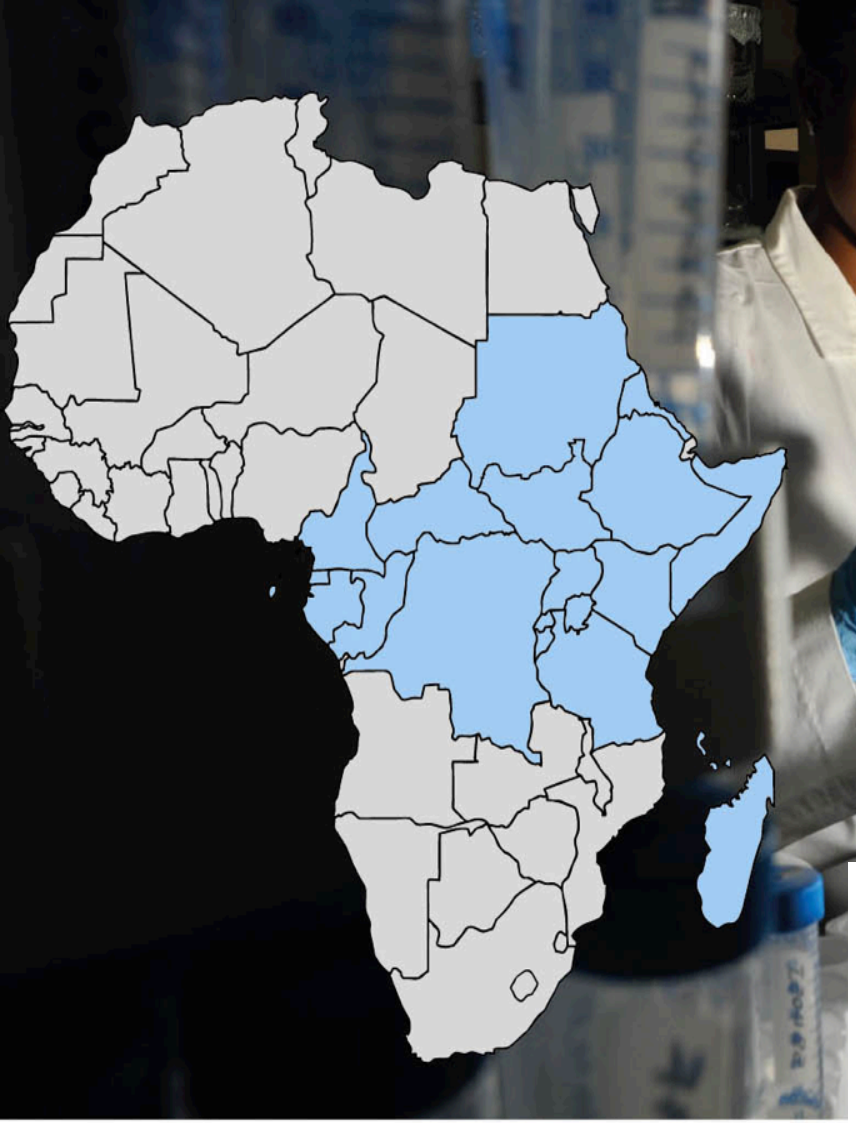


Multiple Sequence Alignments and Phylogenetics

NARL, Kawanda

10 - 14, November, 2014

Joyce Njoki Nzioki
BecA-ILRI Hub, Nairobi, Kenya
http://hub.africabiosciences.org/
http://www.Ilri.org/
j.n.njuguna@cgiar.org

ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

African Union

NEPAD
A PROGRAMME OF THE AFRICAN UNION

biosciences
eastern and central africa

Multiple sequence alignment

- Sequence alignment as earlier discussed is the initial point in functional and structural protein characterization.
- Multiple sequence alignment (MSA) is the alignment of more than two sequences.
- MSA helps to reveal ancestral relationships between organisms and conserved residues and motifs of functional importance.
- Some MSA programs include
  - T-Coffee
  - Muscle
  - Mafft
  - Clustal
  - Promals3D

Phylogenetics

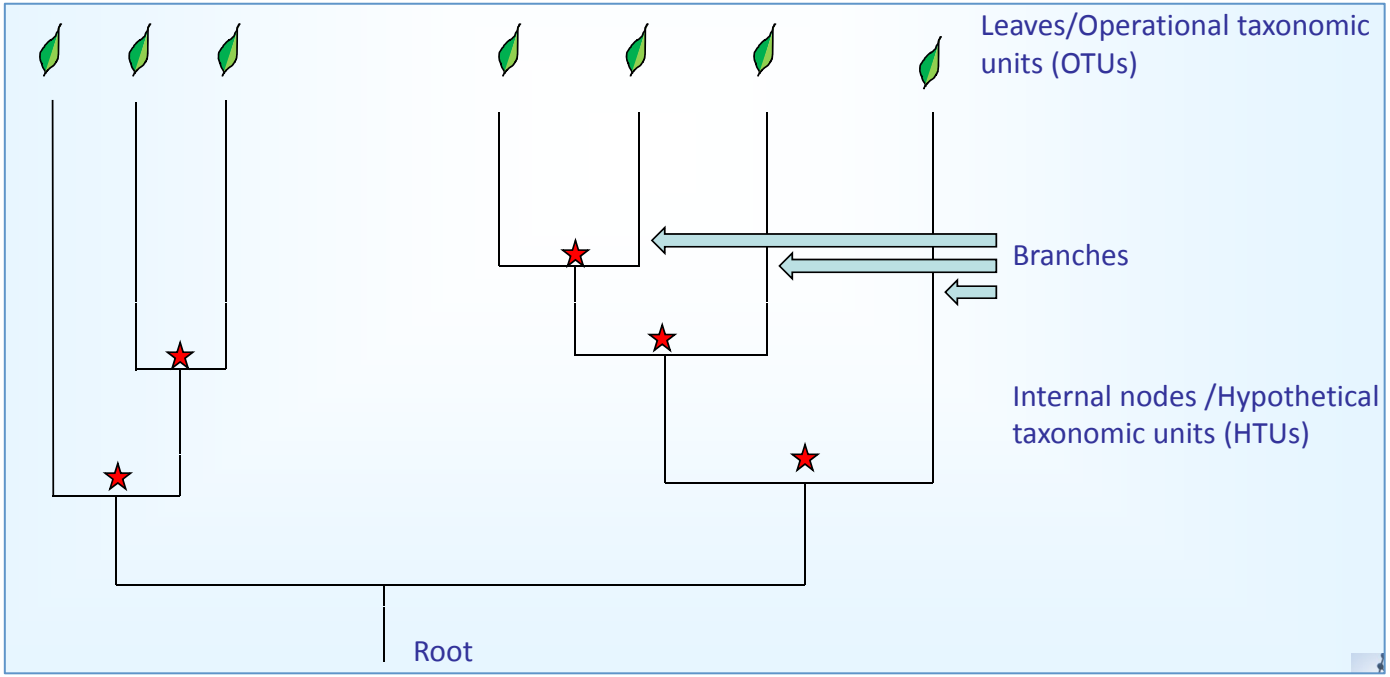
- The study of evolutionary relatedness of organisms. Derived from two Greek words:
  - Phle/Phylon: Tribe/Race
  - Genetikos: Relative to Birth

Phylogenetics

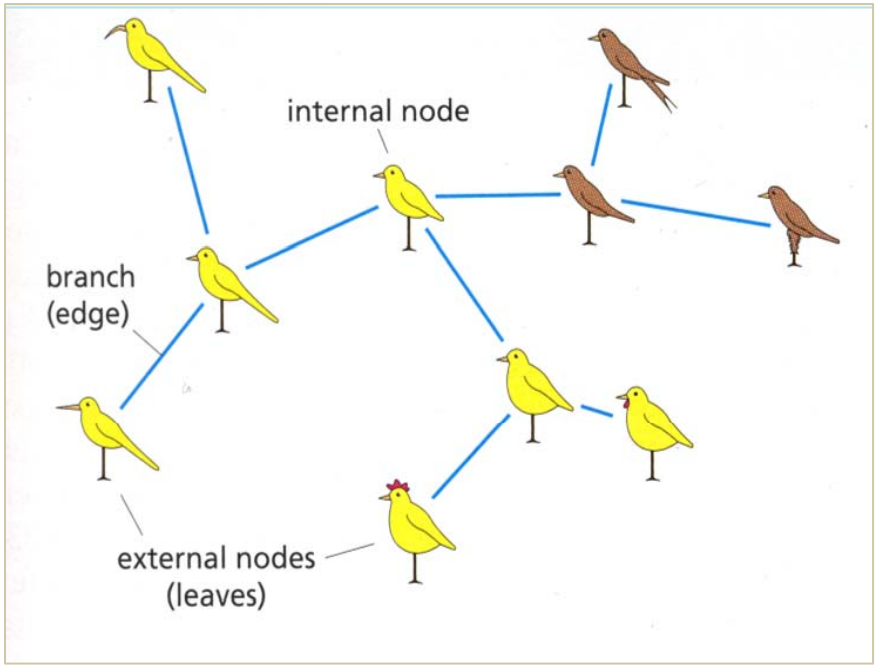
- Evolution is the change in distribution of allele frequencies from one generation to the next.
- Similarity in sequenced data is taken as an indication of evolutionary relatedness. Sequence difference is taken as a measure of evolutionary divergence.
- Progression rules: as an organism is more distant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

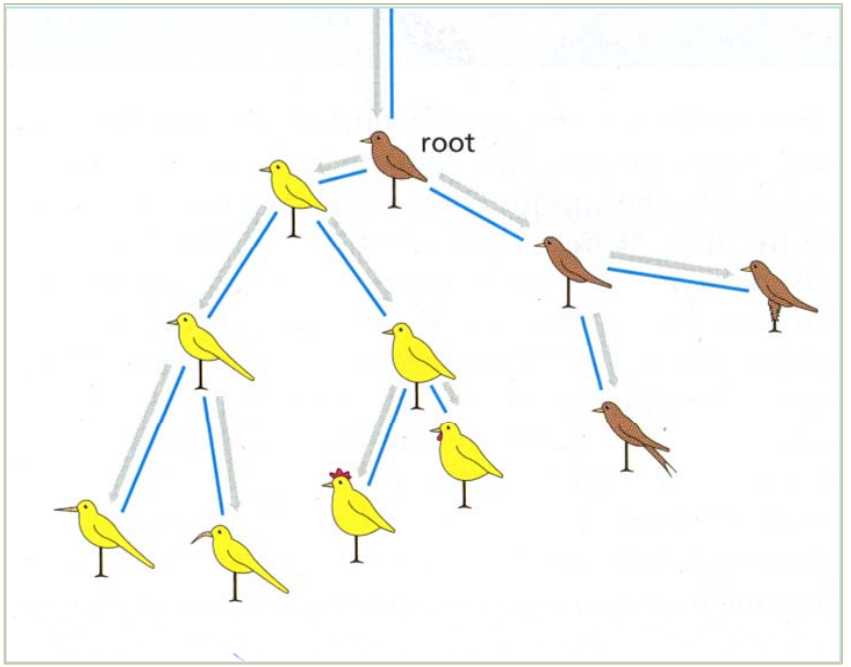
Phylogenetic tree

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



Rooted vs Un-rooted tree





Un-rooted tree
Does not show direction of evolution

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

Rooting a tree

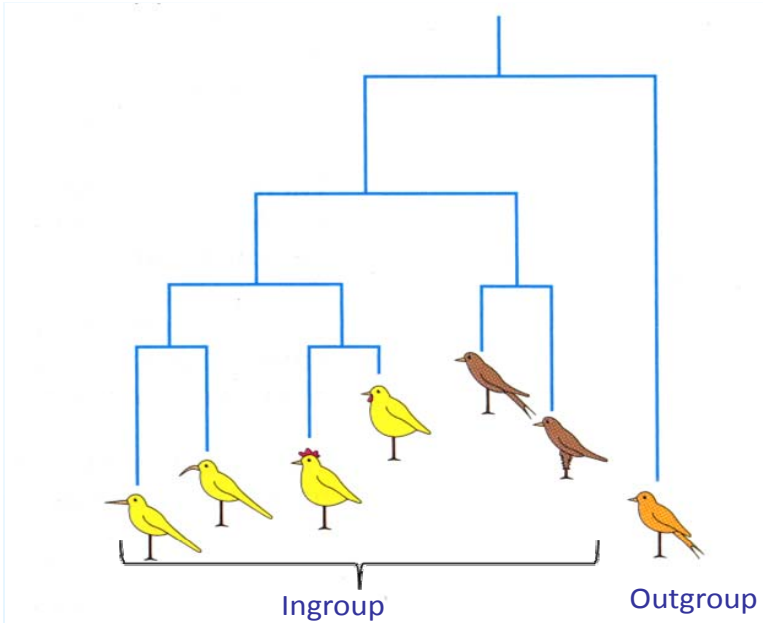
- Two methods are known for tree routing:

1. **Outgroup Criteria**: include in the analysis a group od 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. **Molecular clock**: all lineages are supported to have evolved with the same speed 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.

Introduction to Molecular Phylogeny

- Starting point: a set of homologous, aligned DNA or protein sequences
- Result of the process: a tree describing evolutionary relationships between studied sequences
  = a genealogy of sequences
  = a phylogenetic tree

```
CLUSTAL W (1.74) multiple sequence alignment

Xenopus         ATGCATGGGCCAACATGACCAGGAGTTGGTGTCGGTCCAAACAGCGTT---GGCTCTCTA
Gallus          ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC---CAAAATAACACCAACATGCAAATG
Bos             ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACCCAAAACAGCACCAACGTGCAAATG
Homo            ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACTCAAAACAGCACCAACGTGCAAATG
Mus             ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACTCAAAACAGCACCAACGTGCAAATG
Rattus          ATGCATCCGCCACCATGACCAGCGGGAGGTAGCTCTCAAAACAGCACCAACGTGCAAATG
                ******  **** *********  *  ***  *   * *** * *             *
```

Alignment and Gaps

- The quality of the alignment is essential : each column of the alignment (site) is supposed to contain homologous residues (nucleotides, amino acids) that derive from a common ancestor.

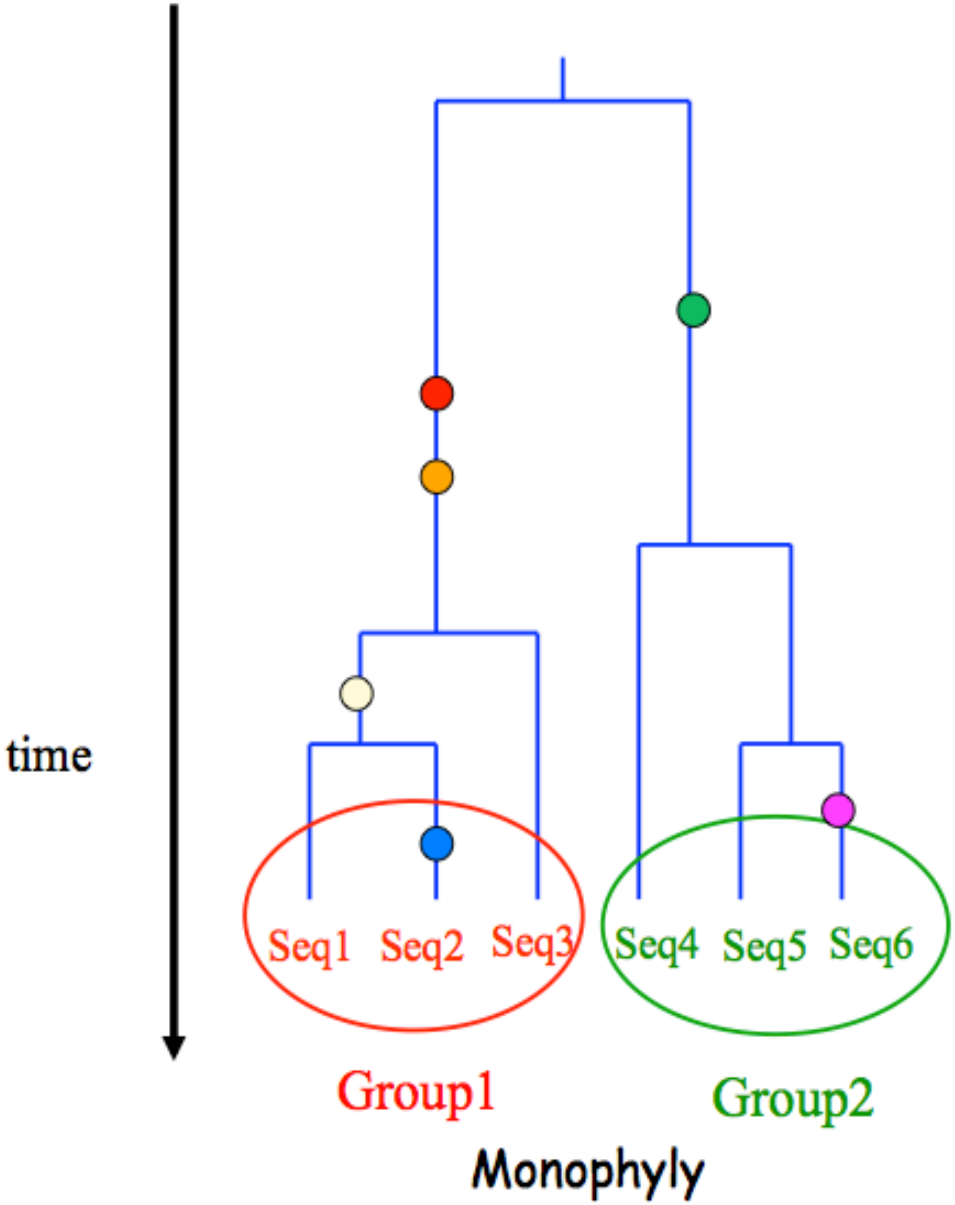
  ==> Unreliable parts of the alignment must be omitted from further phylogenetic analysis.

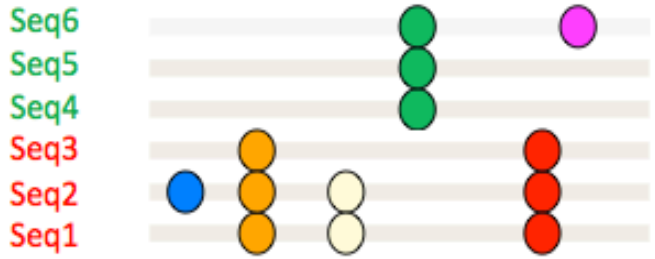
- Most methods take into account only substitutions ; gaps (insertion/deletion events) are not used.

  ==> gaps-containing sites are ignored.

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

Sequences Reflect Relationships





... =derived mutations

Molecular Phylogenies

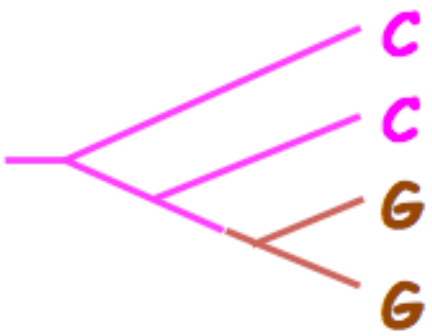
- The gene compared must evolve at a rate comparable to the divergence time of the organism; for example:
  - 18S rRNA gene for phylum-level divergences since it evolves very slowly.
  - Hemoglobin genes for mammalian orders.
  - Mitochondrial DNA for species divergences within a genus.
  - Repetitive DNA sequences (e.g. microsatellites) for individuals within species.

Caveat: homoplasy: independent evolution of the same character

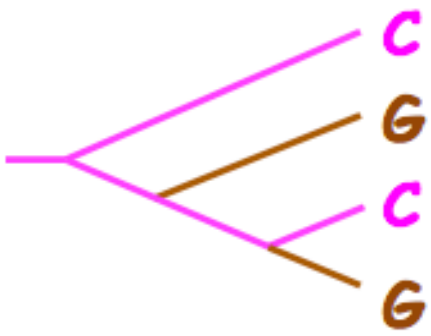
Evolutionary relationship:

Shared ancestral characters

Shared derived characters



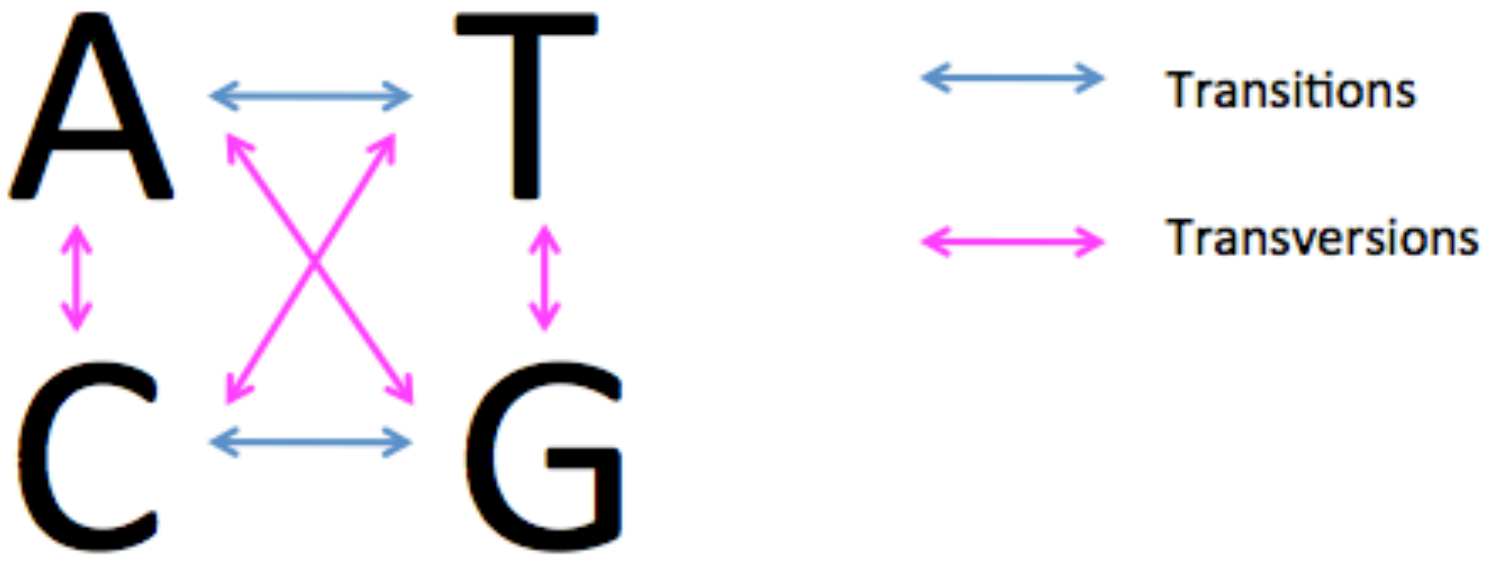
Homoplasy (independent evolution of the same character):



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

Substitution models



Variation in rates

Method of building a tree

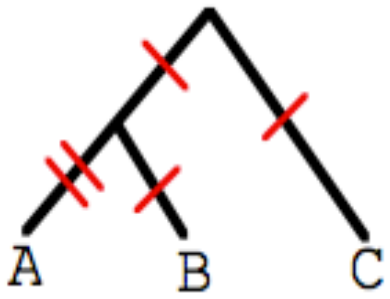
1. Distance methods
2. Character based methods
    1. Maximum parsimony
    2. Maximum likelihood
3. Bayesian inference

Distance methods

- Starts from a multiple sequence alignment
- Makes a matrices of pairwise sequence distances (number of differences)
- Builds a phylogenetic tree

Correspondence between trees and distance matrices

- Any phylogenetic tree induces a matrix of distances between sequence pairs
- "Perfect" distance matrices correspond to a single phylogenetic tree



tree

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

distance matrix

Building phylogenetic trees by distance methods

General principle :

Sequence alignment

↓ (1)

Matrix of evolutionary distances between sequence pairs

↓ (2)

(unrooted) tree

- (1) Measuring evolutionary distances.
- (2) Tree computation from a matrix of distance values.

Multiple sequence alignment

Species A ATGGCTATTCTTATAGTACG
Species B ATCTAGTCTTATATTACA

Aligned sequences

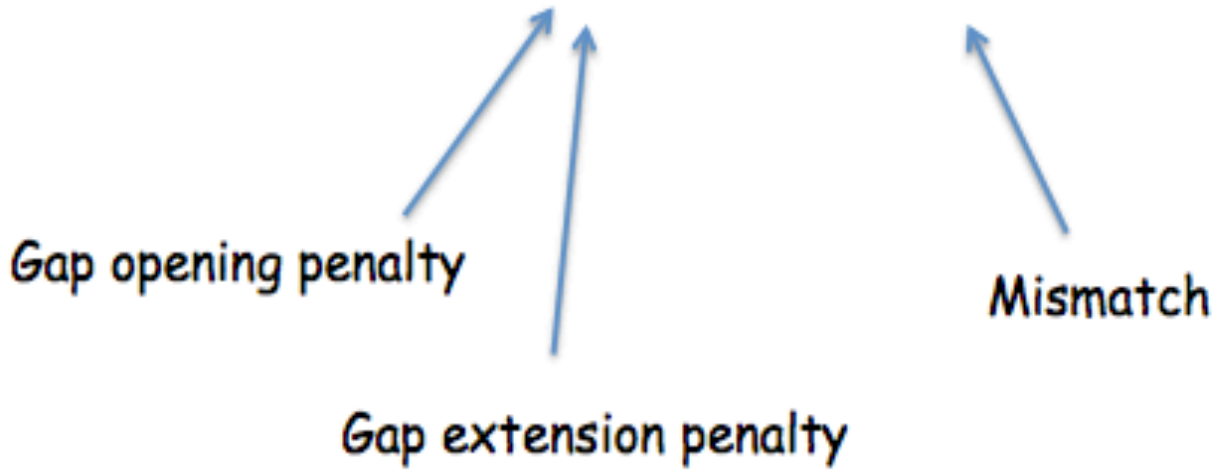
Species A ATGGCTATTCTTATAGTACG
Species B ATC--TAGTCTTATATTACA

Multiple sequence alignment

- Different softwares: ClustalW, ClustalX, Muscle

Species A ATGGCTATTCTTATAGTACG
Species B ATC --TAGTCTTATATTACA

Minimize total score



Principle of distance methods

Taxa	Characters
Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTCG

Transform the sequence data into pairwise distances

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	----	----	0.40	0.55	0.50
Species C	----	----	----	0.15	0.40
Species D	----	----	----	----	0.25
Species E	----	----	----	----	----

Distance methods

- **UPGMA** (Unweighted Pair Group Method with Arithmetic mean): same rate of evolution on each branch

- The **Neighbor Joining** method = most popular method

  does not assume the same rate of evolution on each branch of a 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

Building Trees with 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 and bayesian methods

- Allows for substitution rates to differ on lineages and sites: appropriate for distantly related species
- Estimates the likelihood of a tree=probability of the data given an evolutionary model
- Complex and computationally intensive!

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.

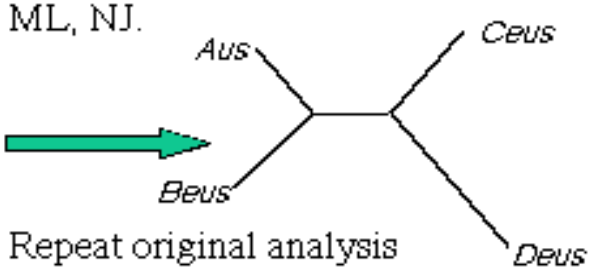
Bootstrapping

- 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

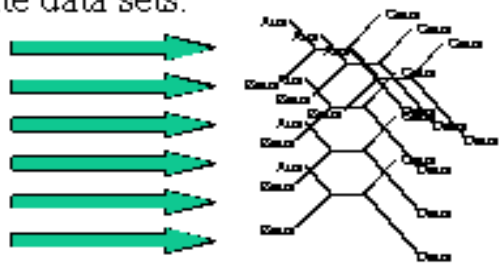
Original analysis, e.g. MP, ML, NJ.



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

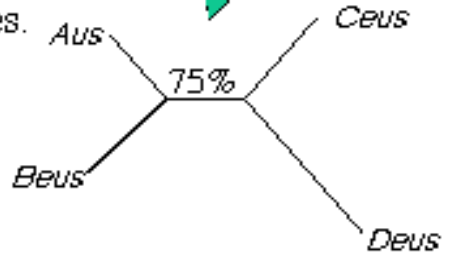
	2	3	13	8	3	19	14	6	20	20	7	1	9	11	17	10	6	14	8	16
Aus	G	A	A	G	A	G	T	G	A	A	T	C	G	C	A	T	G	T	G	C
Beus	G	G	A	G	G	G	T	G	G	G	T	C	A	C	A	T	G	T	G	C
Ceus	G	G	A	G	G	T	T	G	A	A	C	T	T	T	A	C	G	T	G	C
Deus	A	A	G	G	A	T	A	A	G	G	T	T	A	C	A	C	A	A	G	T

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.



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.

WWW resources for molecular phylogeny (1)

- Compilations

⇨ A list of sites and resources:
`http://www.ucmp.berkeley.edu/subway/phylogen.html`

⇨ An extensive list of phylogeny programs
`http://evolution.genetics.washington.edu/phylip/software.html`

- Databases of rRNA sequences and associated software

⇨ The rRNA WWW Server - Antwerp, Belgium.
`http://rrna.uia.ac.be`

⇨ The Ribosomal Database Project - Michigan State University
`http://rdp.cme.msu.edu/html/`

WWW resources for molecular phylogeny (2)

- Database similarity searches (Blast) :

http://www.ncbi.nlm.nih.gov/BLAST/
http://www.infobiogen.fr/services/menuserv.html
http://bioweb.pasteur.fr/seqanal/blast/intro-fr.html
http://pbil.univ-lyon1.fr/BLAST/blast.html

- Multiple sequence alignment

⇨ClustalX : multiple sequence alignment with a graphical interface (for all types of computers).

http://www.ebi.ac.uk/FTP/index.html and go to 'software'

⇨Web interface to ClustalW algorithm for proteins:

http://pbil.univ-lyon1.fr/ and press "**clustal**

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

END

Thank you!