in sequence 4 is calculated as:

$$P_{xl}(t_1+t_2+t_3)$$

Because *x*, *y*, and *z* can be any one of four nucleotides (ACGT), it is necessary to sum over all possibilities to obtain the probability of observing the configuration of nucleotides *i*, *j*, *k*, *l*, in sequences 1, 2, 3, 4, for a given hypothetical tree (see Figure 13.). This likelihood probability is calculated as:

$$h(I,j,k,l)= [\text{å}g_x P_{xl}(t_1+t_2+t_3)]\ [\text{å}P_{xy}(t_1)P_{yk}(t_2+t_3)]\ [\text{å}P_{yz}(t_2)P_{zi}(t_3)\ P_{zj}(t_3)]$$





Different types of model trees for the derivation of the maximum likelihood function

Bayesian MCMC

Requires a stochastic model of evolution, but creates a probability distribution on a set of trees or aspects of evolutionary history

Discrete data methods are generally considered to produce the best estimates of evolutionary history

However, these methods can be computationally expensive, and it can take weeks or months to obtain a reasonable level of accuracy for moderate to large datasets with 100 or more OTUs

Stochastic Models of Evolution

Evolutionary changes in molecular sequences result from mutations, some of which occur by chance, others by natural selection

Rates of change can also differ among OTUs, depending on several factors ranging from GC content to genome size

To accurately estimate phylogenetic trees, assumptions must be made about the substitution process and those assumptions must be stated in the form of a stochastic evolutionary model

These probabilistic models are used to rank trees according to likelihood:

P(data|tree). From a Bayesian perspective, they rank trees according to a posterior probability:
P(tree|data)

The objective of probabilistic models is to find likelihood or posterior probability of a particular taxonomic feature

$$P(x_{\cdot} \mid T, t_{\cdot})$$

Where $x_{\cdot}$ is x^j for j=1…n, T is a tree with *n* leaves with sequence *j* at leaf *j*, and $t_{\cdot}$ are tree edge lengths

A few popular stochastic models of evolution include:

single parameter Jukes-Cantor (JC) method
Kimura 2-parameter (K2P)
Hasegawa-Kishino-Yano (HKY)
Equal-Input

Some software programs, such as Paup*, will automatically use a default model for the tree estimation method chosen

The JC method is the easiest one to comprehend, because it assumes that if a site changes its state, it changes with equal probability to the other states
This is not very realistic, however, as some sites are known to evolve more rapidly than others, and some sites may be invariable and not allowed to change at all

Hidden Markov Models (HMMs)

Profile hidden HMMs are a form of Bayesian network that provides statistical models of the consensus structure of a sequence family

In the HMM format, each position in the model corresponds to a site in the sequence alignment

For each position, there are a number of possible states, each of which corresponds to a different rate of evolution

In addition, transitions between all possible rate-states at adjacent positions

Transition probabilities capture any tendency for patterns of rates to occur in successive sites

Assessing Trees

Tree estimating algorithms generate one or more optimal trees

This set of possible trees is subjected to a series of statistical tests to evaluate whether one tree is better than another and if the proposed phylogeny is reasonable

Common methods for assessing trees include:

Bootstrap and Jackknife Resampling methods

Analytical methods, such as parsimony, distance, and likelihood

Bootstrap Analysis

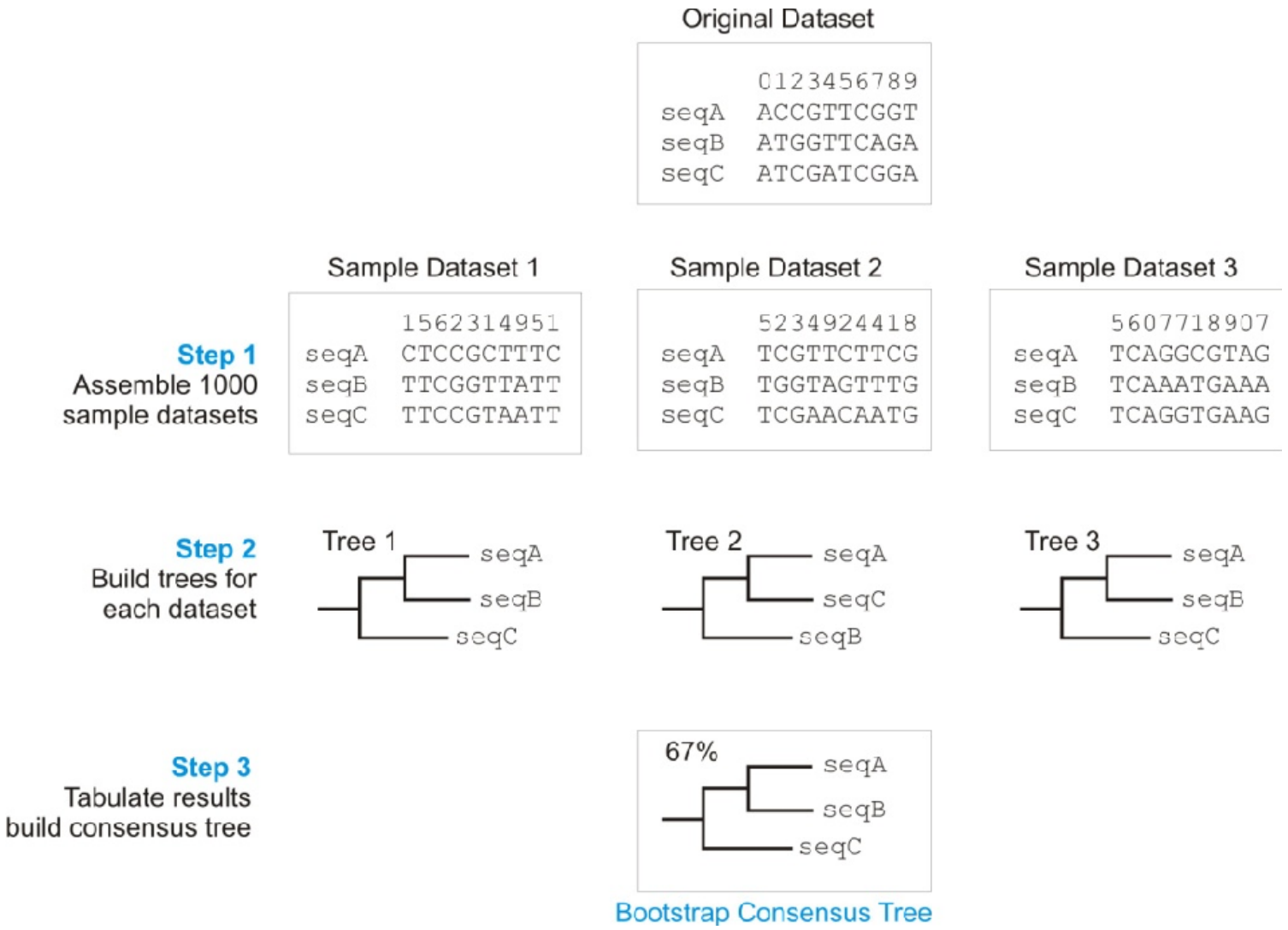
A bootstrap is a statistical method for assessing trees that takes its name from the fact that it can "pull itself up by its bootstraps" and generate meaningful statistical distributions from almost nothing

Using bootstrap analysis, distributions that would otherwise be difficult to calculate exactly are estimated by repeated creation and analysis of artificial datasets

In a *Non-parametric bootstrap*, artificial datasets generated by resampling from original data

In a *parametric bootstrap*, data is simulated according to hypothesis tested

The objective of any bootstrap analysis is to test whether the whole dataset supports the tree



Basic steps in any bootstrap analysis. Sample datasets are automatically generated from an original dataset. Trees are then estimated from each sample dataset. The results are compiled and compared to determine a bootstrap consensus tree

Phylogenetic Analysis Tools

There are several good online tools and databases that can be used for phylogenetic analysis. These include:

PANTHER
P-Pod
Pfam
TreeFam, and
PhyloFacts structural phylogenomic encyclopedia

Each of these databases uses different algorithms and draws on different sources for sequence information, and therefore the trees estimated by PANTHER, for example, may differ significantly from those generated by P-Pod or PFam

As with all bioinformatics tools of this type, it is important to test different methods, compare the results, then determine which database works best (according to consensus results, not researcher bias) for studies involving different types of datasets

In addition, to the phylogenetic programs already mentioned, a comprehensive list of more than 350 software packages, web-services, and other resources can be found at:

http://evolution.genetics.washington.edu/phylip/software.html

PANTHER (pantherdb.org)

Protein ANalysis THrough Evolutionary Relationships, known by its acronym PANTHER, is a library of protein families and subfamilies indexed by function

PANTHER is composed of both a library and index. The library is a collection of "books" that represent a protein family as a collection of multiple sequence alignments, HMMs, and a family phylogenetic tree

Functional divergence within the tree is represented by dividing the parent tree into child trees and HMMs based on shared functions

These subfamilies enable database curators to more accurately capture functional divergence of protein sequences as inferred from genomic DNA

P-POD (ortholog.princeton.edu)

Princeton Protein Orthology Database (P-POD) combines results from multiple comparative methods with curated information culled from the literature

Designed to be a resource for experimental biologists seeking evolutionary information on genes on interest, P-POD employs a modular architecture, based on their Generic Model Organism Database (GMOD)

P-POD can be accessed from their web service or downloaded to run on local computer systems

P-POD accepts FASTA-formatted protein sequences as input, and performs comparative genomic analyses on those sequences using OrthoMCL and Jaccard clustering methods

P-POD database contains both phylogenetic information and manually curated experimental results

The site also provides many links to sites rich in human disease and gene information. This tool may be particularly helpful for bioinformaticists and statisticians developing comparative genomic database tools and resources

Pfam (pfam.sanger.ac.uk/)

PFam is a collection of Protein FAMilies represented by multiple sequence alignments and HMMs

It contains models of protein clans, families, domains, and motifs, and uses HMMs representing conserved functional and structural domains

Pfam can be used to retrieve the domain architectures for a specific protein by conducting a search using a protein sequence against the Pfam library of HMMs

This database is also helpful for proteomes and protein domain architecture analysis

There are two versions of the Pfam database:

Pfam–B is generated automatically from ProDom, using PsiBLAST
Pfam-A is hand-curated from custom multiple sequence alignments

TreeFam (TreeFam.org)

TreeFam is a curated database of phylogenetic trees and orthology predictions for all animal gene families that focuses on gene sets from animals with completely sequenced genomes

Orthologs and paralogs are inferred from phylogenetic tree of gene family

Release 4 contains curated trees for 1314 families and automatically generated trees for another 14351 families

TreeFam is a two-part database:

TreeFam-B contains automatically generated trees, and
TreeFam-A consists of manually curated trees

PhyloFacts (phylogenomics.berkeley.edu/phylofacts)

PhyloFacts is an online phylogenomic encyclopedia for protein functional and structural classification

It contains more than 57,000 "books" for protein superfamilies and structural domains

Each book contains heterogenous data for protein families, including multiple sequence alignments, one or more phylogenetic trees, predicted 3-D protein structures, predicted functional subfamilies, taxonomic distributions, GO annotations, and PFAM domains

HMMs constructed for each family and subfamily permit novel sequences to be classified to different functional classes

PhyloFacts seeks to correct and clarify annotation errors associated with computational methods for predicting protein function based on sequence homology

PhyloFacts can be used to search for protein structure prediction or functional classification for a particular protein sequence

Applied Molecular Phylogenetics

Tracing the evolution of man

To compare and analyze variation in DNA sequences using modern human and Neanderthal mitochondrial DNA (mtDNA)

For this study, 206 modern human mtDNAs and parts of two Neanderthal mtDNAs sequences derived from skeletal remains were used to generate an initial dataset

Genetic distance was first estimated using the Jukes-Cantor single parameter model followed by Kimura 2-Parameter model to distinguish between transition and transversion probabilities

A phylogenetic tree representing primate evolution was generated using pairwise genetic distances between primate Hypervariable regions I and II of mtDNA

Chasing an epidemic: SARS

Available genomic data was used to reconstruct the progression of the SARS epidemic over time and space

To conduct this phylogenetic analysis, researchers used the neighborjoining method to construct a phylogenetic tree of spike proteins in various coronaviruses and identify the viral host (a Himalyan palm civet)

They then obtained 13 SARs genome sequences with documented information on the date and location of the sample

The neighbor-joining method and a distance matrix based on Jukes-Cantor model, were used to generate an epidemic tree, from which it was possible to identify the origin (date and location) of the virus by observing progression of mutations over time

Barking up the right tree

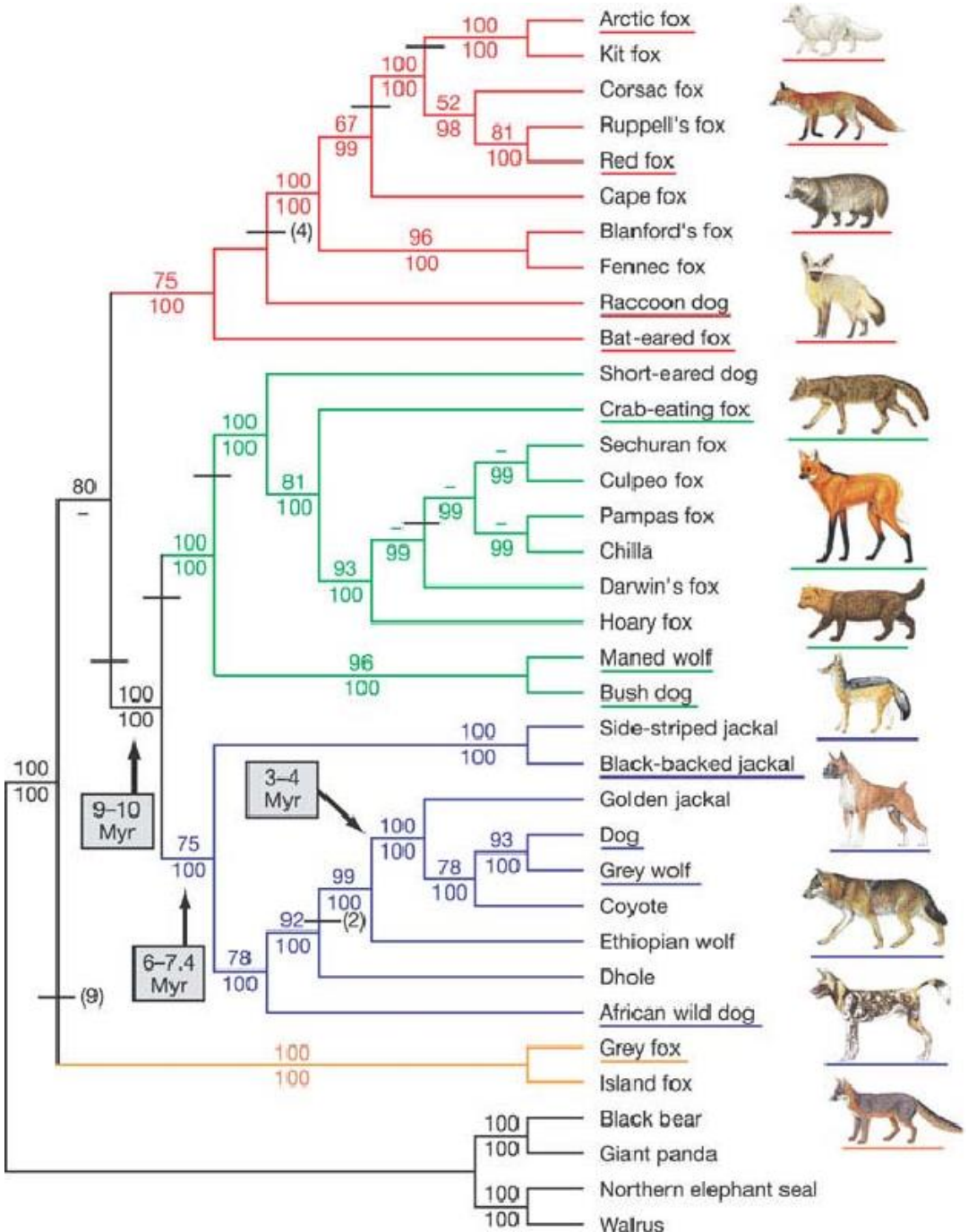
The canid family phylogenetic tree is based on 15 kb of exon and intron sequence

It was constructed using the maximum parsimony method and represents the single most parsimonious tree

A good example of how phylogenies are referenced in the literature, this tree includes bootstrap values and Bayesian posterior probability values listed above and below internodes, respectively

Dashes indicate bootstrap values below 50%

In addition, divergence time in millions of years (Myr) is indicated for three nodes



Phylogenetic Tree of the Canid family