Phylogenetics

IMBB 2016

BecA-ILRI Hub, Nairobi

May 9 – 20, 2016

Joyce Nzioki

Phylogenetics

The study of evolutionary relatedness of organisms.

Derived from two Greek words:

- Phle/Phylon: Tribe/Race
- Genetikos: Relative to Birth

Phylogenetics

- Evolution is the change in distribution of allele frequencies from one generation to the next.
- Similarity in sequenced data is taken as an indication of evolutionary relatedness. Sequence difference is taken as a measure of evolutionary divergence.
- Progression rules: as an organism is more distant from its ancestor their characters are more evolved.

Phylogenetics

- Phylogeny can be drawn from molecular data (DNA/RNA/Proteins) or morphological data

- Taxonomy is informed from phylogenetics

- Phylogenetic trees are graphical representation of the evolutionary relationships through time or genetic distance.

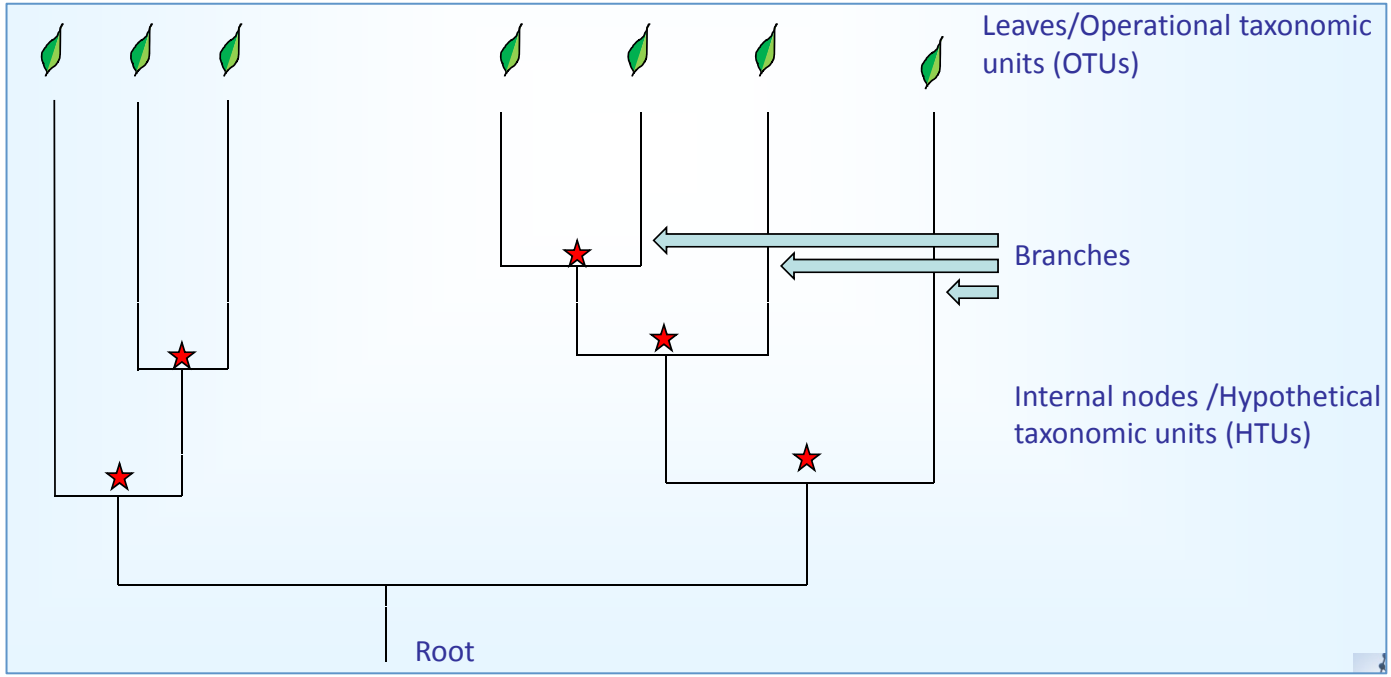
Aspects of phylogeny

We will look at the major aspects of phylogenies and how they can be interpreted.

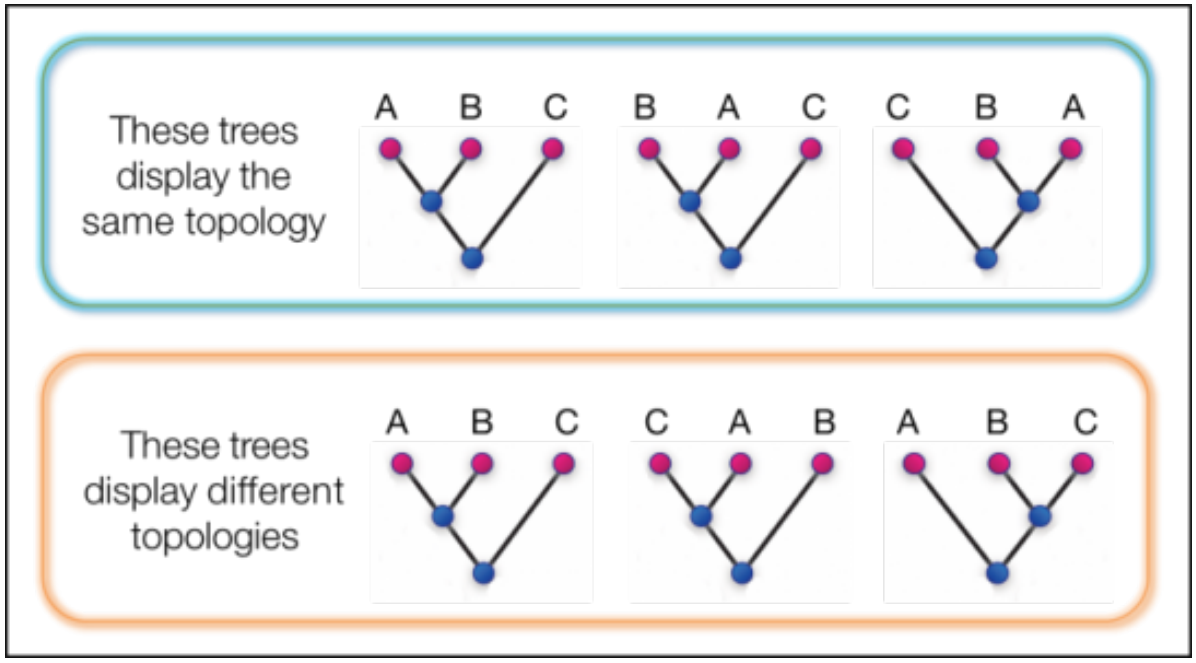
1. Topology
2. Branches
3. Nodes
4. Confidence

Phylogenetic tree

This is a branching diagram that infers evolutionary relationship of various species based on their physical or genetic traits.

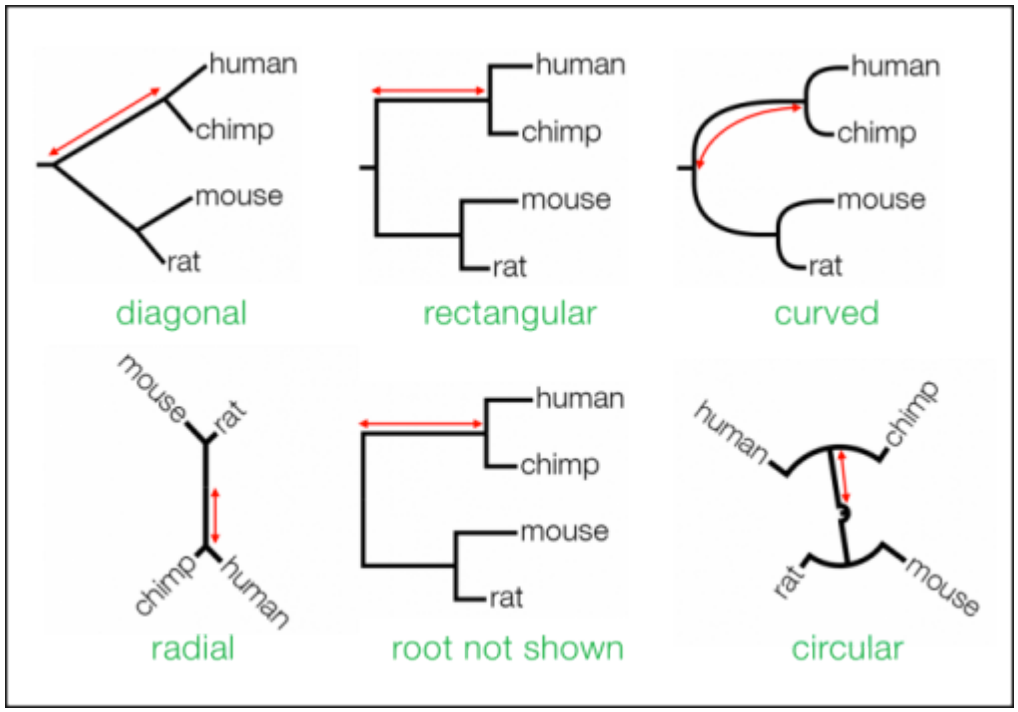


Topology – branching structure of a tree



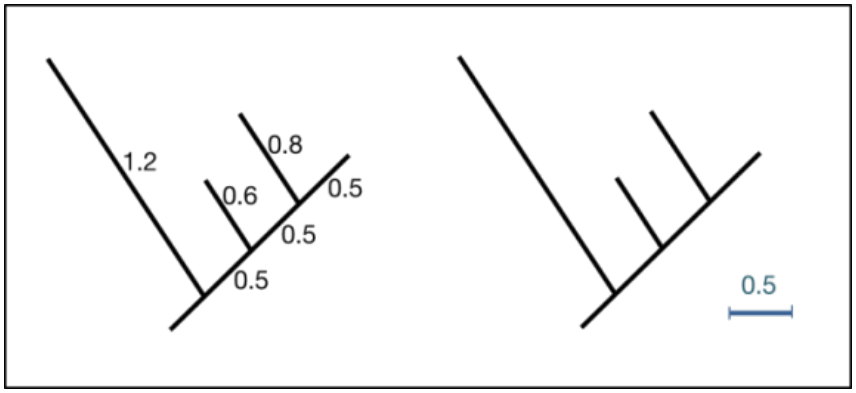
- The three top trees have common topologies. “A and B are more closely to one another than to C”.
- The bottom box had different relationships

Alternative representative of phylogenies



- The same topology can be drawn in different ways the common format are shown above.
- Diagonal and rectangular formats commonly used for publication
- Curved format used in review papers to represent summaries of phylogenies
- Radial format is used to show un-rooted trees
- Circular format is used to represent large phylogenies.

Branches

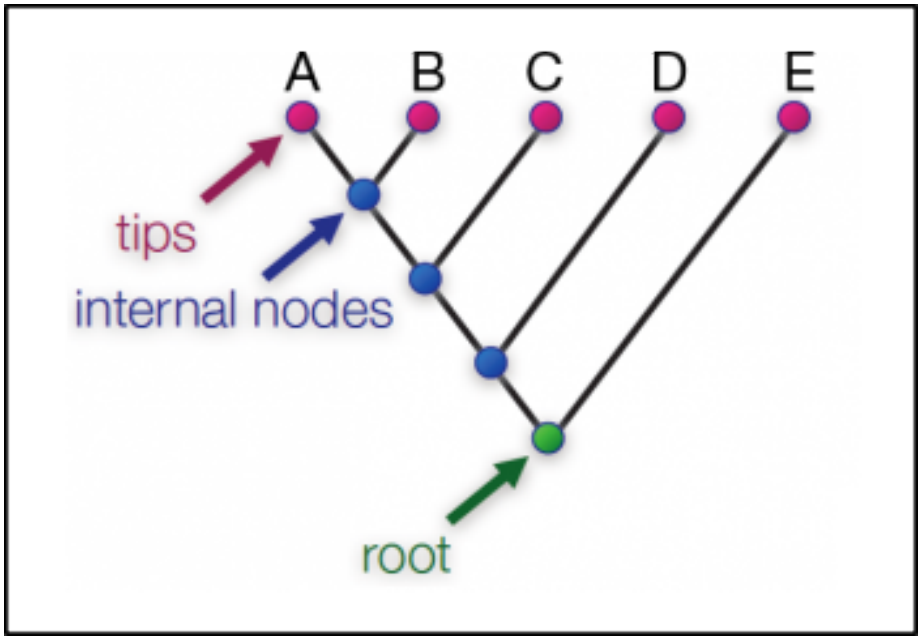


Human	ATGTTGACTC
Mouse	ATGCTGACTC

- Branches show the path of transmission of genetic information from one generation to the next.
- Branch lengths indicate genetic change i.e. the longer the branch, the more genetic change has occurred.
- Informative branch lengths are drawn to scale and indicate the number of substitutions per site

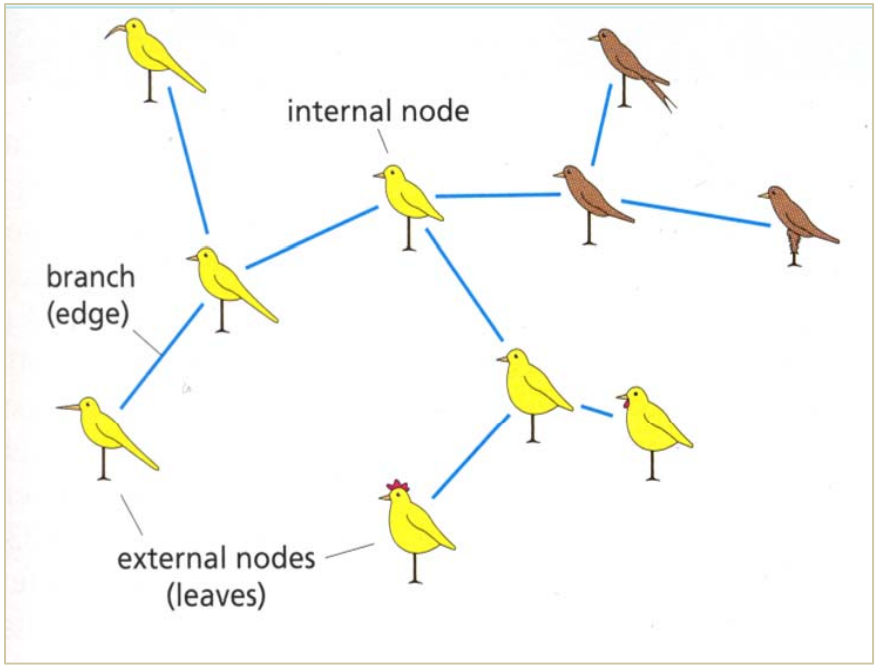
- A naive method of estimating the genetic change is counting the number of differences and dividing by the sequence length
- i.e. (1/10 = 0.1) though this doesn't account for multiple substitutions.
- To overcome this issues we use evolutionary models to infer genetic changes that occur

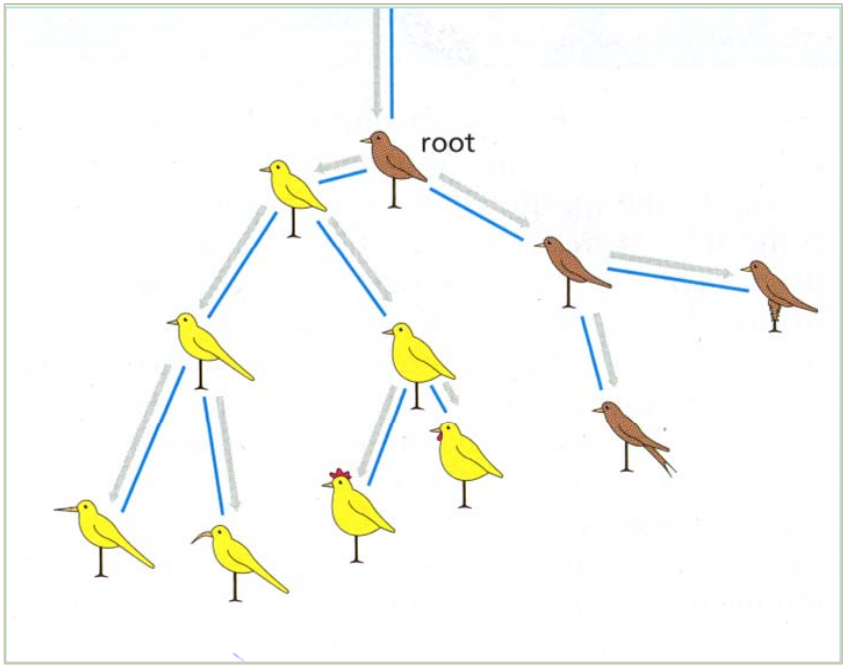
Nodes



- Tips /OTUs /external nodes – the sequences sampled for phylogeny.
- Internal nodes – occur at the point where more that one branch meet and represent the ancestral sequence / hypothetical common ancestor.
- The root – represents the most recent common ancestor.

Rooted vs Un-rooted tree





Un-rooted tree
Does not show direction of evolution / ancestry

Rooted tree
Direction of evolution indicated as moving away from the root

Rooting a tree

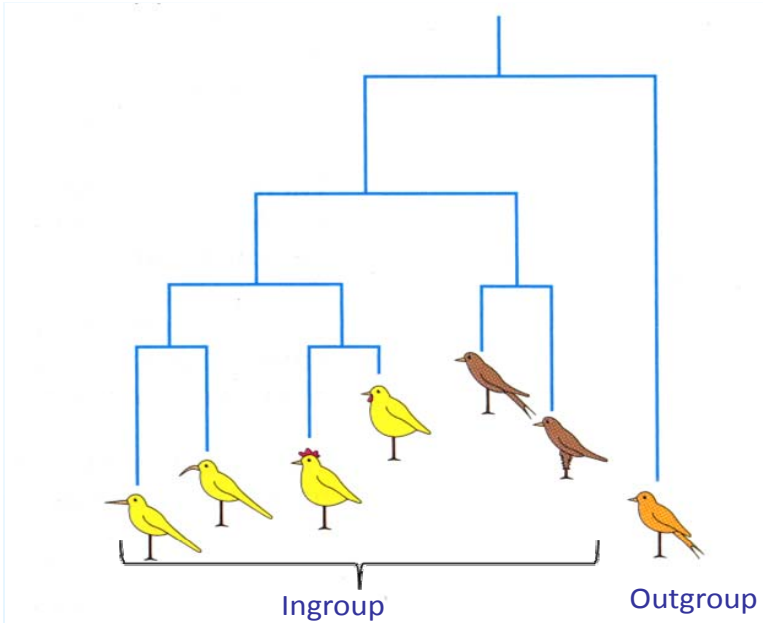
Two methods are known for tree routing:

1. **Outgroup Criteria**: include in the analysis a group of sequences known as *a priori* to be external to the group in study; the root is by necessity the branch joining the outgroup and the other sequences

2. **Midpoint rooting**: this method makes the assumption that all lineages are supported to have evolved with the same rate since divergence from their common ancestor. The root is at the equidistant point from all tree leaves .

Rooting a tree with an outgroup

- This is the use of an organism or group of organisms (outgroup) that are more evolutionary distant to the group in study (internal group).
- The common ancestor is therefore placed between the internal group and the outgroup. This effectively roots the tree and evolutionary distances will be relative to this point. (gives a direction of evolution)



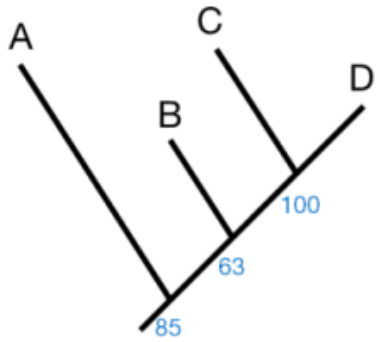
Selecting an outgroup

- An outgroup should not be too distantly related to the internal group, this results in very long branch lengths that distort the remaining branches rendering the topology unreliable.

- The outgroup should also not be too closely related to the internal group this may not make a true outgroup.

- Using various outgroup species may better balance the final tree branching.

Confidence



- Inferring phylogenies is inherently an uncertain process.
- Approaches used to estimate our confidence in the inferred tree topology: bootstraps, likelihood and Bayesian approaches.
- This course we will focus on bootstrap.
- Typically these are shown on the phylogeny. Indicating the percentage confidence of a branch.
- Interpreting the exact meaning of confidence values is still an area of debate but generally >50% is acceptable provided an appropriate evolutionary model was used.

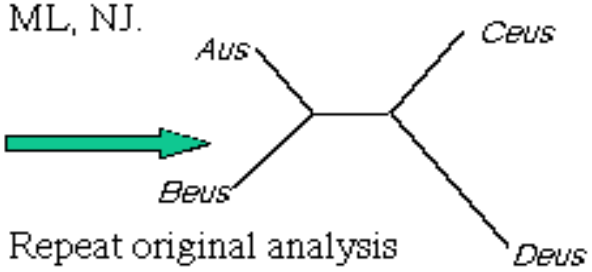
Bootstrapping

- Bootstrapping is commonly used test of reliability of inferred phylogenetic tree.
- A single tree may not be credible given the dependencies involved: (characters, evolutionary model, parameters).
- Bootstrapping is done by generating 100-1000 replicas of your data (arrange character positions at random, to create a series of bootstrap samples of same size as original data)
- The bootstrap datasets are analyzed looking for consistency. Variation among the datasets is used to estimate error involved in making estimates in the original data

Original data set with *n* characters.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Aus	C	G	A	C	G	G	T	G	G	T	C	T	A	T	A	C	A	C	G	A
Beus	C	G	G	C	G	G	T	G	A	T	C	T	A	T	G	C	A	C	G	G
Ceus	T	G	G	C	G	G	C	G	T	C	T	C	A	T	A	C	A	A	T	A
Deus	T	A	A	C	G	A	T	G	A	C	C	C	G	A	C	T	A	T	T	G

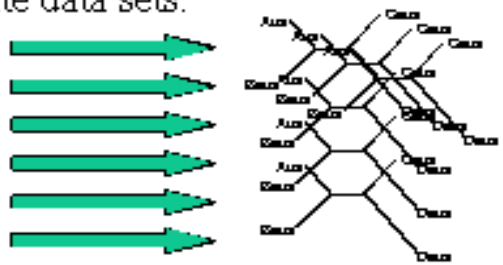
Original analysis, e.g. MP, ML, NJ.



Draw *n* characters randomly with replacement. Repeat *m* times.

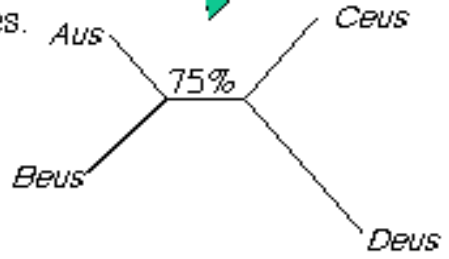
	2	3	13	8	3	19	14	6	20	20	7	1	9	11	17	10	6	14	8	16
Aus	G	A	A	G	A	G	T	G	A	A	T	C	G	C	A	T	G	T	G	C
Beus	G	G	A	G	G	G	T	G	G	G	T	C	A	C	A	T	G	T	G	C
Ceus	G	G	A	G	G	T	T	G	A	A	C	T	T	T	A	C	G	T	G	C
Deus	A	A	G	G	A	T	A	A	G	G	T	T	A	C	A	C	A	A	G	T

Repeat original analysis on *each* of the pseudo-replicate data sets.

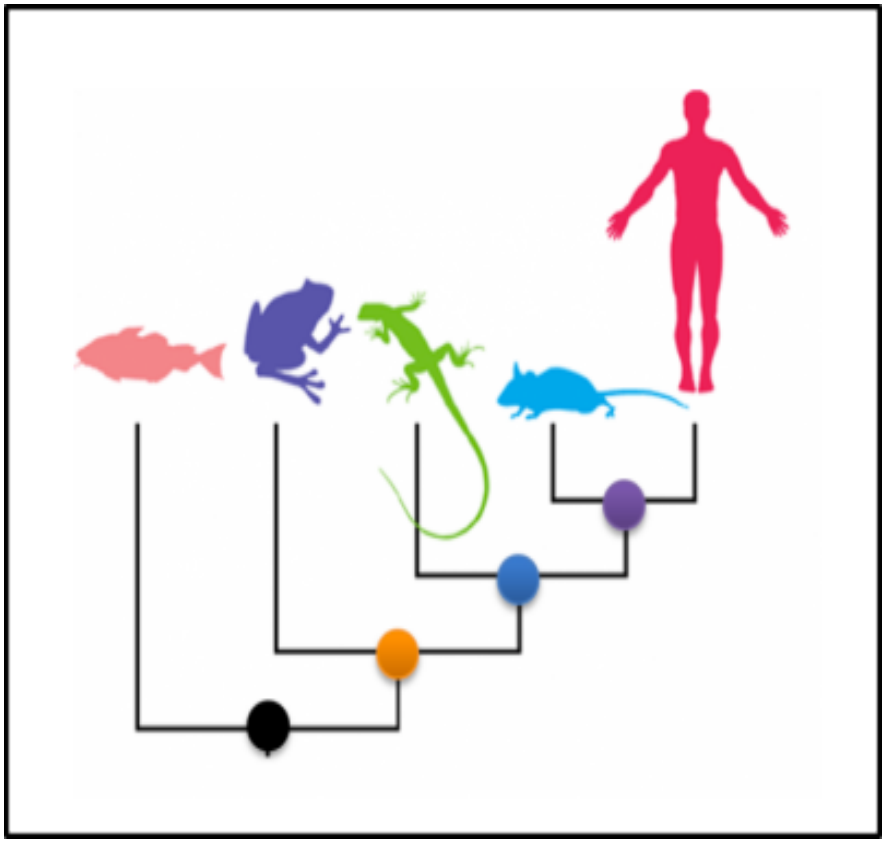


m pseudo-replicates, each with *n* characters.

Evaluate the results from the *m* analyses.

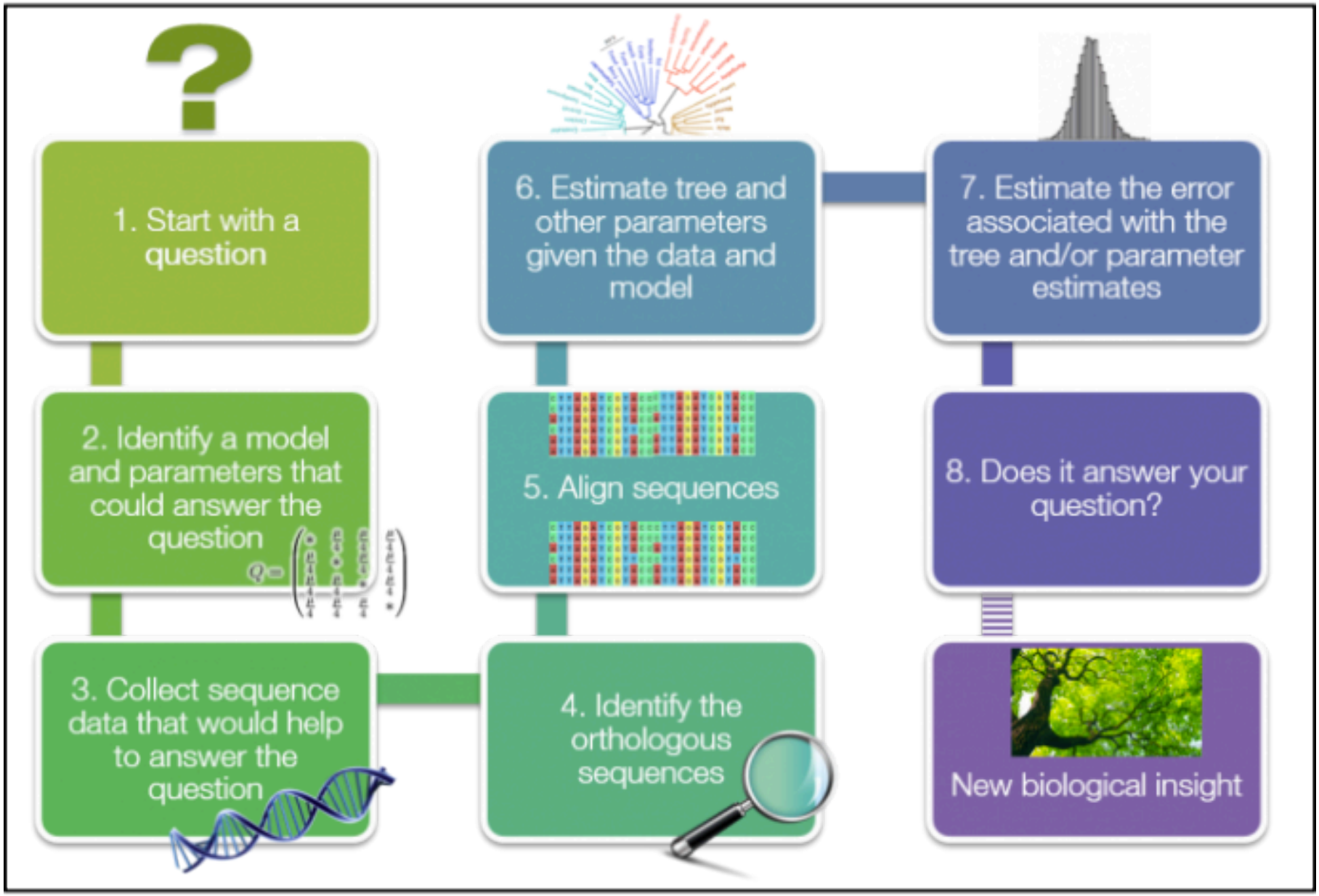


Interpreting patterns of relatedness



- Humans are more closely related to mice than lizard (share a more recent common ancestor).
- Frogs are more closely related to lizards than they are to fish (as they share a common ancestor with lizard more recent than fish).
- Fish are equally related to mice as they are to frogs. Mice and frogs both share the same common ancestor (black spot) with fish.
- If you rotate the branches to change topology, the biological meaning remains.

Major stages in phylogenetic analysis

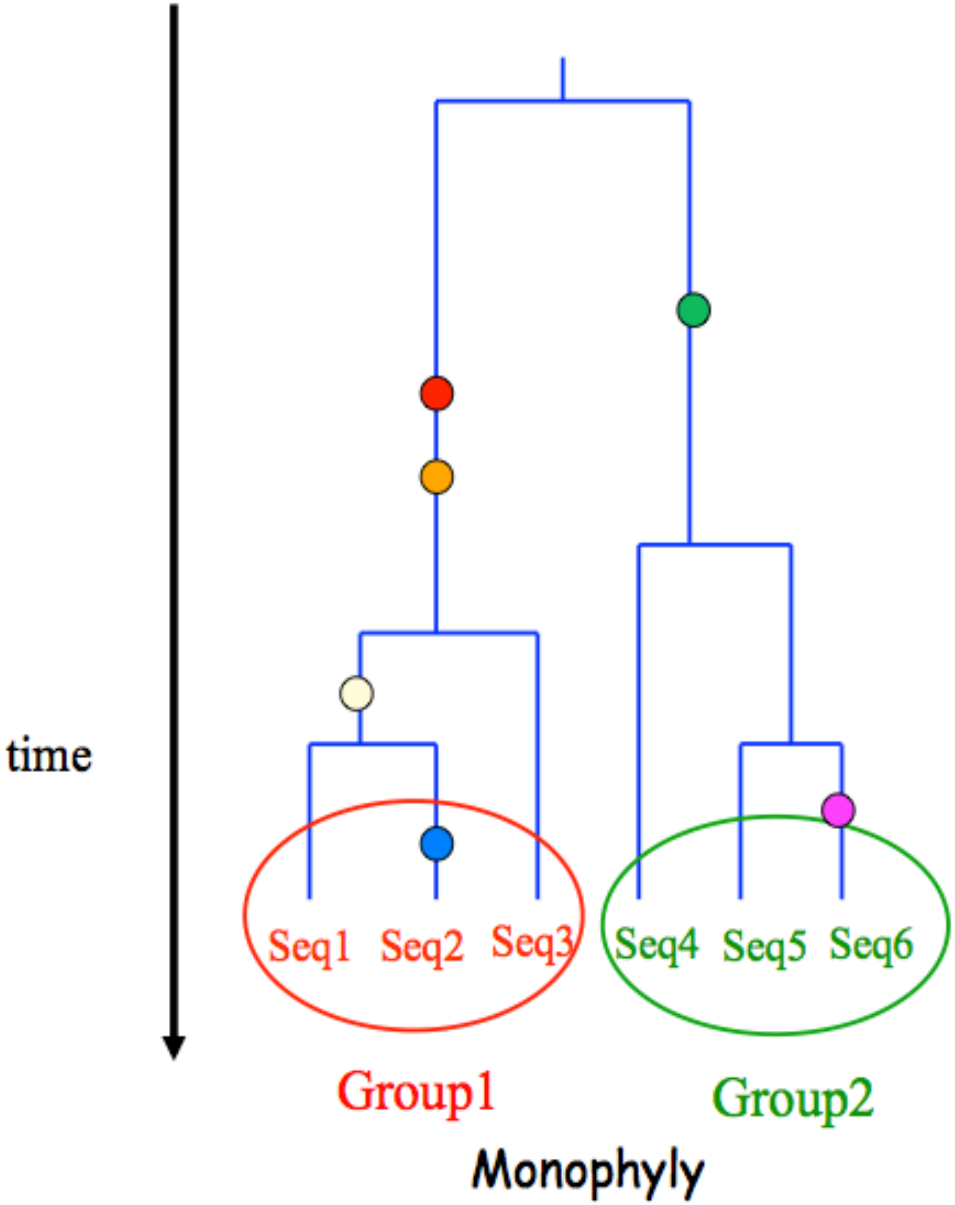


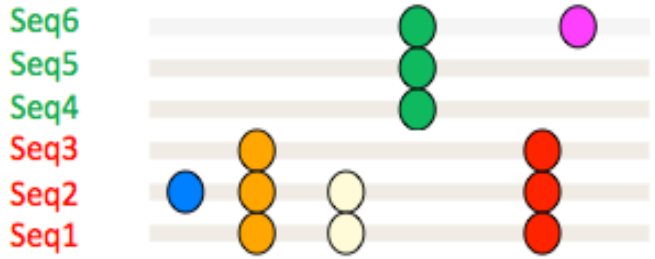
Building a phylogenetic tree

- Staring point :a set of homologous aligned DNA or protein sequences
- Quality of the alignment is essential, unreliable parts of the alignment are omitted
- Most methods only take into account substitutions gaps(deletions/insertions) are not used.

```
Xenopus     ATGCATGGGCCAACATGACCAGGAGTTGGTGTCggtCCAAACAGCGTT---GGCTCTCTA
Gallus      ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC---CAAAATAACACCaacATGCAAATG
Bos         ATGCATCCGCCACCATGACCAGCAGGAGGTAGCagtCAAAACAGCACCaacGTGCAAATG
Homo        ATGCATCCGCCACCATGACCAGCAGGAGGTAGCagtCAAAACAGCACCaacGTGCAAATG
Mus         ATGCATCCGCCACCATGACCAGCAGGAGGTAGCactCAAAACAGCACCaacGTGCAAATG
Rattus      ATGCATCCGCCACCATGACCAGCGGGAGGTAGCtctCAAAACAGCACCaacGTGCAAATG
```

Sequences Reflect Relationships





... =derived mutations

The gene compared must evolve at a rate comparable to the divergence time of the organism; for example:

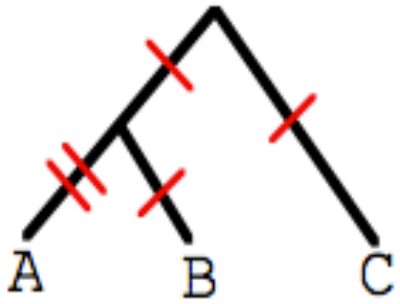
- 18S rRNA gene for phylum-level divergences since it evolved slowly
- Hemoglobin genes for mammalian orders.
- Mitochondrial DNA for species divergences within a genus
- Repetitive DNA sequences (e.g. microsatellites) for individuals within a species

Method of building a tree

1. Distance methods
2. Character based methods
    1. Maximum parsimony
    2. Maximum likelihood
3. Bayesian inference

Distance methods

- Start from a multiple sequence alignment
- Make a matrix of pairwise distances
- Build a phylogenetic tree.
- E.g. Neighbor joining and UPGMA trees



tree

	A	B	C
A	**0**		
B	**3**	**0**	
C	**4**	**3**	**0**

distance matrix

Distance methods

- To estimate evolutionary distances between sequences there is need for statistical / evolutionary models.
- Statistical models estimate for evolutionary distance while accounting for residue substitution and homoplasy.
  - Juke-Cantors: good for distances <10%
  - Kimura-2: distance 10-30% and transitions ~= transversions
  - Tamura: distances 10-30% and strong G+C bias
  - Jin-Nei γ: distance 10-30% and varying transition-transvertion rates
  - Tajima-Nei: distances 30-100%
- These evolutionary distances are then converted into a distance matrix used in building the tree

Character based methods

- This analyses any set of discrete character, that is each position in an aligned sequence character.
- All character can be analyzed separately and independently of one another.

- These include:
  1. Maximum Parsimony (MP)
  2. Maximum Likelihood (ML)
  3. Bayesian methods

Maximum Parsimony

- Parsimony involves evaluating all possible trees and giving each a score based on the number of evolutionary changes that are needed to explain the observed data.

- The best tree is the one that requires the fewest base changes for all sequences to derive from a common ancestor

Maximum Likelihood

- Maximum Likelihood evaluates the topologies of different trees given a particular evolution model and picks the best one according to the likelihood score. (tree with the highest likelihood)
- It considers all characters and looks for trees that best suit a given evolution model.
- It is possibly more accurate than Maximum parsimony if the appropriate model is chosen.

summary

- UPGMA assumes molecular clock, so provides a rooted tree (this assumption may be too strong in some cases)
- Neighbor joining has been proved to create correct trees when evolutionary rates vary.
- Maximum Parsimony is good for closely related sequences
- Maximum likelihood methods is the general of all three.

Applications of Phylogenetics

1. **Classification** – phylogenetics gives accurate patterns of relatedness that now inform the Linnaen classification of new species.
2. **Forensics** - Access DNA evidence presented in court.
3. **Origin of pathogens** – learn about new pathogen outbreak and source of transmission.
4. **Conservation** - can inform conservation policies of species becoming extinct.

WWW resources for molecular phylogeny (3)

- Sequence alignment editor

⇨ SEAVIEW : for windows and unix
http://pbil.univ-lyon1.fr/software/seaview.html

- Programs for molecular phylogeny

⇨ PHYLIP : an extensive package of programs for all platforms
http://evolution.genetics.washington.edu/phylip.html

⇨ CLUSTALX : beyond alignment, it also performs NJ

⇨ PAUP* : a very performing commercial package
http://paup.csit.fsu.edu/index.html

⇨ PHYLO_WIN : a graphical interface, for unix only
http://pbil.univ-lyon1.fr/software/phylowin.html

⇨ MrBayes : Bayesian phylogenetic analysis http://
morphbank.ebc.uu.se/mrbayes/

⇨ PHYML : fast maximum likelihood tree building http://www.lirmm.fr/
~guindon/phyml.html

⇨ WWW-interface at Institut Pasteur, Paris
http://bioweb.pasteur.fr/seqanal/phylogeny

Thanks

- Thanks to Anne Jores and EBI online training for the some of the slides presented