E-content

M.Sc. Zoology (Semester II)
CC8- Biochemistry

Unit: 3

1. Variation in the evolution of protein and DNA sequences

2. Molecular phylogenies

Dr Gajendra Kumar Azad
Assistant Professor
Post Graduate Department of Zoology
Patna University, Patna
Email: gkazadpatnauniversity@gmail.com

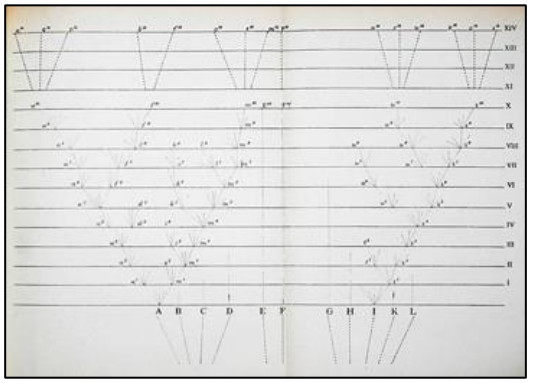
What is Molecular Evolution ?

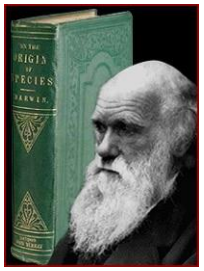
- Molecular evolution refers to the changes in living systems that have occurred with time subsequently leading to the formation of unicellular and multicellular organisms. Molecular evolution deals with the mechanisms underlying evolution at the molecular level.

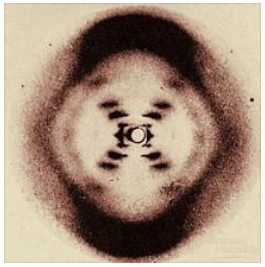
- Molecular evolution address broad range of problems:
  1. Use **DNA/protein** to study the evolution of **organisms**, e.g. population structure, geographic variation and phylogeny
  2. Use different **organisms** to study the evolution process of **DNA/protein**

A brief historical perspective

- Darwin first came up with the idea that living organisms are evolutionarily related
- Molecular evolution became a science following discovery of DNA and crack of genetic code
- Insulin: first protein sequenced (Sanger, 1955), and sequence compared across species.
- Neutral theory: Motoo Kimura, Thomas Jukes (1968,69)
- Effect of population size: Michael Lynch (2000s)

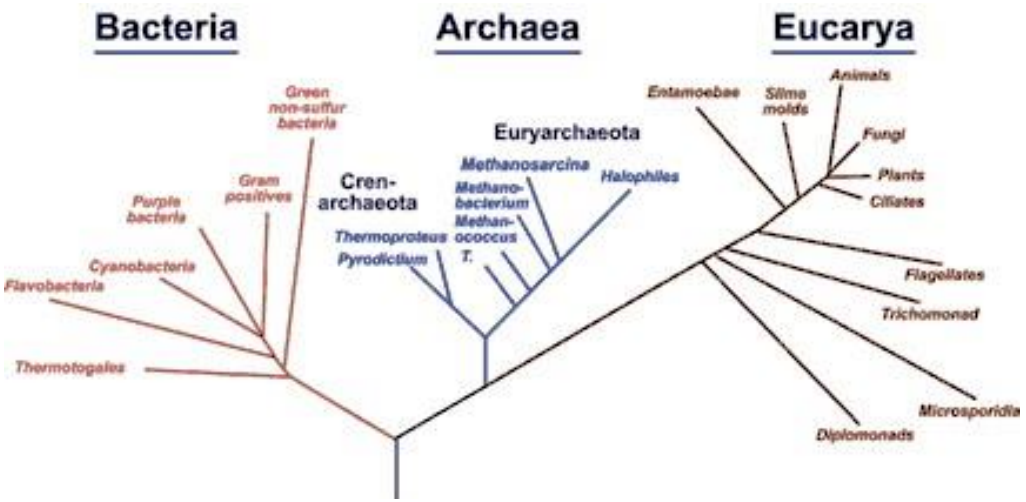


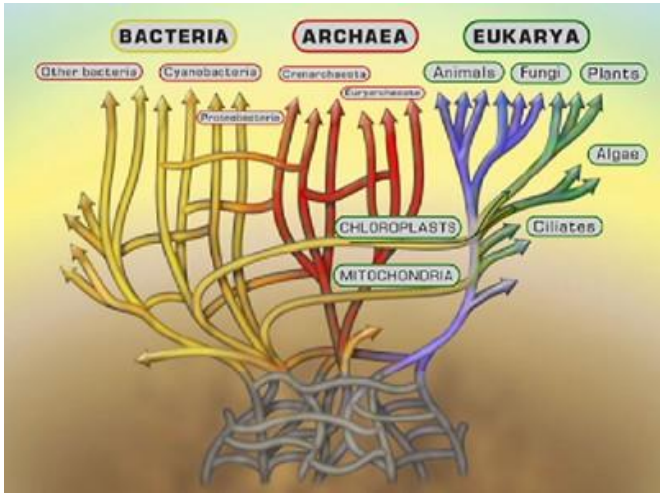




- Until 1970s, cellular organisms were divided into eukaryotes (have nucleus) and prokaryotes (no nucleus)

- Using 16S rRNA gene sequence, Carl Woese redefined three domains





Ford Doolittle

- To recover evolutionary relationships from amino acid or nucleotide sequences, rigorous models of molecular evolution are needed.

Mutations in DNA and protein

The composition of any organism genome is the consequence of the molecular and population genetic forces that act upon that particular genome. Novel genetic variants will arise through mutation and will spread and be maintained in populations due to genetic drift or natural selection.

Mechanism of molecular evolution at DNA level is generally due to these three processes: mutation, insertion, and deletion

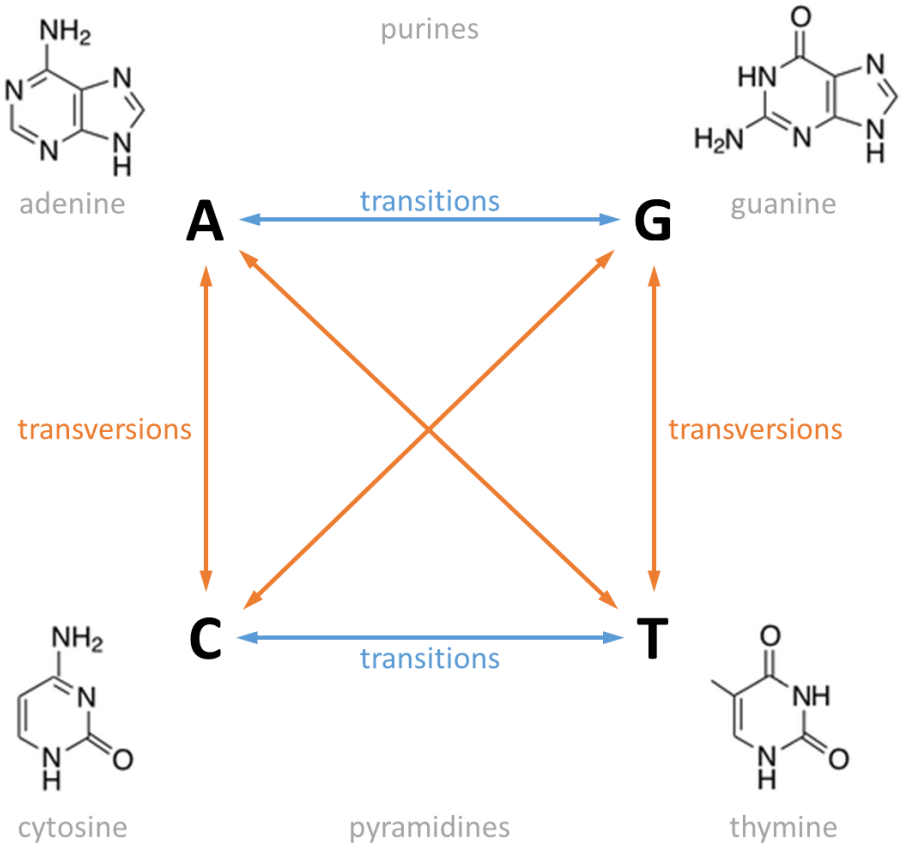
GACGACCATAGACCAGCATAG

GACTACCATAGA-CTGCAAAG

- **Transition**: A ↔ G, C ↔ T
- **Transversion:** purine ↔ pyrimidine

In genetics, a distinction is made between two types of **mutations** (the replacement of one nucleotide by a different nucleotide).
A **transition** changes a purine nucleotide (two rings) into another purine (A ↔ G), or changes a pyrimidine nucleotide (one ring) into another pyrimidine (C ↔ T). All other mutations in which a purine is substituted for a pyrimidine, or *vice versa*, are called **transversions**.

Although in theory there are only four possible transitions and eight possible transversions, in practice transitions are more likely than transversions because substituting a single ring structure for another single ring structure is more likely than substituting a double ring for a single ring. Also, transitions are less likely to result in amino acid substitutions (due to wobble base pair) and are therefore more likely to persist as **silent substitutions** in populations.



- **Synonymous mutation** -> do not change amino acid
- **Nonsynonymous mutation** -> change amino acid

- **Nonsense** mutation: point mutation resulting in a pre-mature stop codon
- **Missense** mutation: resulting in a different amino acid
- **Frameshift** mutation: insertion / deletion of 1 or 2 nucleotides
- **Silent** mutation: the same as nonsynonymous mutation

- Neutral mutation: mutation has no fitness effects, invisible to evolution (neutrality usually hard to confirm)
- Deleterious mutation: has detrimental fitness effect
- Beneficial mutation: *has beneficial effect on fitness*

Fitness = ability to survive and reproduce

Negative Selection and Positive Selection

- There are two types of natural selection in biological evolution: Positive (Darwinian) selection promotes the spread of beneficial alleles, and negative (or purifying) selection hinders the spread of deleterious alleles
- **Negative selection (purifying selection)**
  - Selective removal of deleterious mutations (alleles)
  - Result in conservation of functionally important amino acids
  - Examples: ribosomal proteins, RNA polymerase, histones

- **Positive selection (adaptive selection, Darwinian selection)**
  - Increase the frequency of beneficial mutations (alleles) that increase fitness (success in reproduction)
  - Examples: male seminal proteins involved in sperm competition, membrane receptors on the surface of innate immune system
  - Classic examples: Darwin's finch, rock pocket mice in Arizona (however the expression level of these genes instead of their protein sequence are targeted by selection)

Synonymous mutation rate (Ks): Mutations/substitutions of DNA base pairs that do not result in a change of amino acid sequence. Also known as a silent mutation.

Non-synonymous mutation rate (Ka): Mutations/substitutions of DNA base pairs that result in a single amino acid change on a given polypeptide. Also known as a substitution mutation.

Non-synonymous/Synonymous mutation ratio (Ka/Ks): Ratio of mutations that change a specific protein structure (non-synonymous, Ka) to mutations that do not change a specific protein.

This ratio is used to estimate the selection pressure a given protein or section of DNA experiences.

Neutral selection (Ka/Ks equal to 1)
Positive selection (Ka/Ks more than 1)
Purifying selection (Ka/Ks less than 1)

Purifying or negative selection

```
Seq1   AAG  ACT  GCC  GGG  CGT  ATT
Seq2   AAA  ACA  GCA  GGA  CGA  ATC
```

DNA

```
Seq1    K    T    A    G    R    I
Seq2    K    T    A    G    R    I
```

Protein

 # Synonymous substitutions = 6
 # Non-synonymous substitutions = 0

Ka / Ks
= Non-synonymous / Synonymous substitutions
= 0

Neutral Selection

Seq1	AAG	ACT	GCC	GGG	CGT	ATT
Seq2	AAA	ACA	GAC	GGA	CAT	ATG

DNA

Seq1	K	T	A	G	R	I
Seq2	K	T	D	G	H	M

Protein

 # Synonymous substitutions = 3
 # Non-synonymous substitutions = 3

Ka / Ks
= Non-synonymous/Synonymous substitutions
= 1

Positive Selection

Seq1	AAG	ACT	GCC	GGG	CGT	ATT
Seq2	AAA	ATT	GAC	GAG	CAT	ATG

DNA

Seq1	K	T	A	G	R	I
Seq2	K	I	D	E	H	M

Protein

 # Synonymous substitutions = 1
 # Non-synonymous substitutions = 5

Ka / Ks
= Non-synonymous/Synonymous substitutions
=5

Synonymous substitutions are NOT always neutral

Different codons for the same amino acid may have different functional constraints and fitness effects

- Translational efficiency: codon usage bias
- RNA stability and correct folding of secondary structures
- RNA editing
- Protein folding
- Exon splicing regulatory motifs
- Binding sites for microRNA and RNA binding proteins (RBP)

Neutral theory of evolution

- Using sequence data of hemoglobin, insulin, cytochrome *c* from many vertebrates, Motoo Kimura calculated on average sequence evolution in mammals had been very rapid: 1 amino acid change every 1.8 years

- Such a high mutation frequency suggest the majority of substitutions have no fitness effects, i.e. selectively neutral, and are created by genetic drift.

- Rate of molecular evolution is equal to the neutral mutation rate, this gives rise to the concept of "molecular clock"

Genes evolve at different rates

Rates of nucleotide substitution (per site per billion years)

Gene	Non-synonymous rate	Synonymous rate
Histone H4	0.00	3.94
Histone H2	0.00	4.52
Actin a	0.01	3.68
Ribosomal protein S14	0.02	2.16
Insulin	0.13	4.02
a-globin	0.78	2.58
Myoglobin	0.57	4.10

Molecular clock

Proteins undergo changes in their amino acid sequences over evolutionary time, as a result of the accumulation of nonsynonymous mutations in their encoding genes.

Genes and proteins act as “molecular clocks”, accumulating changes at a relatively constant rate, as mutations occur with a certain probability each time a nucleotide is replicated.

From the very beginning of molecular evolution studies, it became apparent that different proteins evolve at very different rates, each evolving according to its own “molecular clock”.

- Different proteins have different rates
- Different domains of the same protein may have different rate
- Same protein in different organisms may have different rates

Phylogenetic analysis

Phylogeny refers to the evolutionary history of species. Phylogenetics is the study of phylogenies—that is, the study of the evolutionary relationships of species. Phylogenetic analysis is the means of estimating the evolutionary relationships. In molecular phylogenetic analysis, the sequence of a common gene or protein can be used to assess the evolutionary relationship of species. The evolutionary relationship obtained from phylogenetic analysis is usually depicted as branching, treelike diagram—the phylogenetic tree.

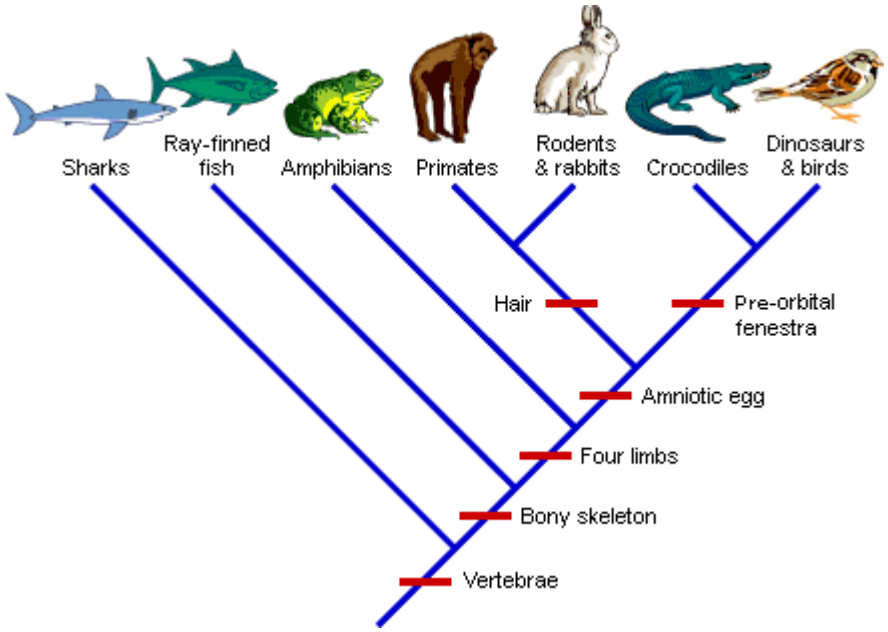
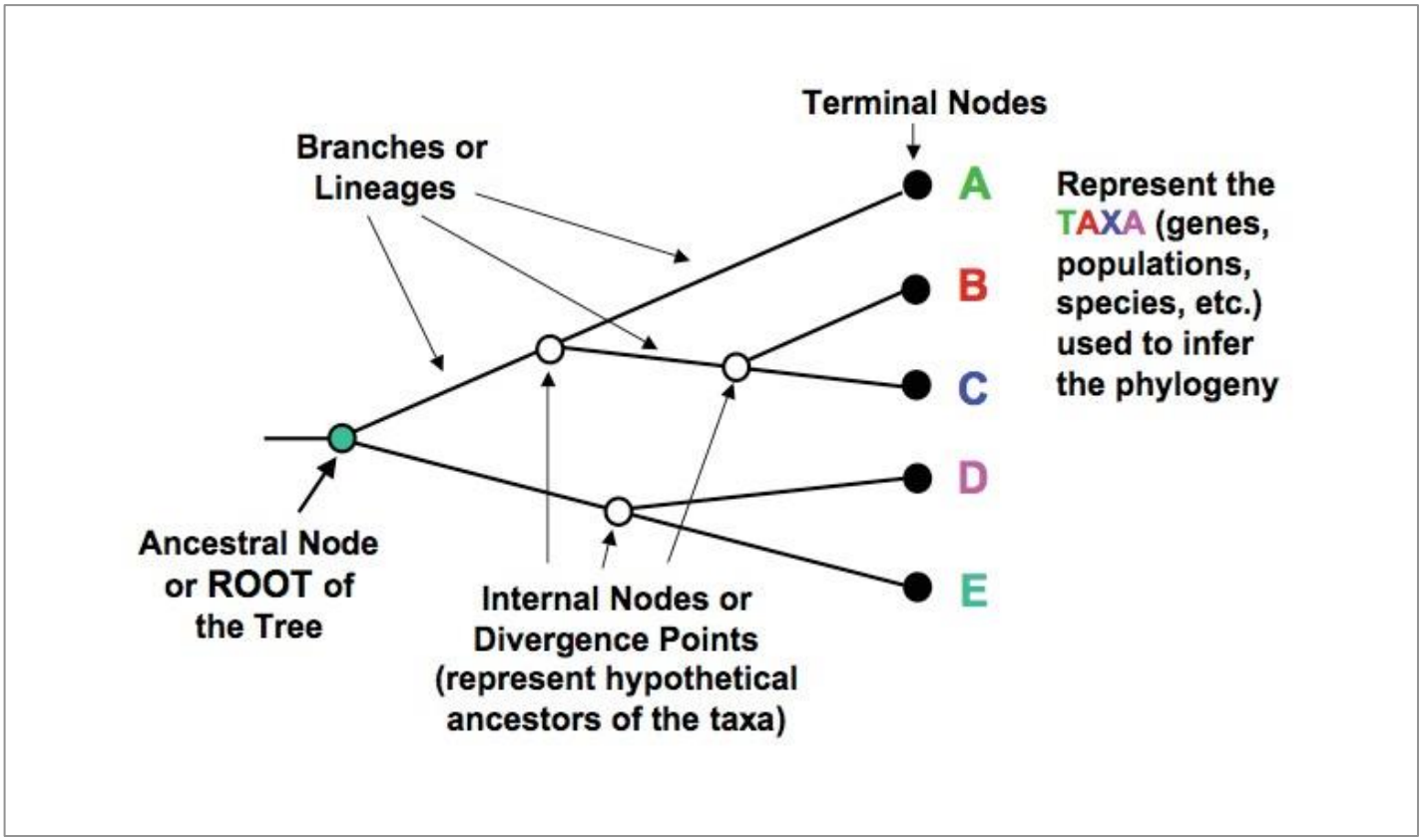


Figure: A phylogeny represents the evolutionary relationships among a set of organisms or groups of organisms, called taxa (singular: taxon) that are believed to have a common ancestor.

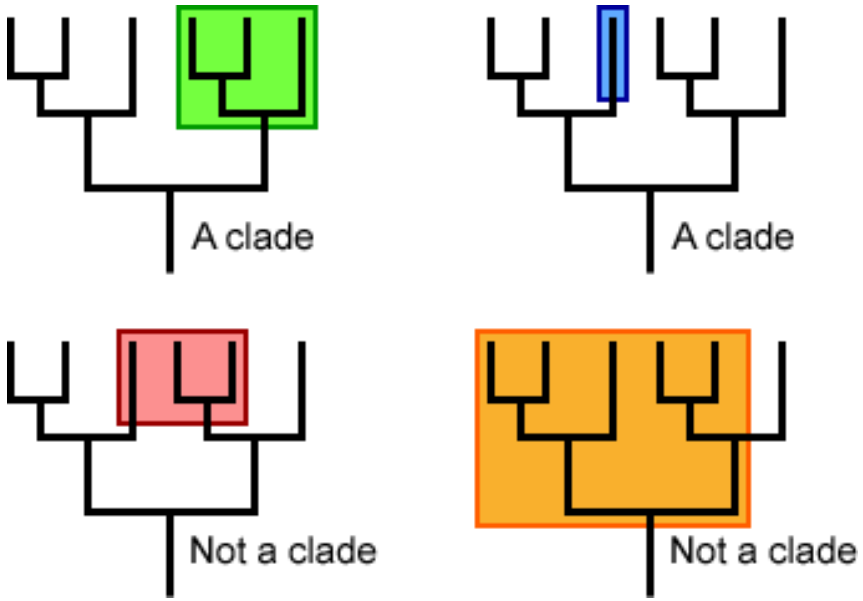
Phylogenetic Tree Terminology



In this tree, the vertical **branches** represent a **lineage**, which is a taxon, shown at the **tip**, and all its ancestors. The **nodes** are where lineages diverge, representing a speciation event from a common ancestor. Time in this particular style of tree is oldest in the left and the most recent in the right.

The trunk at the base of the tree is actually called the **root**, and the root node represents the **most recent common ancestor** of all of the taxa represented on the tree.

A group of taxa that includes a common ancestor and *all* of its descendants is called a **monophyletic group**, or a **clade**. Groups that *exclude one or more descendants* or that *exclude the common ancestor* are not monophyletic groups (clades); these groups are called paraphyletic and polyphyletic, respectively.



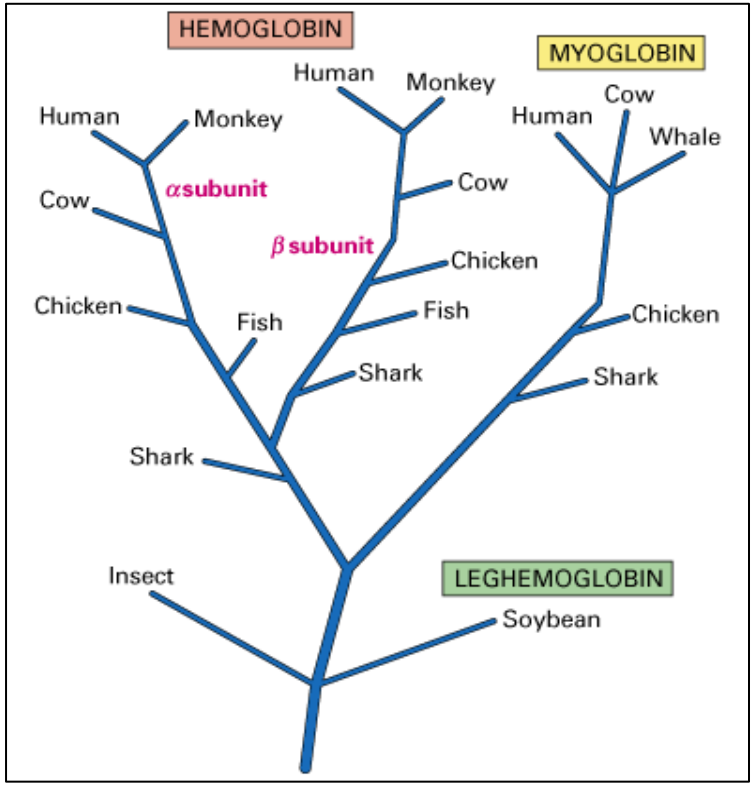
Two Areas in Phylogenetic analysis

- Phylogenetic inference or "**tree building**":
  - To infer the branching orders and lengths between "taxa" (or genes, populations, species etc).
  - For example, can DNA tell us giant panda more similar to bear or to dog, and when did they diverge ?

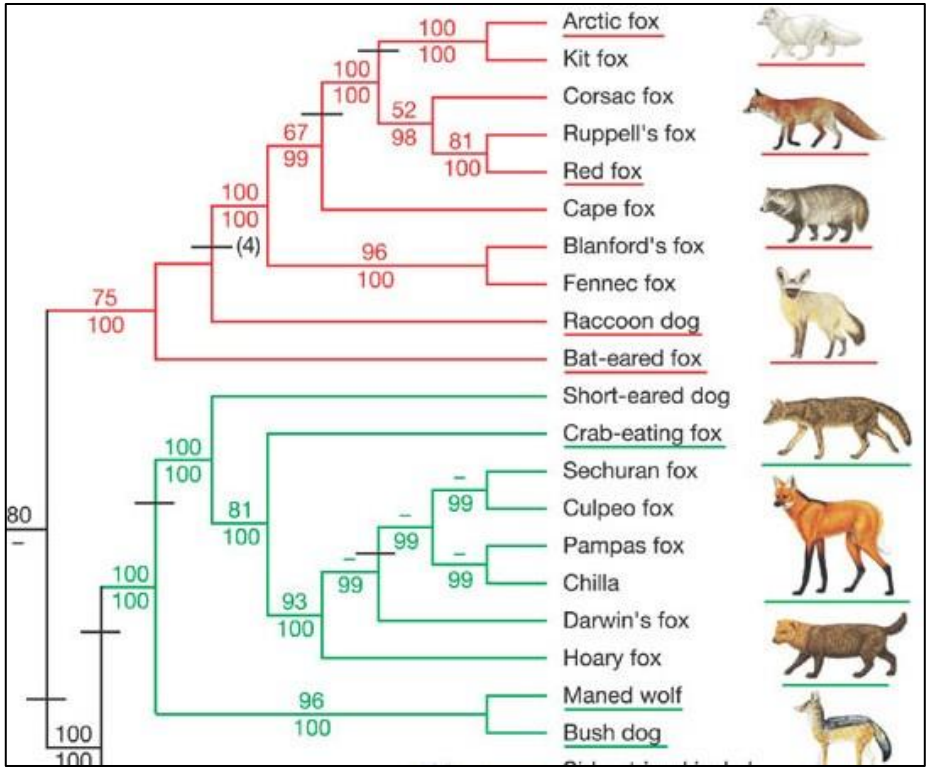
- **Character and rate** analysis:
  - Using phylogeny as a framework to understand the evolution of traits or genes.
  - For example, is gene X under positive or purifying selection ?

Phylogenetic Tree

Gene tree

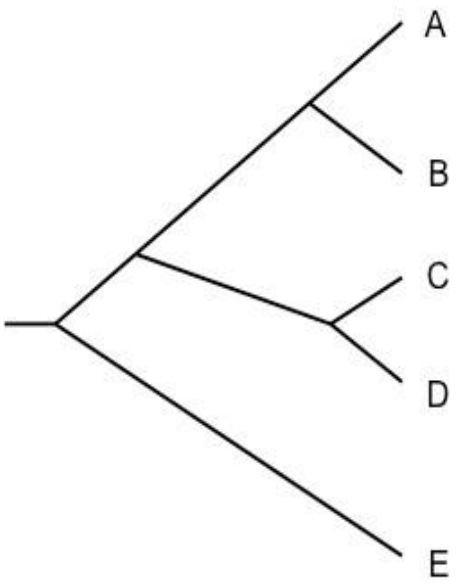


Species tree

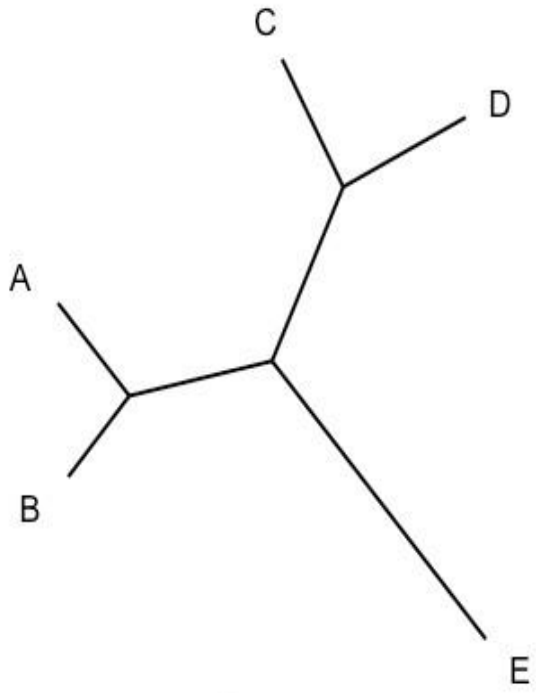


Gene/ species trees aim to represent the evolutionary history of gene families/ species, which evolved from a common ancestor.

Rooted and unrooted trees



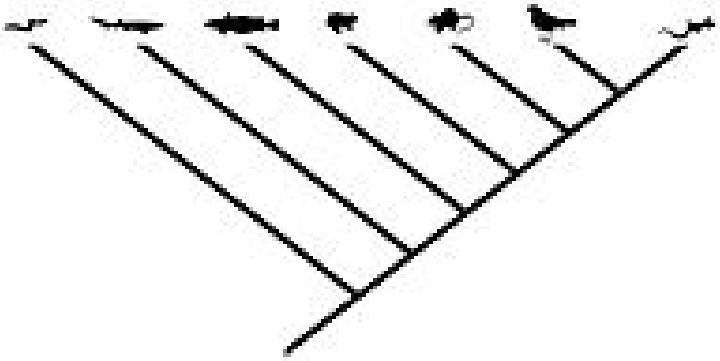
ROOTED

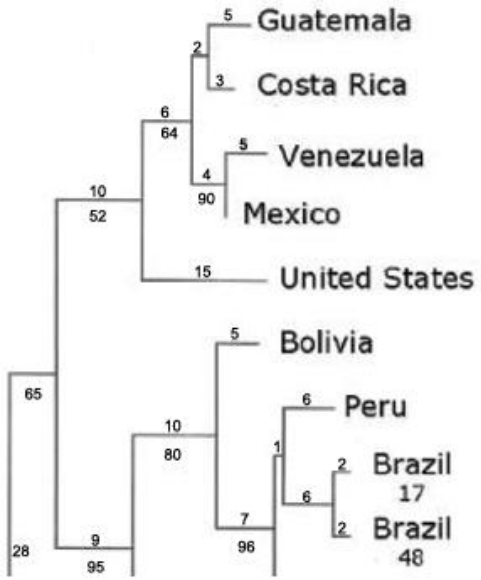


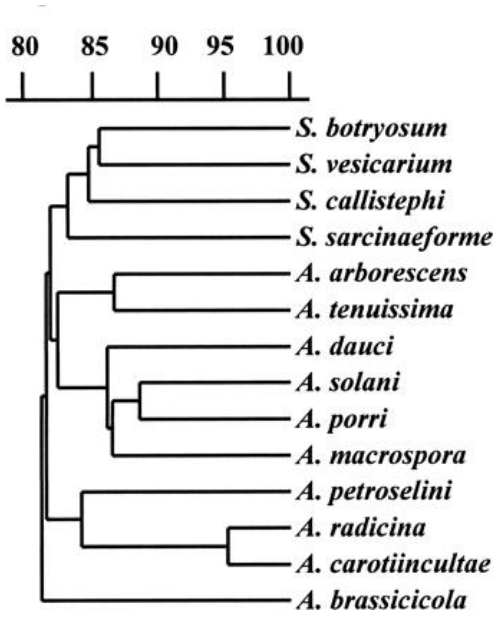
UNROOTED

Phylogenetic trees take several forms: Such as they can be *rooted* or *unrooted*. A *rooted* tree is a tree in which one of the nodes is stipulated to be the root, and thus the direction of ancestral relationships is determined. An *unrooted* tree, as could be imagined, has no pre-determined root and therefore induces no hierarchy. A tree can show edge lengths, indicating the genetic distance between the connected nodes.

Different types of trees







Cladogram

Most simple tree

Just shows relative recency

Branch length has no meaning

Phylogram

Additive tree

Branch length reflects number of changes

Dendrogram

Ultrametric tree

Special form of a phylogram

All tips equal length from root Axis = divergence time assuming molecular clock

Methods of Tree reconstruction

- Maximum Parsimony methods
- Distance based methods
- Maximum Likelihood methods
- Bayesian methods

Parsimony Methods

- Optimality criterion: The "most-parsimonious" tree is the one that requires the fewest number of evolutionary events (e.g. nucleotide substitutions, amino acid replacements) to explain the observed sequences.

- Advantages:
  - Intuitive, logical and simple (can be done with pencil-and paper)
  - Can be used on molecular and other (morphological, language) data.
  - Can be used to infer the sequences of extinct (hypothetical) ancestors

- Disadvantages
  - Can be fooled by high levels of homoplasy ("same events")
  - Can be problematic when the real tree is mixed with very short and long branches, e.g. long-branch attraction

Distance based methods

- Estimate the number of substitutions between each pair of sequences in a group of sequences.
- Try to build a tree so that the branch lengths represent the pair-distances.
- What are these “distances” ? Example: sequence identity between two protein and DNA sequences

Make trees from pair-wide distances

- **Neighboring joining**
  - Pair with the smallest branch lengths chosen to be joined
  - A new distance table is created with joint sequences entered as a composite.
  - Repeat process to select next pair to join.
  - Repeat process until correctly branched tree and distances identified

- UPGMA
  - Unweighted Pair Group Method with Arithmetic Mean

- **Maximum likelihood methods:**
  - ML methods evaluate phylogenetic hypothesis in terms of the **probability** that a proposed model and the parameters gave rise to the observed data. The tree found to have the highest likelihood is considered to be the optimal tree.

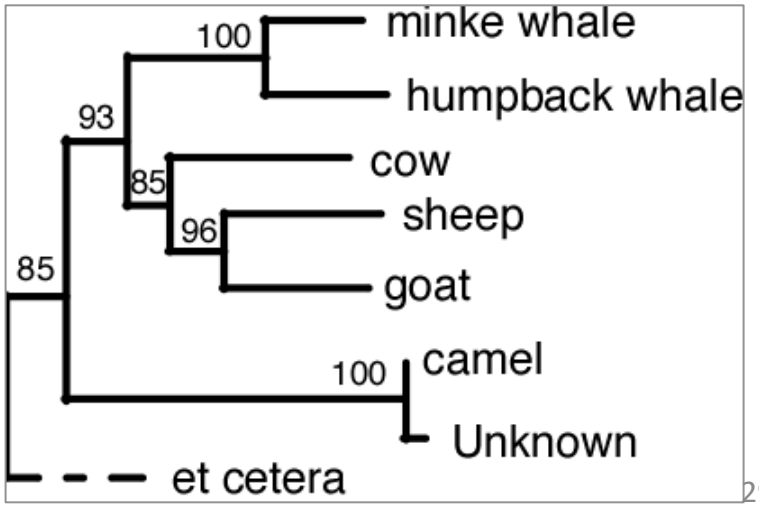
- **Bayesian Markov chain Monte Carlo methods**
  - Generate a large population of trees, then take a random walk through the "tree space" until a perfect tree is found.

How well supported is the tree?

Bootstrapping

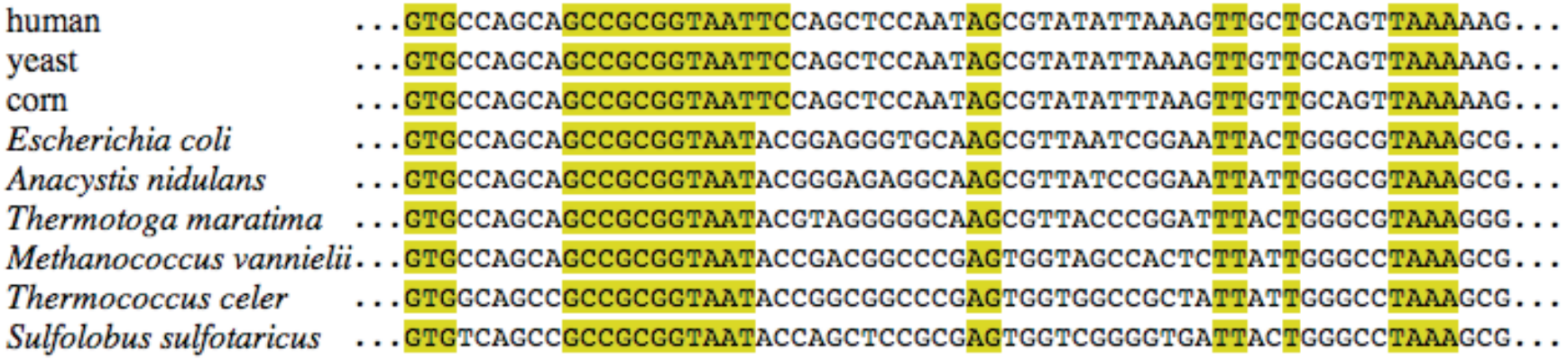
- How robust is the tree ? How much does the data support the tree ? How confident are we about a particular branch point ?

- To test this, we repeatedly re-sampled the data with the replacement and re-calculate the tree, and ask how many times do we still see the original tree or branch point.



Constructing organism phylogeny from specific genes

- The gene must be present in all organisms
- The gene cannot be subject to horizontal transfer
- The gene must display an appropriate level of sequence conservation for the divergences of interest, i.e. evolving not too fast and not too slow.
- The gene must be sufficiently large to carry a record of the historical information.



16s rRNA

Application of phylogenetics

1. Classification: Phylogenetics based on sequence data provides us with more accurate descriptions of patterns of relatedness than was available before the advent of molecular sequencing. Phylogenetics now informs the Linnaean classification of new species. Most importantly, trees provide an efficient structure for organizing knowledge of biodiversity and allow one to develop an accurate, nonprogressive conception of the totality of evolutionary history.

2. Identifying the origin of pathogens: Molecular sequencing technologies and phylogenetic approaches can be used to learn more about a new pathogen outbreak. This includes finding out about which species the pathogen is related to and subsequently the likely source of transmission. This can lead to new recommendations for public health policy.

3. Conservation: Phylogenetics can help to inform conservation policy when conservation biologists have to make tough decisions about which species they try to prevent from becoming extinct.

4. Cancer biology: In the last few years, developments in sequencing technology have allowed us to sequence individual cells. Applying single-cell sequencing technology to tumours revealed that the cells within a tumour were not only very phenotypically diverse, but they were also very genetically diverse. In this way, tumours can be thought of in a similar way to a communal population of individual organisms, and as such, the tumour evolves over time.

References

Evolution by Douglas J. Futuyma

Molecular Evolution: A phylogenetic approach by Roderic D. M. Page

Fundamentals of molecular evolution Book by Dan Graur

End