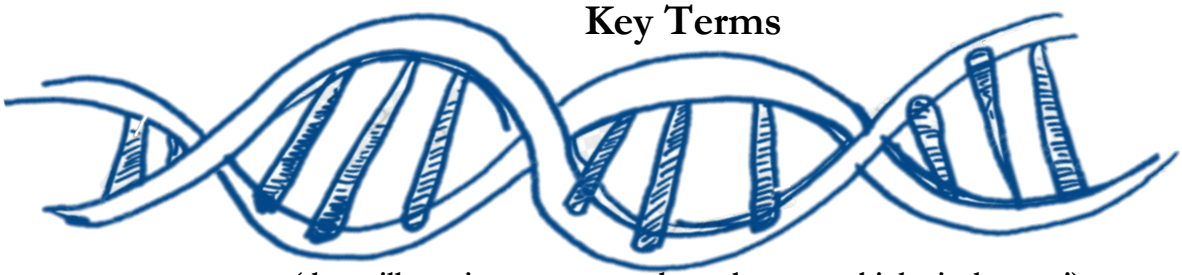
Genes & Evolution

Key Terms



(that will continue to appear throughout your biological career!)

BLAST – Short for "Basic Local Alignment Search Tool." This is an algorithm used for comparing similarity between two or more biological sequences (i.e. DNA). This is often used to compare a sequence of interest to an entire database of sequence such as Genbank. It can also be used to compare two specific sequences to each other. This is a term that you will see a LOT.

Central dogma of Biology – The concept that DNA is transcribed into RNA. RNA is then translated into a protein. Protein is then used as the functional unit by the cell. i.e. DNA -> RNA -> Protein

FASTA data format or file – A very common way to store biological sequence data, such as DNA sequences, in a text format. Often this is the universal/base format that might need to be converted into more specific formats for use by some software that analyze DNA sequences. FASTA format consists of a descriptor line that start with a ">" symbol followed by the sequence data starting on the next line (example below). A common mistake of student new to using FASTA format is forgetting to include the ">" but this symbol is crucial for software to recognize and read a FASTA file. The descriptor line is limited to 80 characters and typically includes data such as: species name, catalog or accession number, and any other important data to identify what the sequence is (such as an individual ID for the specific organism that was sequenced). These type of files are a text file and typically with ".fas," ".fa," or ".fasta" extension i.e. *exampleDNA.fas*

Example FASTA:

>*Homo sapien* | accession#:XKY_869605750 | gene: COI
AGCGCGCGCTCGCGACGGGTCTAGCTAGCTAGTCGATCGTAGTCGAT
CGTAGCTGACTGATCGTAGCTAGCTGATCGATCGTAGCTAGCTAGCTG
ATCGATGCTAGCTAGTCGATCGTAGCTAGCTAGCTAGTCGATCGATGC
TAGCTAGCTAGCGTGCTGATCGCATGCATGCATGCTAGCTAGCTGATC
GATGCTAGCTGATCGATGCTAGCTAGCTAGCTAGCTGATCGTAGCTAG
CTAGCTAGTCGATCGTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCT
AGCTAGCTACTAGCTAGCTAGCTAGCTAGCTACGTAGCTAGCTAGTCA
GTCGATCGTAGCATCGATTTTTTTTTTTTCGATGCTAGCTAGCT

Gap – A gap in a multiple sequence alignment is a portion of a biological sequence (i.e DNA) that is found in at least one individual but missing in at least one other.

Example:

Frog: AATCGG
Salamander: A---CGG

Genetic Database Repository – Databases that are used to house genetic information generated by researchers. Most peer-reviewed journals require genetic information from a study to be uploaded to one of these databases as a stipulation of publishing. The information contained on these databases is made publicly available to other researchers. Many studies are published based entire on genetic sequences that have been accessed from one of these types of databases. Example databases include: the European Nucleotide Archive (https://www.ebi.ac.uk/ena), International Barcode of Life (https://ibol.org/), DNA databank of Japan (https://www.ddbj.nig.ac.jp/index-e.html), and Genbank (https://www.ncbi.nlm.nih.gov/genbank/). Genbank is the most commonly used database in North America and the one we will focus on in this course.

Genbank – This a **massive** repository of genetic data maintained by The National Center for Biotechnology Information (NCBI). The data is publicly available and free to access from the Genbank website: https://www.ncbi.nlm.nih.gov/genbank/

Genetic distance – The number of mutations/substitutions that have accumulated between two species after they have diverged. A simplistic way to think about it: how genetically different two species are from each other.

Genetic variation – The differences in DNA amongst individuals, populations, or species. Random mutation is main cause of genetic variation however other factors play a role to maintaining the amount of variation.

Indel – An indel is either **in**sertion or **del**etion of a portion of a biological sequence (i.e DNA) in at least one sequence contained in a multiple sequence alignment. Indels lead to gaps in in the alignment and should be heavily scrutinized. Indels and gap formation between closely related species are highly unlikely from a biological standpoint and often the result of error.

Ingroup – The species/individuals of interest to your study.

Maximum likelihood tree – This method is much more complex mathematically than neighbor joining or maximum parsimony trees. The algorithm generates numerous trees and assigns a likelihood score to each. The best tree is the tree with the highest likelihood score. This method is can be computational demanding and be slow to run. It is however, very commonly used in modern phylogenetic studies and considered a robust method even for very large datasets.

Maximum parsimony tree – This is a character-based method of phylogenetic tree construction. The maximum parsimony method builds numerous trees from the genetic sequences provided and scores them. The "best" tree is the tree that requires the fewest number of mutations/substitutions for all sequences to be derived a common ancestor. Maximum parsimony is a fairly rapid method of tree construction and is useful for exploratory purposes. While still being useful this method can suffer from improper branch lengths. An additional problem associated with this method is that there can be multiple "best" trees because there might be multiple tree topologies that have the same number of mutations/substitutions.

Multiple sequence alignment – An alignment of three or more biological sequences including DNA, RNA, or protein sequences. Multiple sequence alignments are commonly used to infer homology of a particular sequence (same evolutionary origin) and infer phylogenetic relationships between organisms.

Mutation – A random change in a DNA sequence that take place between generations. Mutations can have a negative, positive, or neutral impact on the organism. Mutation is the way in which new variation is formed and is the driving force of evolution.

Negative selection – Also known as purifying selection. The selective force in which deleterious (or bad) alleles (for survival and reproduction) are removed. Most mutations of a non-synonymous site will be negatively selected against; therefore - will not be passed to the next generation. Very often this selective force will prevent a fetus from ever being born in the first place. If the organism is born it will be at a disadvantage and unlikely to pass its genetic make-up to next generation. This is important for understanding how mutation accumulates across time and how most genetic variation is neutral.

Neighbor joining tree – This is a distance-based method of phylogenetic tree construction. Neighbor join trees are useful for exploratory purposes because they are computational easy and can be constructed very rapidly even for large datasets. This method of tree construction is not used as frequently in published modern genetic studies because the potential for bias in tree construction.

Neutral Theory – The concept that the vast majority of genetic variation (i.e. differences in DNA sequences between individuals/populations/species) is a result of random change in DNA between generations that has no impact on the survivorship or fitness of the individual. This commonly takes place in the synonymous sites in the gene, therefore has no impact on the resulting protein.

Non-synonymous mutation - A mutation in a gene that alters the resulting protein produced. This type of mutation usually results in a negative fitness or survivorship cost to the individual. However, sometimes the altered protein can give a positive fitness advantage to the organism.

Nucleotides – The nitrogenous bases in nucleic acids (i.e. DNA or RNA). The four nucleotides found in DNA are adenine (A), thymine (T), guanine (G) and cytosine (C). These are the characters that make up the DNA or gene sequences used to build phylogenetic trees and assess evolutionary relationships between species. Example short gene sequence: >
AGGGGCTCGAGCTCGGCTAGCTGCGCTAGCGGCTAGC

Outgroup – The outgroup in a phylogenetic tree is distantly related species used for rooting a phylogenetic tree or establish a common ancestor to all species included in tree. Correctly choosing an outgroup is not always a particularly easy thing to do. You need to choose a "distantly" related species, but it cannot be *too* distant, or you may create error in tree construction. For example, if you are investigating the evolutionary relationship of some different mice species found in Oklahoma, you would not want to choose a Nile crocodile as your outgroup. You would also not want to choose another mouse that lives near-by in Northern Texas that looks and acts similarly to the mice you're interested in. The choice of an appropriate outgroup often requires spending time reading the published literature to understand taxonomic relationship between your species of interest. It may also require some trial and error to decide what is the best outgroup to use. A good suggestion is to ask what another biologist would suggest for an outgroup to get multiple opinions when starting out!

Parsimony – The concept that the simplest explanation that can explain the data is the most likely. This is a fundamental concept across biological systems. Parsimony is a similar principle to the philosophical concept of "Occam's razor." In regard to phylogenetics, it is more likely that a particular trait evolved once and was then passed onto descendants rather than a trait continually evolved multiple times in a descendant. In a DNA sequence this would mean that a particular mutation/substitution or allele arose by chance and was passed down to descendants.

Phylogenetic tree – A technique used to infer evolutionary history and relatedness of different organisms for example species, populations, or families.

Polymerase Chain Reaction (PCR) - Molecular technique used to isolate, amplify, and sequence a particular region of genome or specific gene of interest.

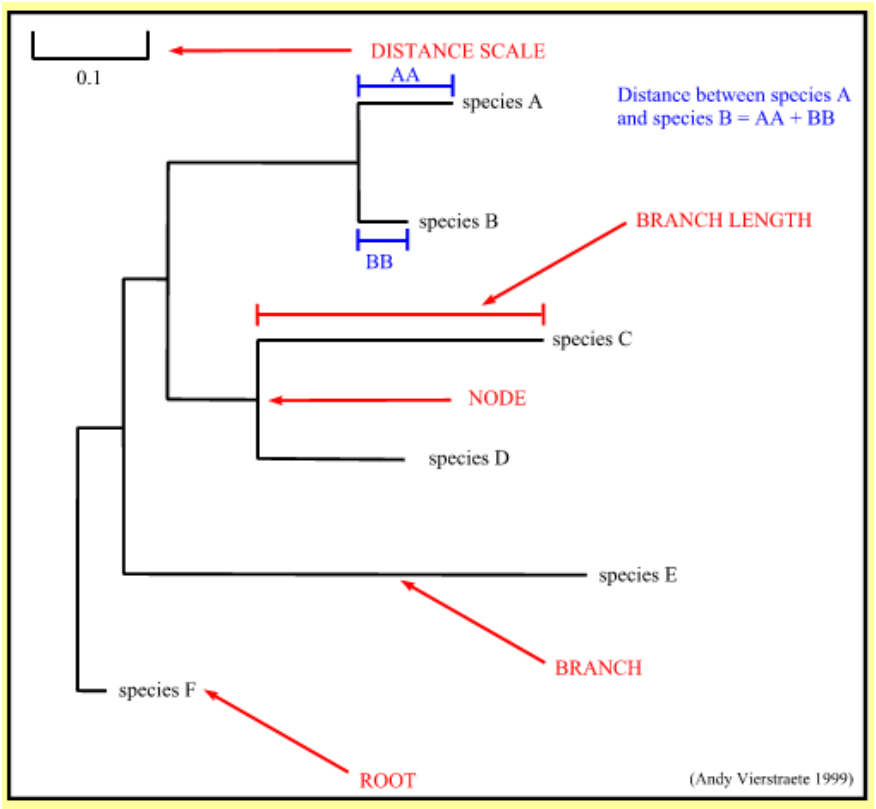
Positive selection – A much less common occurrence than negative selection. In positive selection a random mutation actually gives the individual an advantage over others. This means that the individual is more likely to breed and pass the advantage to its offspring and those offspring to their offspring and so on. The result of this is that the mutation will proliferate through the population.

Root – The root of a phylogenetic tree is the common ancestor for all the species included in the tree. An outgroup species is commonly used to generate a root. Rooting a tree allows directional inferences of evolution to be inferred i.e. timing of evolutionary events.

Synonymous mutation – A mutation in a gene that does NOT alter the resulting protein after translation. Synonymous mutation account for the majority of genetic diversity. Personal study tip: it helped me to remember the difference between synonymous vs. non-synonymous by remember you get the SAME protein with synonymous mutations and both "Same" and "Synonymous" start with the letter "S."

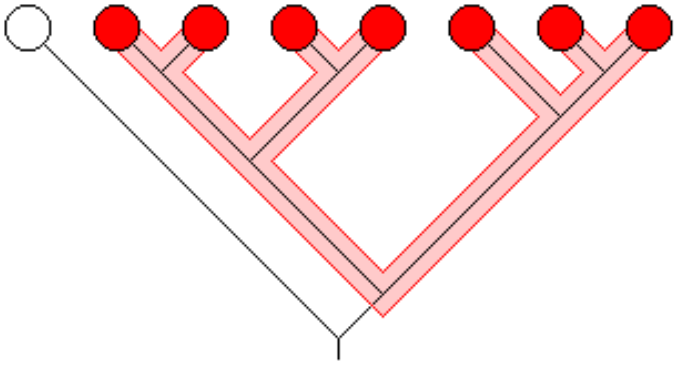
Topology / Tree Topology – The branching pattern observed in a phylogenetic tree.

Basic Phylogenetic Tree Diagrams
&
Important Terms



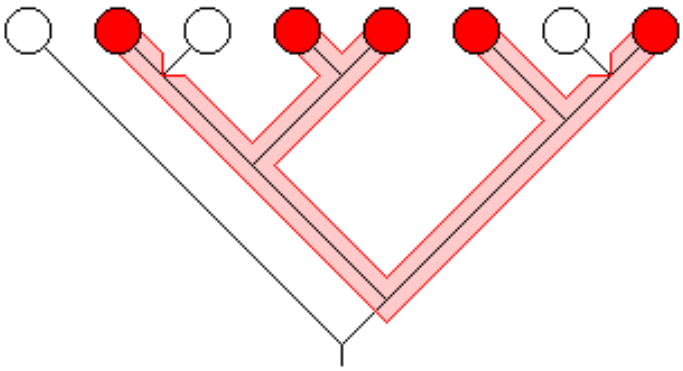
- **Node :** a node represents a taxonomic unit. This can be a taxon (an existing species) or an ancestor (unknown species : represents the ancestor of 2 or more species).
- **Branch :** defines the relationship between the taxa in terms of descent and ancestry.
- **Topology :** is the branching pattern.
- **Branch length :** often represents the number of changes that have occurred in that branch.
- **Root :** is the common ancestor of all taxa.
- **Distance scale :** scale which represents the number of differences between sequences (e.g. 0.1 means 10 % differences between two sequences)

Monophyletic taxon (clade) :



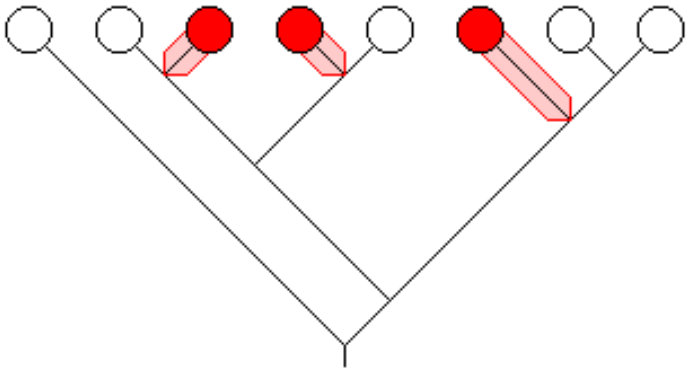
Monophyletic taxon: A group composed of a collection of organisms, including the most recent common ancestor of all those organisms and all the descendants of that most recent common ancestor. A monophyletic taxon is also called a clade.

Paraphyletic taxon :



Paraphyletic taxon: A group composed of a collection of organisms, including the most recent common ancestor of all those organisms. Unlike a monophyletic group, a paraphyletic taxon does not include all the descendants of the most recent common ancestor.

Polyphyletic taxon :



Polyphyletic taxa: A group composed of a collection of organisms in which the most recent common ancestor of all the included organisms is not included, usually because the common ancestor lacks the characteristics of the group. If your samples show this pattern the name of the overall taxon must be assessed.

Image and definition sources:
https://ucmp.berkeley.edu/glossary/gloss1/phyly.html

https://users.ugent.be/~avierstr/principles/phylogeny.html

Notes