Molecular Phylogeny

Ernst Haeckel, who coined the word 'phylogeny' illustrated evolutionary relationships using trees

Tree of Life, Ernst Haeckel, 1866
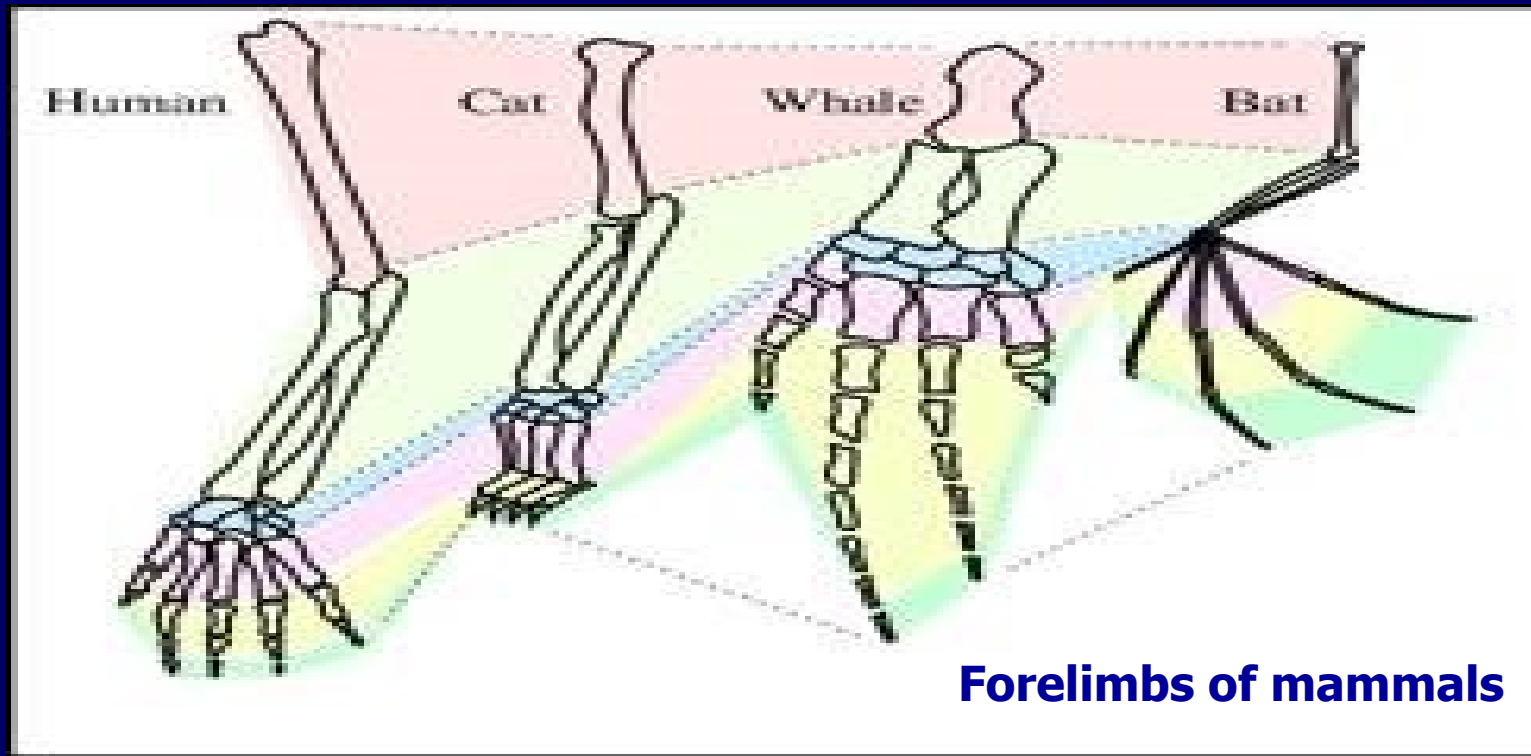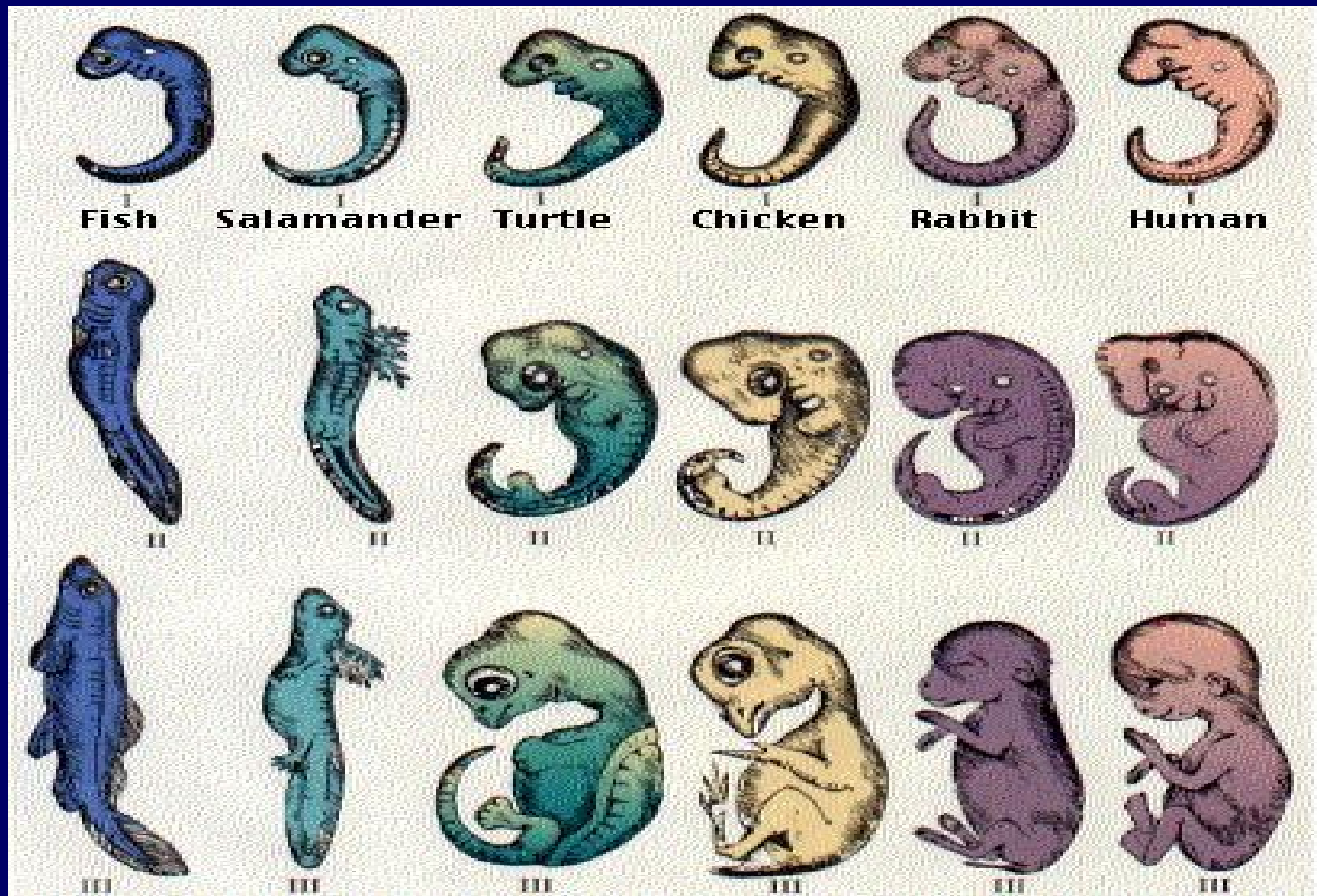
Phylogeny by Anatomy

- Before the birth of molecular biology, phylogeny was derived by looking for homologous **anatomical** structures or pattern of embryonic development
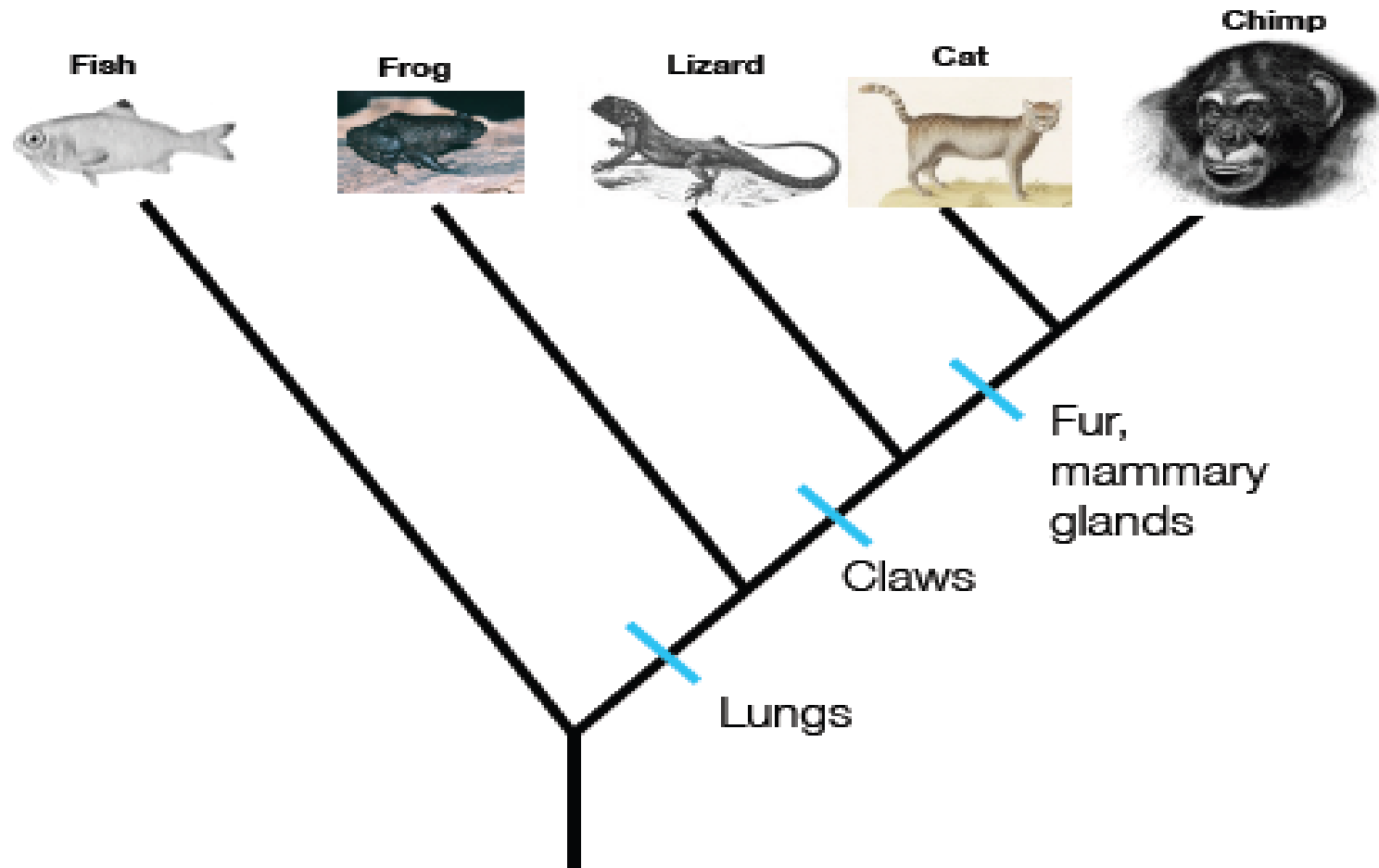

Forelimbs of mammals

Anatomical Homology

Embryonic Development of Vertebrates

Fish   Salamander   Turtle   Chicken   Rabbit   Human

Reconstructing a tree

Molecular homology

Number of mutations between the human haemoglobin alpha chain (142 amino acids long) and

- Human beta chain   0
- Gorilla   1
- Gibbon   2
- Rhesus monkey   8
- Dog   15
- Horse, cow   25
- Mouse   27
- Gray kangaroo   38
- Chicken   45
- Frog   67
- Lamprey   125
- Sea slug (a mollusk)   127
- Soybean (leghemoglobin)   124

Now phylogeny is based on proteins & DNA sequences

Molecular Phylogeny

- **Mol. Phylogeny is the study of evolution of <u>biological sequences</u> (nucleic acid / protein)**

- **An MSA is the first step for phylogenetic analysis**

- **Aligned sequences are represented as <u>outer branches</u> on an evolutionary tree**

- **In a gene family phylogenetic relationships**

  - **one can determine, family membership and how the family might have evolved over time**

  - **helps in identifying genes with equivalent function, (can then be tested by genetic experiments)**

Molecular Phylogeny

- **Phylogenetic analysis may also be used to follow changes occurring in a rapidly changing species *e.g.* influenza**

  - **Study rapidly changing genes**

  - **Next year's strain can be predicted**

  - **Flu vaccination can be developed**

- **Analysis of the types of changes (synonymous / nonsynonymous) within a population can reveal, for example, whether a particular gene is under selection**

Concept of Evolutionary Tree

- **2-dimensional graph showing evolutionary relationship between organisms, or genes from organisms**

- **Each sequence referred to as taxon, defined as phylogenetically distinct unit**

- **Taxa lie as outer branches (leaves) of tree, while internal nodes and branches represent relationship between taxa;**

  - closely related taxa are found on neighbouring branches
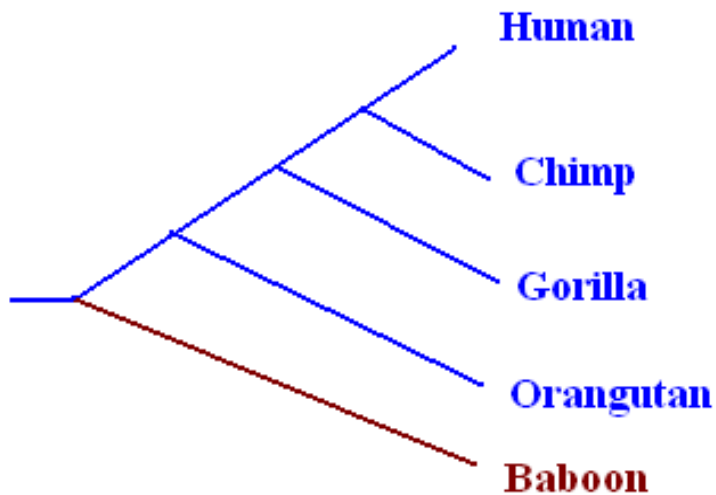
Concept of Evolutionary Tree

- **Branch length indicates the No. of mutations (a measure of evolutionary time) before the next level of separation**

- **Internal nodes represent the splitting of evolutionary path of the gene into two different species that are reproductively isolated**

- **<u>Molecular clock hypothesis</u> assumes <u>uniform</u> rate of mutation, suitable for closely related species**

 **– represented by <u>equal</u> branch lengths from the common ancestor**

Concept of Evolutionary Tree

- **Trees can be rooted or unrooted**

- **To construct a rooted tree include a taxon that we are reasonably sure branched off earlier than others, also termed as <u>outgroup</u>**

- **The sum of all branch lengths is termed as the <u>length</u> of the tree**

- **Trees are usually binary with two branches bifurcating at each node; may have more branches if taxa are too close to be resolved**
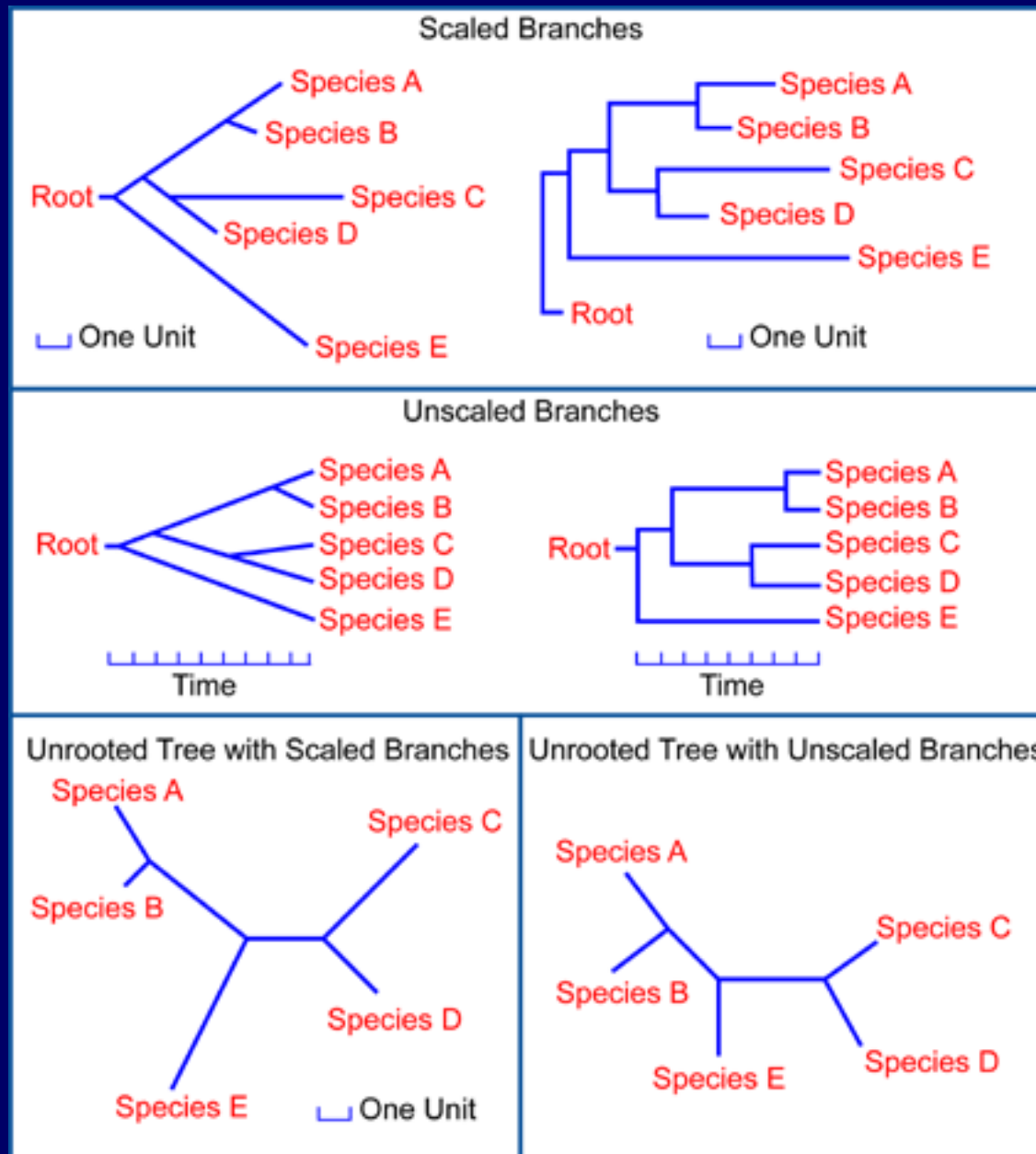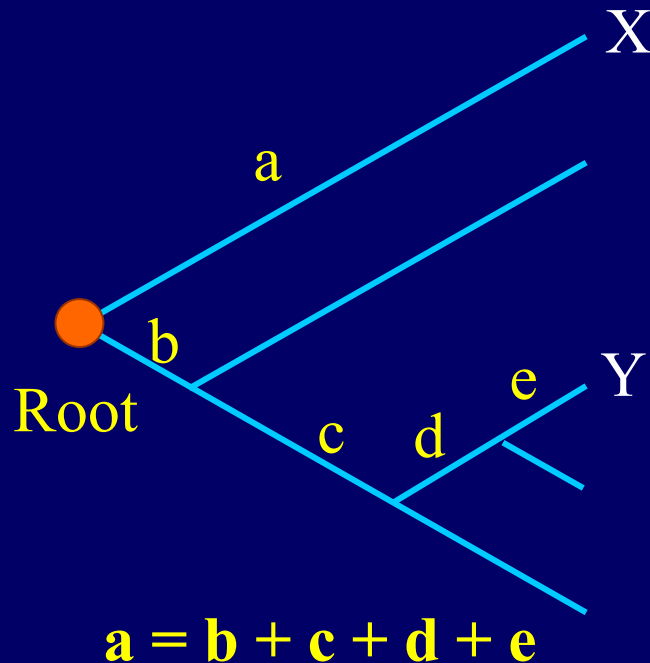
Example: Rooted & Unrooted

Rooted
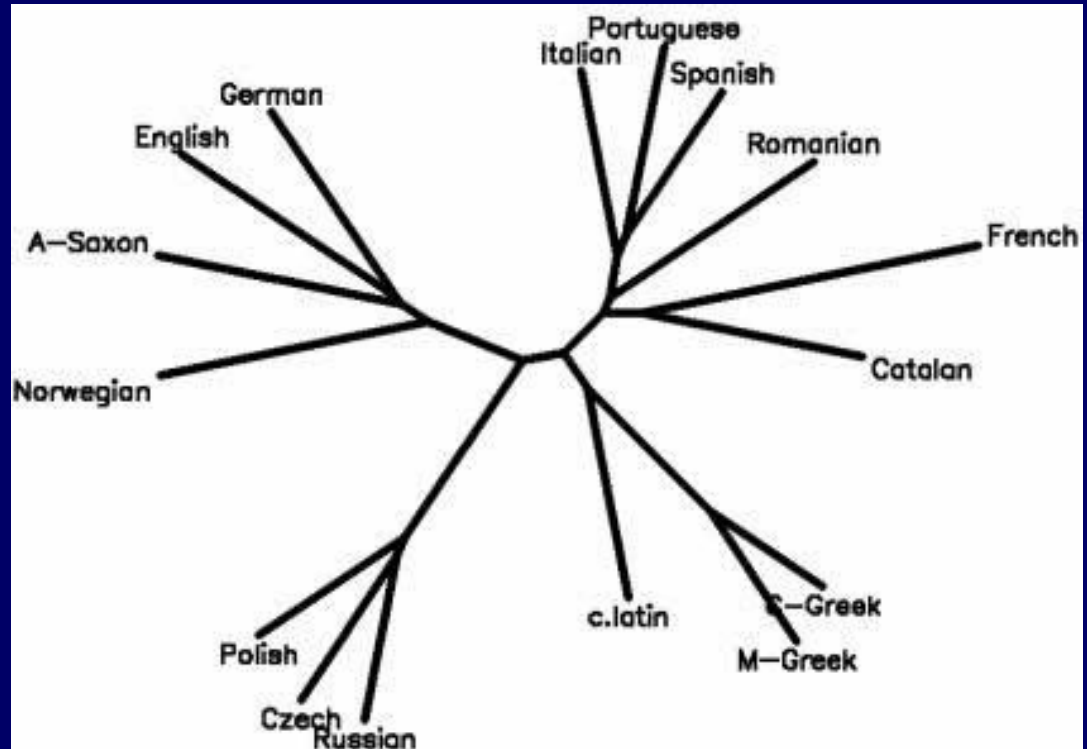
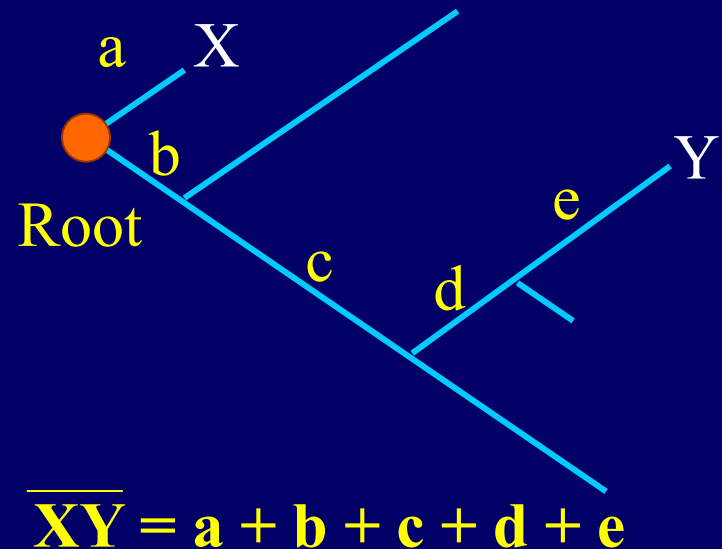**Unrooted**

Example: Scaled & UnScaled

Concept of Evolutionary Tree

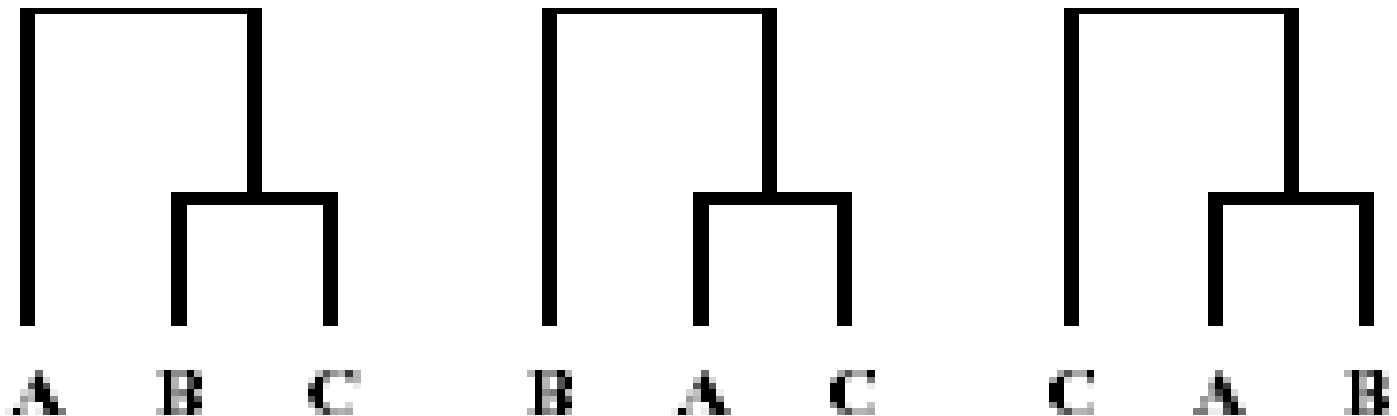## Ultrametricity

All tips are at equal distance from the root.

$$a = b + c + d + e$$

Additivity

Distance between any two tips equals the total branch length between them.

$$\overline{XY} = a + b + c + d + e$$

Tree reconstruction

- Tree reconstruction becomes difficult if all possible trees have to be scored before deciding on the optimal tree

- For three species there are only three possible trees

Tree reconstruction

- **With 4 species the No. of rooted trees becomes 15, 105 for 5 species and over 13 billion for 12 species**

- **For *n* species, the No. of rooted trees is given by**

$$(2n - 3)!/(2^{(n-2)}[n - 2]!)$$

- **For unrooted trees, the No. is significantly smaller (1 for n=3, 3 for n=4, 15 for n=5, and over half a billion for n=12):**

$$(2n - 5)!/(2^{(n-3)}[n - 3]!)$$

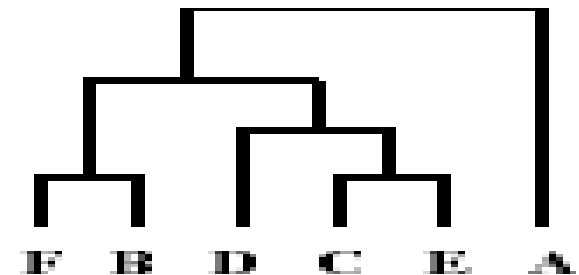- **This problem has been shown to be NP-hard**
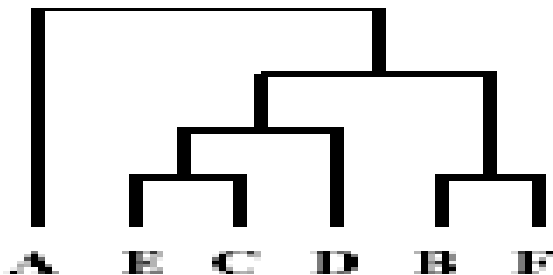
Number of possible evolutionary trees to consider as a function of number of sequences

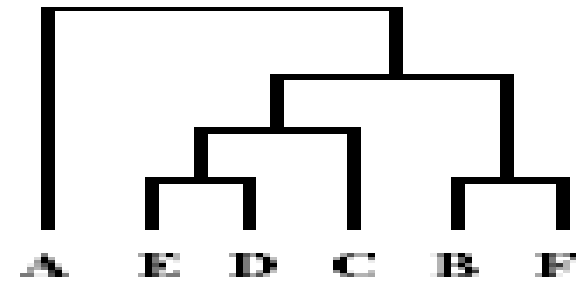| Taxa or sequence no.	No. of rooted trees	No. of unrooted trees
3	3	1
4	15	3
5	105	15
—	—	—
7	10,395	954

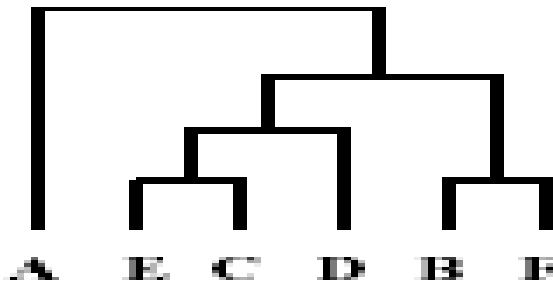# Terminology

- **Tree topology: ordering of species independent of branch lengths**

Same topology

Different topology

METHODS

```
┌─────────────┐     ┌─────────────┐     ┌─────────────┐   ┌─────┐   ┌─────────────┐
│ Choose      │     │ Obtain      │     │ Is there    │   │ Yes │   │ Maximum     │
│ set of      │ ──> │ multiple    │ ──> │ strong      │──>│     │──>│ parsimony   │
│ related     │     │ sequence    │     │ sequence    │   └─────┘   │ methods     │
│ sequences.1 │     │ alignment   │     │ similarity?3│             └─────────────┘
└─────────────┘     │ (Chapter 4).2│    └─────────────┘
                    └─────────────┘          │
```

Choose set of related sequences.[1] → Obtain multiple sequence alignment (Chapter 4).[2] → Is there strong sequence similarity?[3]

Yes → Maximum parsimony methods

No → Is there clearly recognizable sequence similarity?[4]

Yes → Distance methods

No → Maximum likelihood methods[5]

→ Analyze how well data support prediction.[6]

Methods for Phylogeny

- **Maximum parsimony approach is used for closely related sequences, and the variation is similar among all pairs of sequences**

- **Distance methods are used when more variation and intermediate level of similarity is present**

- **Maximum likelihood methods work for any set of sequences, but they are useful when sequences are very variable**
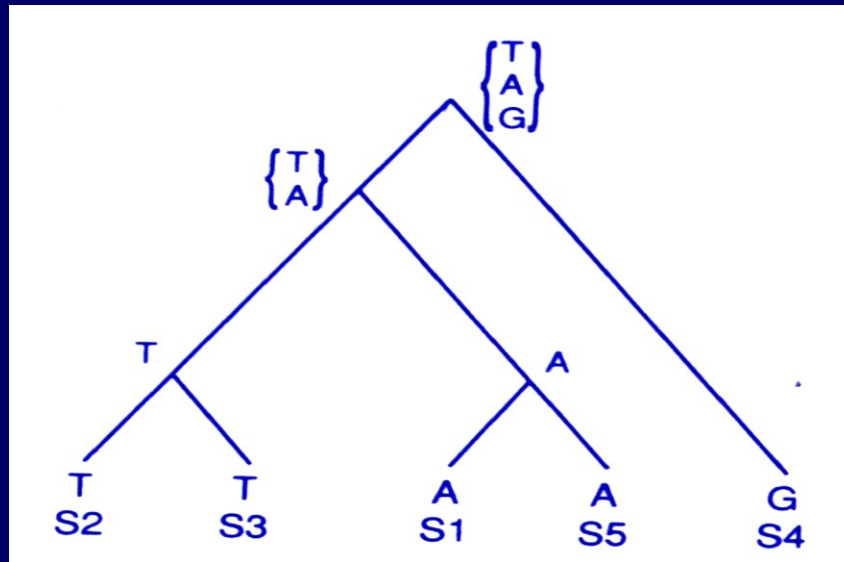
Maximum Parsimony Method

- **Predicts trees that <u>minimize</u> the No. of steps to generate observed variation**

- **For each aligned position, trees that require smallest no. of evolutionary steps to explain the observed changes are identified**

- **Finally the trees that produce the smallest no. of changes overall for all aligned positions are identified**

- **Works best for a small no. of closely related sequences**

parsimony score = 3
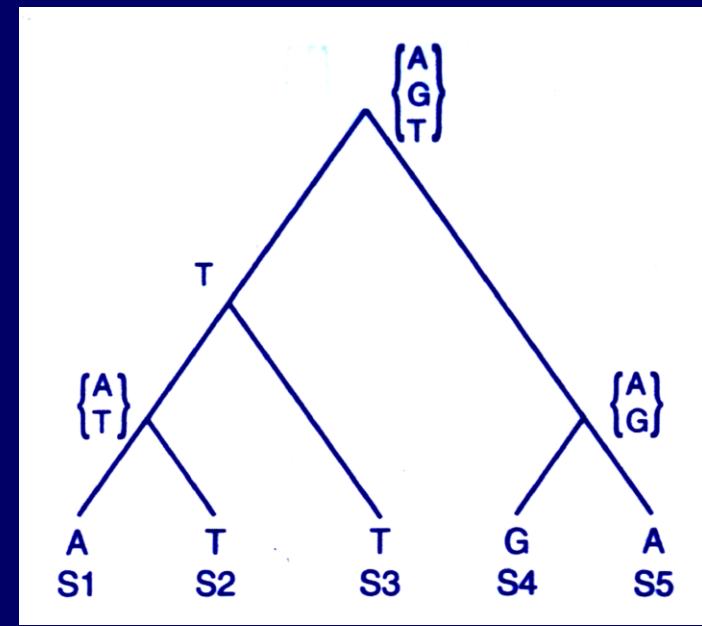
Maximum Parsimony

Consider another tree relating the same 1-base sequences: S1: A, S2: T, S3: T, S4: G, S5: A

Parsimonious Score = 2

Labeling the internal vertices as before, this tree has a parsimony score of 2; only two mutations are needed.

$\Rightarrow$ this tree is more parsimonious than the earlier one

Maximum Parsimony Method

- **Maximum Parsimony method does not use any model of molecular evolution.**

- **It is based on an implicit assumption that mutation is rare, and the best explanation of evolutionary history is the one that requires the least mutation.**

Phylip Programs: Parsimony

- **dnapars**: treats gaps as 5th state

- **dnapenny**: uses branch and bound search method

- **dnacomp**: based on compatibility criteria; finds tree that supports largest number of sites

- **dnamove**: performs parsimony and compatibility interactively

- **protpars**: based on no. of mutations to change a codon for aa1 to codon for aa2 for non-synonymous changes only

Distance Methods

- **Uses No. of changes between pairs of sequences in a group to construct a tree**

- **Sequences with fewest changes are neighbours, *i.e.* they share a node to which they are joined by a branch**

- **Aim is to position neighbours correctly and to compute branch lengths that best fit the data**

- **Methods for tree construction: UPGMA, Fitch-Margoliash, Neighbour-Joining, etc., based on distance matrix that gives substitution rates.**

Distance Methods

- **Distance is computed based on conditional probability of observing a change.**

- **A simplest model, called Jukes-Cantor assumes all substitutions have same rate of mutation, while more complex one define different rates for each type of substitution.**

$$M = \begin{pmatrix} p_{A|A} & p_{A|G} & p_{A|C} & p_{A|T} \\ p_{G|A} & p_{G|G} & p_{G|C} & p_{G|T} \\ p_{C|A} & p_{C|G} & p_{C|C} & p_{C|T} \\ p_{T|A} & p_{T|G} & p_{T|C} & p_{T|T} \end{pmatrix}$$

Descendent base ← → **Ancestral base**

Molecular clock hypothesis: No. of mutations observed is proportional to the evolutionary time

Jukes-Cantor model

Transition matrix for Jukes-Cantor model:

$$M = \begin{pmatrix} 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{pmatrix}$$

$$d_{\mathrm{JC}}(S_0, S_1) = t\alpha \approx -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

Value of α depends on the time step we use and features of the particular DNA sequence being modeled.

α is a probability, but can be interpreted as a rate at which observable base substitutions occur over one time step and is measured in units of

α = (substitutions per site) / time step

d = tα = expected no. of subs. per site during elapsed time

Kimura Models

Kimura 2-parameter model allows for different rates for transitions (β) and transversions (γ), while for Kimura 3-parameter model, all substitution rates are different:

$$M = \begin{pmatrix} * & \beta & \gamma & \gamma \\ \beta & * & \gamma & \gamma \\ \gamma & \gamma & * & \beta \\ \gamma & \gamma & \beta & * \end{pmatrix}$$

$$M = \begin{pmatrix} * & \beta & \gamma & \delta \\ \beta & * & \delta & \gamma \\ \gamma & \delta & * & \beta \\ \delta & \gamma & \beta & * \end{pmatrix}$$

with diagonal entries being 1-β-2γ, and 1-β-γ-δ, respectively.

$$d_{K2} = -\frac{1}{2}\ln(1 - 2\mathbf{p}_1 - \mathbf{p}_2) - \frac{1}{4}\ln(1 - 2\mathbf{p}_2)$$
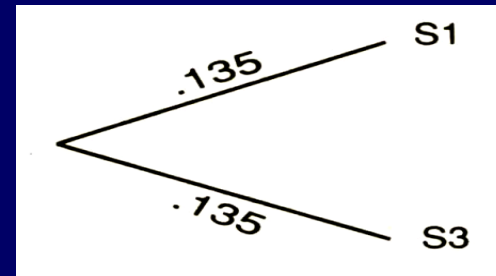
$$d_{K3} = -\frac{1}{4}(\ln(1 - 2\beta - 2\gamma) + \ln(1 - 2\beta - 2\delta) + \ln(1 - 2\gamma - 2\delta))$$

Tree Construction: UPGMA

Unweighted Pair-Group Method with Arithmetic Means (UPGMA), also called the average distance method.

- assumes that rate of change along the branches of a tree is a constant, i.e., it assumes a molecular clock.

	S_1	S_2	S_3	S_4
S_1		.45	.27	.53
S_2			.40	.50
S_3				.62

Distances Between Taxa

Step -1: pick the two closest taxa, S1 and S3, .27 distance apart.

Draw the edges, each of length 0.27/2 = 0.135, equidistant from the common ancestor.

UPGMA

Combine S1 & S3 into a group

	S_1	S_2	S_3	S_4
S_1		.45	.27	.53
S_2			.40	.50
S_3				.62

e.g., distance between S1-S3 and S2 is

(.45 + .40)/2 = .425

The table then collapses to:

	S1-S3	S2	S4
S1-S3		.425	.575
S2			.50

UPGMA

Step-2: Repeat the process, using the collapsed distance table. Because the closest taxa and/or groups in the new table are S1-S3 and S2, .425 apart:
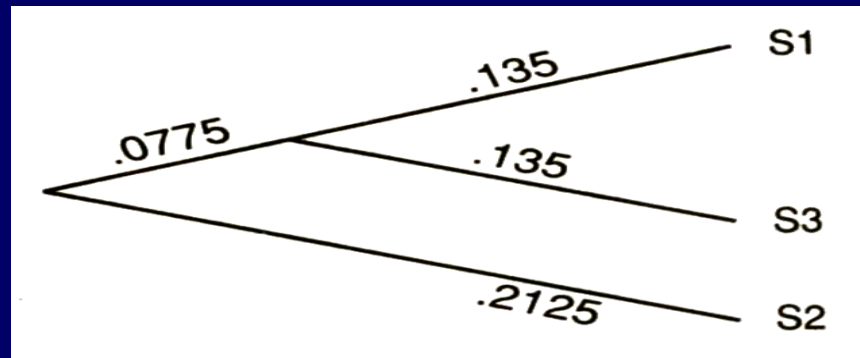
Edge to S2 will have length .425/2 = .2125, while the other new edge will be (.425/2) - .135 = .0775

	S1-S3	S2	S4
S1-S3		.425	.575
S2			.50

UPGMA

Step-3: **Again combining taxa, we form a group S1-S2-S3, and compute its distance from S4 by averaging the original distances from S4 to each of S1, S2, and S3: (.53 + .5 + .62)/3 = .55**

	S1-S2-S3
S4	.55

Final tree is drawn by estimating S4 as .55/2 = .275 from the root.

The other edge has length 0.275 – 0.2125 = .0625, since that places all other taxa .275 from the root as well.

UPGMA

	S_1	S_2	S_3	S_4
S_1		.45	.27	.53
S_2			.40	.50
S_3				.62
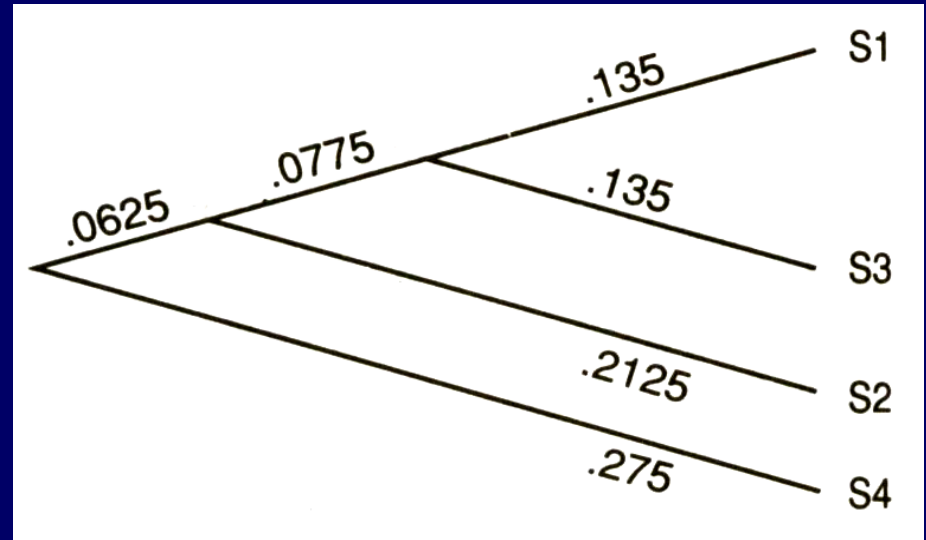
Does the constructed tree exactly fit the data?

Distance on the tree from S3 to S4 =.55, while according to the original data, it is.62!

To an approximation, the tree distances are reasonably close to the distances given by the data

UPGMA

Note: molecular clock assumption is implicit in UPGMA.

In this example, when we placed S1 & S3 at the ends of equal length branches, we assumed that the amount of mutation each underwent from their common ancestor was equal.

UPGMA always places all the taxa at the same distance from the root, so that the amount of mutation from the root to any taxon is identical.

UPGMA always produces a rooted tree.

Phylip Programs: Distance Methods

- **dnadist** is the Phylip program for computing distance among DNA sequences

- **protdist** is the program for computing distances for protein sequences

- **fitch** estimates branch length assuming additivity of branch lengths using Fitch-Margoliash method; molecular clock not assumed

- **kitsch** same as fitch but assumes molecular clock

- **neighbor** estimates phylogeny using the neighbour-joining or UPGMA method

Maximum Likelihood

- **Maximum likelihood is similar to maximum parsimony, in that analysis is performed for each column of the alignment, all possible trees are considered, and trees with fewest changes are usually more likely**

- **ML allows corrections for variations in the mutation rates by considering explicit evolutionary models**

- **Method can be used to explore relationships among more diverse sequences**

Maximum Likelihood

The basic approach of maximum likelihood is:

- **Specify a particular model of molecular evolution (e.g., Jukes-Cantor (all substitutions are equally likely), Kimura (different rates for transitions and transversions).**

- **Consider a specific tree for relating our taxa. Assuming the model of evolution and specific tree are correct, compute the probability that the DNA sequence in our data could have been produced.**

- This is the likelihood of the tree, given our data.

Maximum Likelihood

- Compute the likelihood for each tree

- Probability of each tree is the product of mutation rates in each branch

- Likelihoods given by each site (col) are multiplied to give the likelihood of the tree

- Choose the tree with the greatest likelihood as the tree best fitting the data.

- Phylip programs dnaml and dnamlk (same as dnaml except that is assumes a molecular clock) and proml and promlk (for proteins)
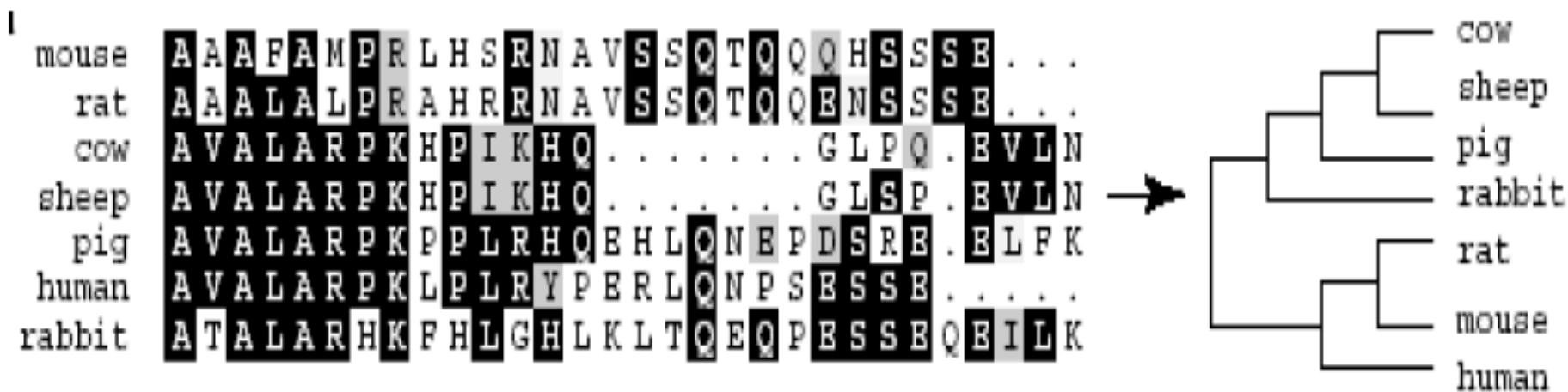
Bootstrapping

- Once a tree has been chosen by some method, it would be desirable to quantify how **confident** one is of it.

- This is given by the statistical technique - **bootstrapping**.

- In this procedure, the true data sequences are used to create a set of new **pseudo-replicate** sequences of the same length.

- Bases at a particular site in the new sequences are chosen to be the bases appearing in a randomly chosen site in the original sequences.

Bootstrapping

Initial alignment and tree:

In bootstrap analysis:

- **All sites are considered independent as if freely available in a 'hat' to pick**

- **Available sites are picked up randomly to reconstruct a new alignment of the <u>original size</u> and a new phylogeny**

Bootstrapping

- **Process is repeated many times to ascertain the strength of clustering. New "alignment" may contain several sites multiple times while some other sites may be absent - sampling with replacement**
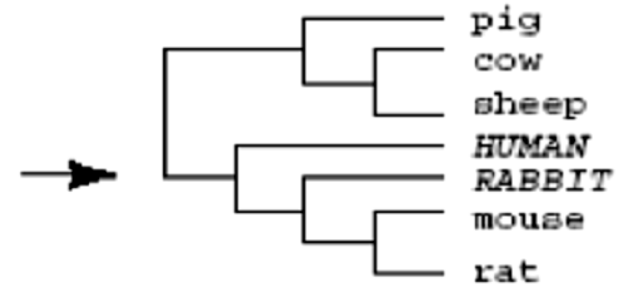
Bootstrapping

- **Phylogenies are compared to calculate values [Bootstrap value] that signify the number of times a given branch/cluster occurred in the Multiple bootstrap trees**

- **Higher the value - higher the confidence of phylogenetic inference**

- **In general values < 50% provide very poor support**

Which method is the Best?

One of the difficulties of picking a method is that one can find good arguments for and against them all.

Cautious approach - always use a number of different methods on the data.

Rather than trusting a single method to give an accurate tree, check to see if different methods give roughly the same results.

They often do, and if they do not, it is worth investigating why they don't.

Summarize

There are many approaches to phylogenetic inference that are **not** sequence-based.

Evidence of **all** (e.g., morphological similarities, identification of common ancestors from fossils) should be weighed before making too strong a statement about the phylogeny of the species under consideration.

Application

Let's consider a few examples of applications of phylogeny inference.

| Investigating whether the evolution of several species parallel one another:

(i) Evolution of hosts & parasites can be studied by constructing separate phylogenetic trees for each.

Similarity of the tree topologies can indicate whether the parasites evolved with the host, or if the parasites "jumped" from one host species to another.

Application

(ii) Likewise, trees for two symbiotic species, such as fungus-growing ants and the fungus they grow,

- help indicate how far back in evolutionary history the symbiotic partnership stretches.

Application

II Determining likely infection sources of human immunodeficiency virus (HIV) by constructing trees from HIV sequences from a number of infected individuals:

There have been several forensic applications of this

(the Florida Dentist AIDS cases, and the case of a doctor accused of intentionally injecting HIV into a former lover).
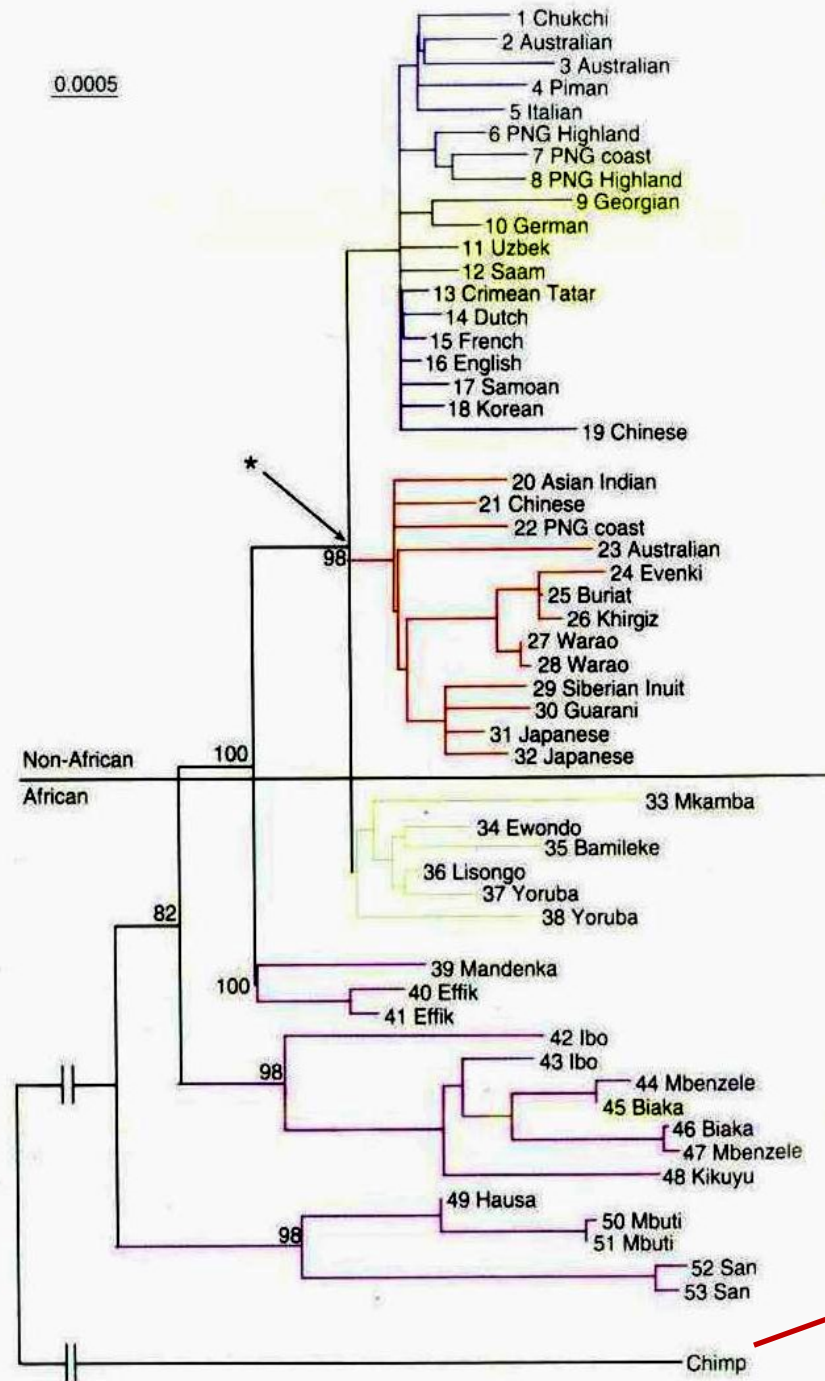
Application

III **Investigating the "Out of Africa" hypothesis of human origins:**

The clustering pattern on a tree constructed from human DNA sequences from ethnic groups around the world should help indicate how human populations are related and hence how and from where they spread.

Populations of humans that **share** particular sets of mutations are understood to be more closely related to each other than they are to populations lacking these mutations.

Phylogenetic relationships among representatives of human populations based upon complete mtDNA sequences

Scale: no. of nucleotide differences per site

0.0005

1 Chukchi
2 Australian
3 Australian
4 Piman
5 Italian
6 PNG Highland
7 PNG coast
8 PNG Highland
9 Georgian
10 German
11 Uzbek
12 Saam
13 Crimean Tatar
14 Dutch
15 French
16 English
17 Samoan
18 Korean
19 Chinese

*

98

20 Asian Indian
21 Chinese
22 PNG coast
23 Australian
24 Evenki
25 Buriat
26 Khirgiz
27 Warao
28 Warao
29 Siberian Inuit
30 Guarani
31 Japanese
32 Japanese

Non-African
African

100

33 Mkamba
34 Ewondo
35 Bamileke
36 Lisongo
37 Yoruba
38 Yoruba

82

39 Mandenka
40 Effik
41 Effik

100

98

42 Ibo
43 Ibo
44 Mbenzele
45 Biaka
46 Biaka
47 Mbenzele
48 Kikuyu
49 Hausa
50 Mbuti
51 Mbuti
52 San
53 San

98

Chimp

Tree suggests native Australians are more closely related to Japanese than they are to the major groups of Africans.

Note: diversity of African populations: Kikuyu people differ more from Effik than Europeans differ from Papua New Guinean highlanders.
 - they have had a longer time to evolve and accumulate mutations

It is observed that populations from the rest of the world are a subset of African populations, supporting the "Out-of-Africa" hypothesis that

- modern humans in other parts of the world are descendants of migrants that originated from an African population.

References

- **Bioinformatics Sequence & Genome Analysis, David W. Mount**

- **Biological Sequence Analysis, Probabilistic Models of Proteins and Nucleic Acids, R. Durbin, S.R. Eddy, A. Keoghs and G. Mitchison**