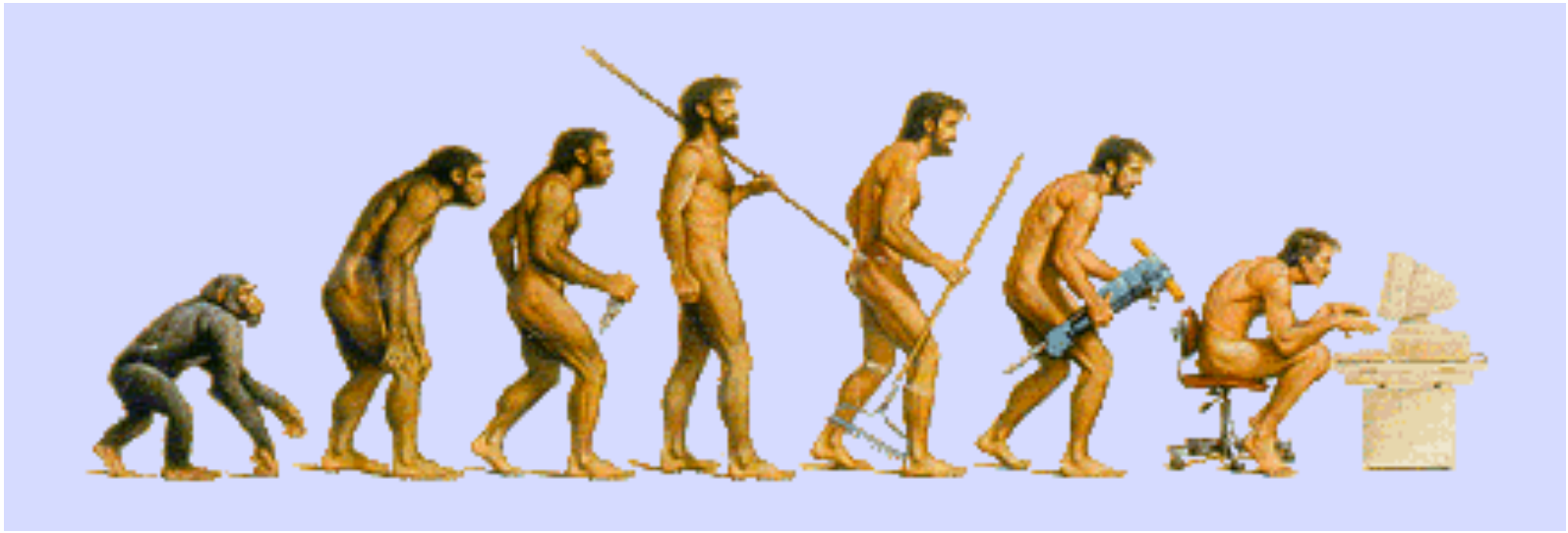
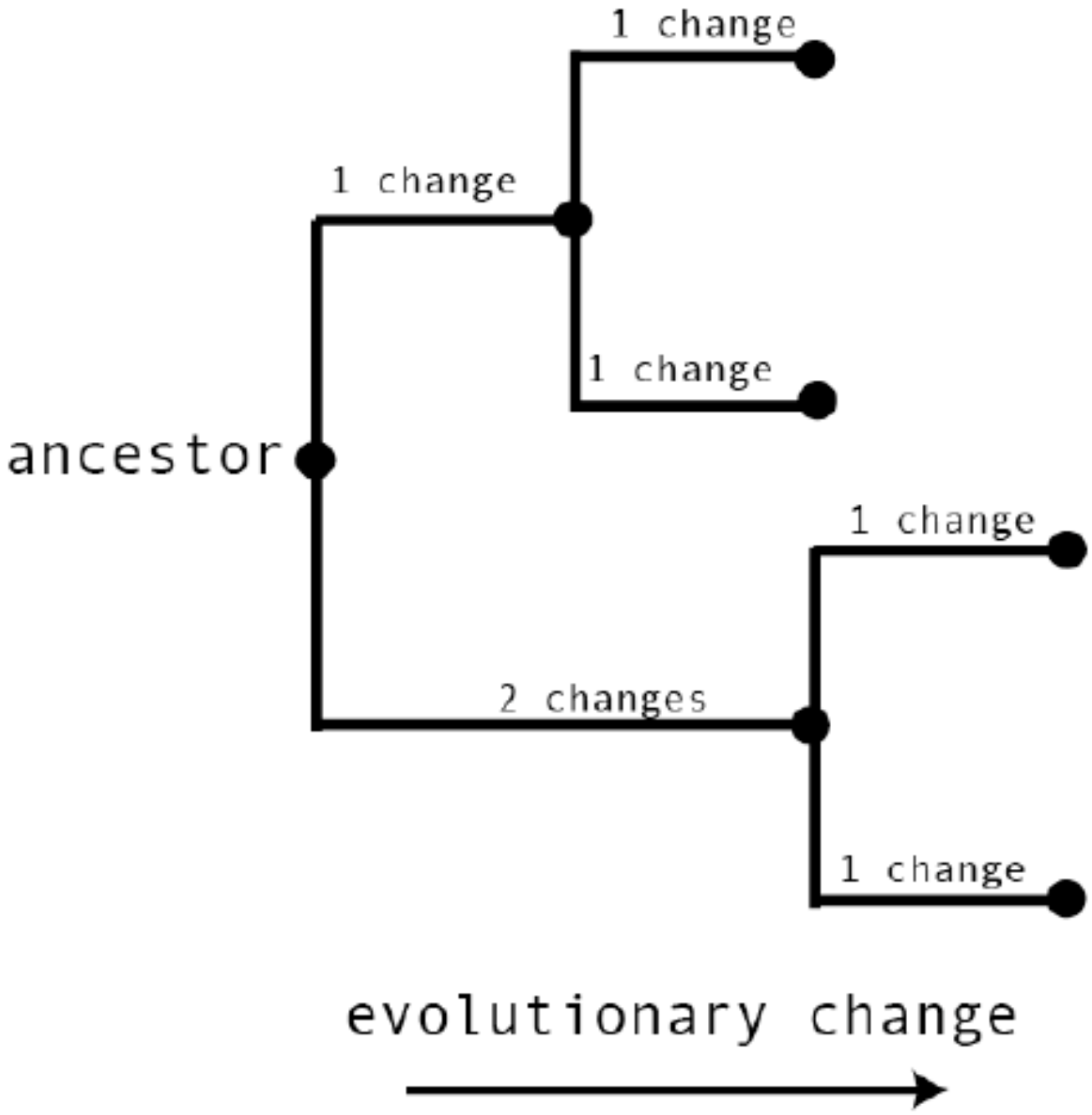
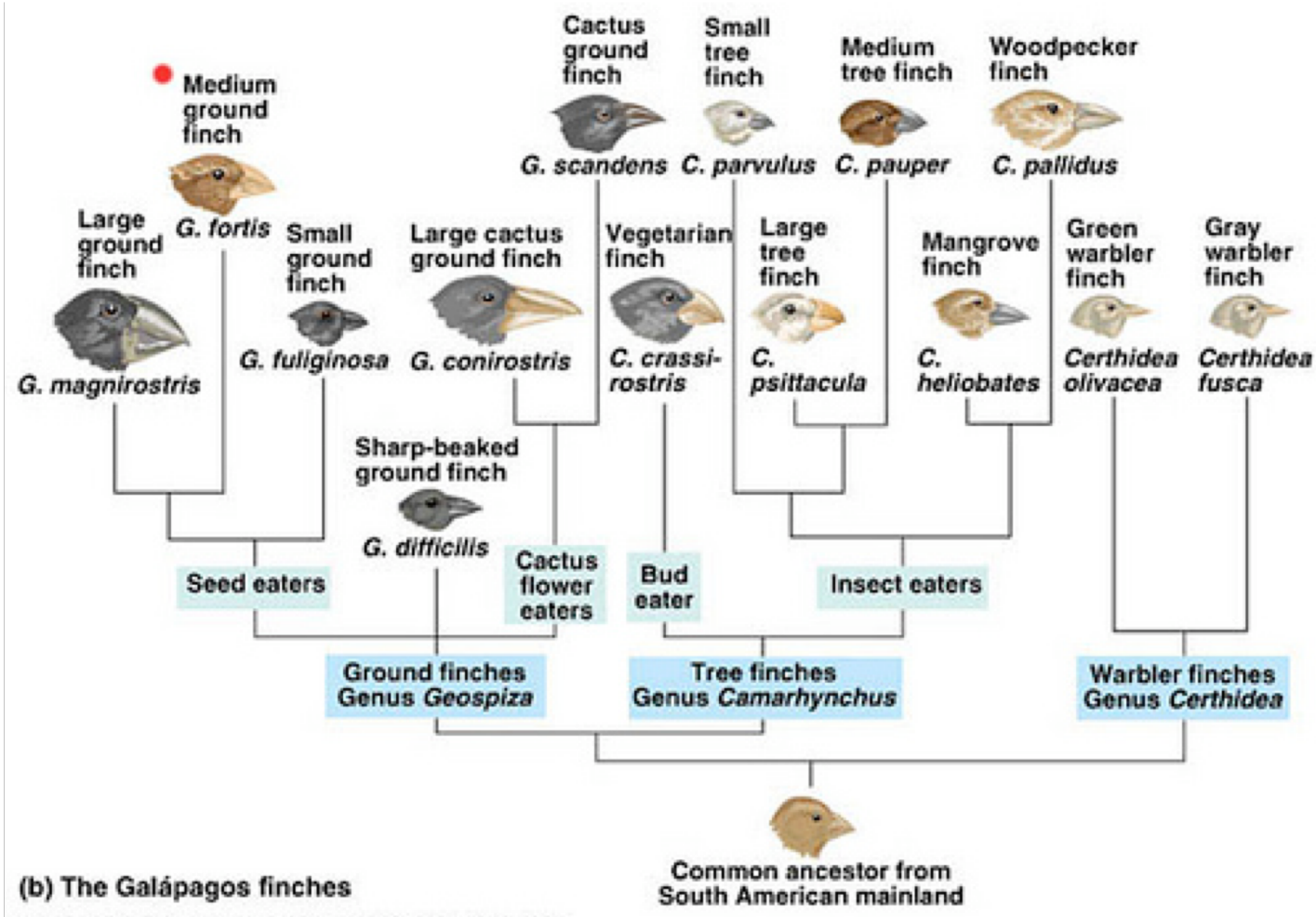
Phylogenetics
(a re-review)



The basics

- In general, the closer the two species are evolutionarily, the closer their genomes will be (Anthrax homework example)
- We will assume all life comes from a common ancestor
- Relationships can be illustrated using trees
  - Phylogenetics' task is to infer these trees


1 change
1 change
1 change
ancestor
1 change
2 changes
1 change
evolutionary change



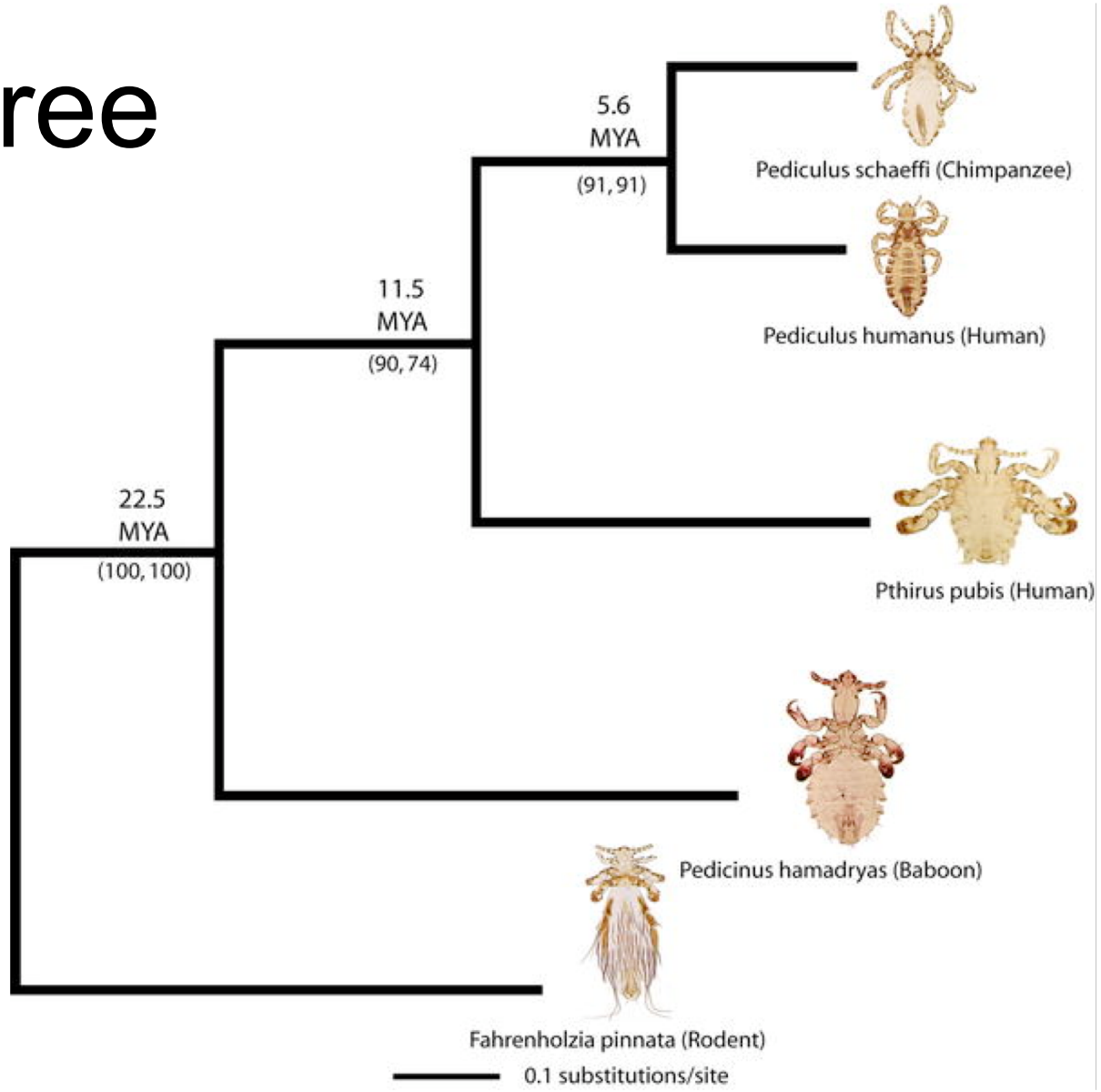
(b) The Galápagos finches

Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Interesting facts

- Darwin's finches are 14 species.
- They descend from one common ancestor that arrived in the Galapagos Island archipelago within the past 2-3 million years.
- “Incipient” species: nothing biological keeping them from forming hybrids, but happens very rarely in the wild due to mating call differences.

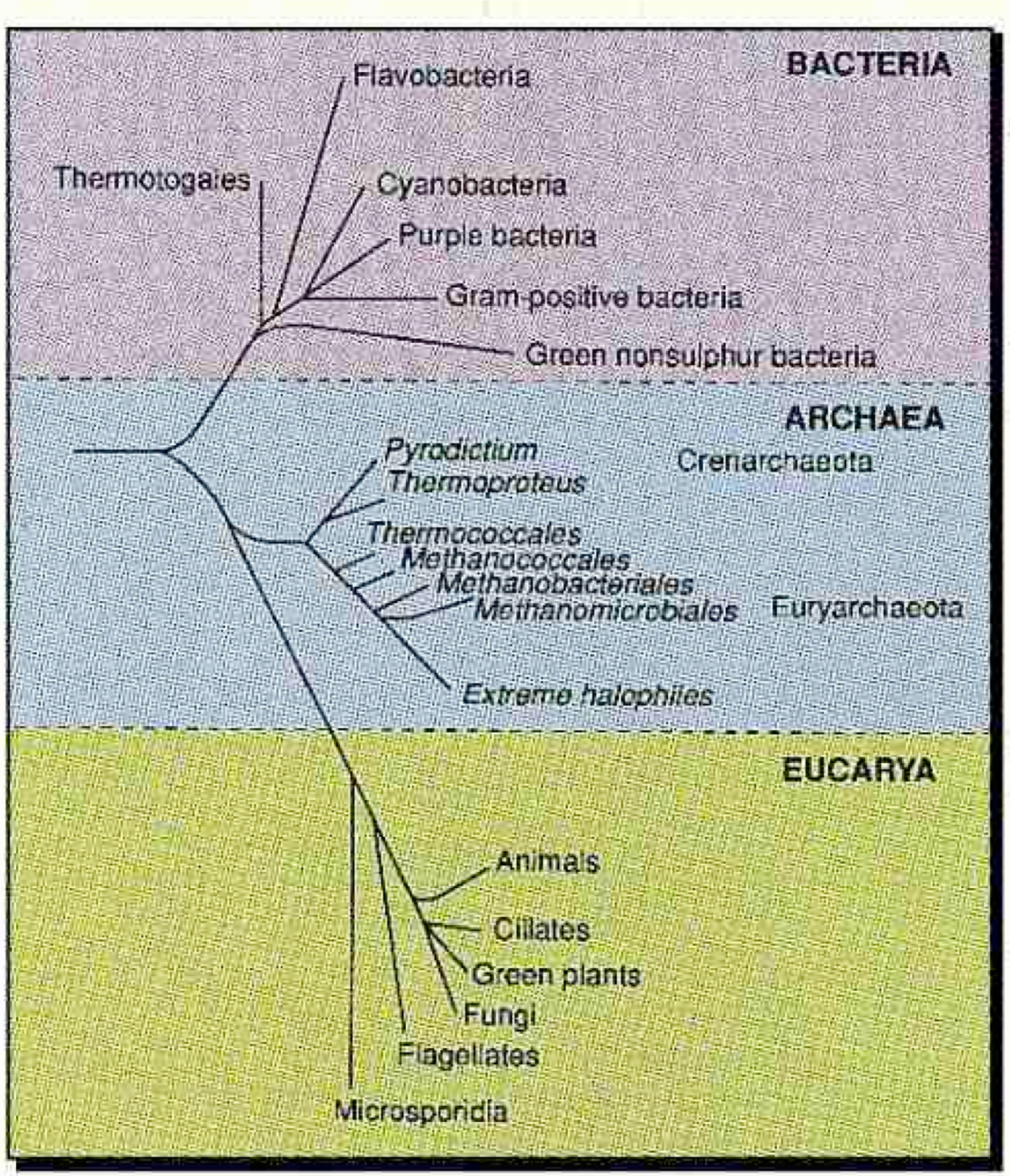
A sample tree



Reed et al. (2004)

Carl Woese

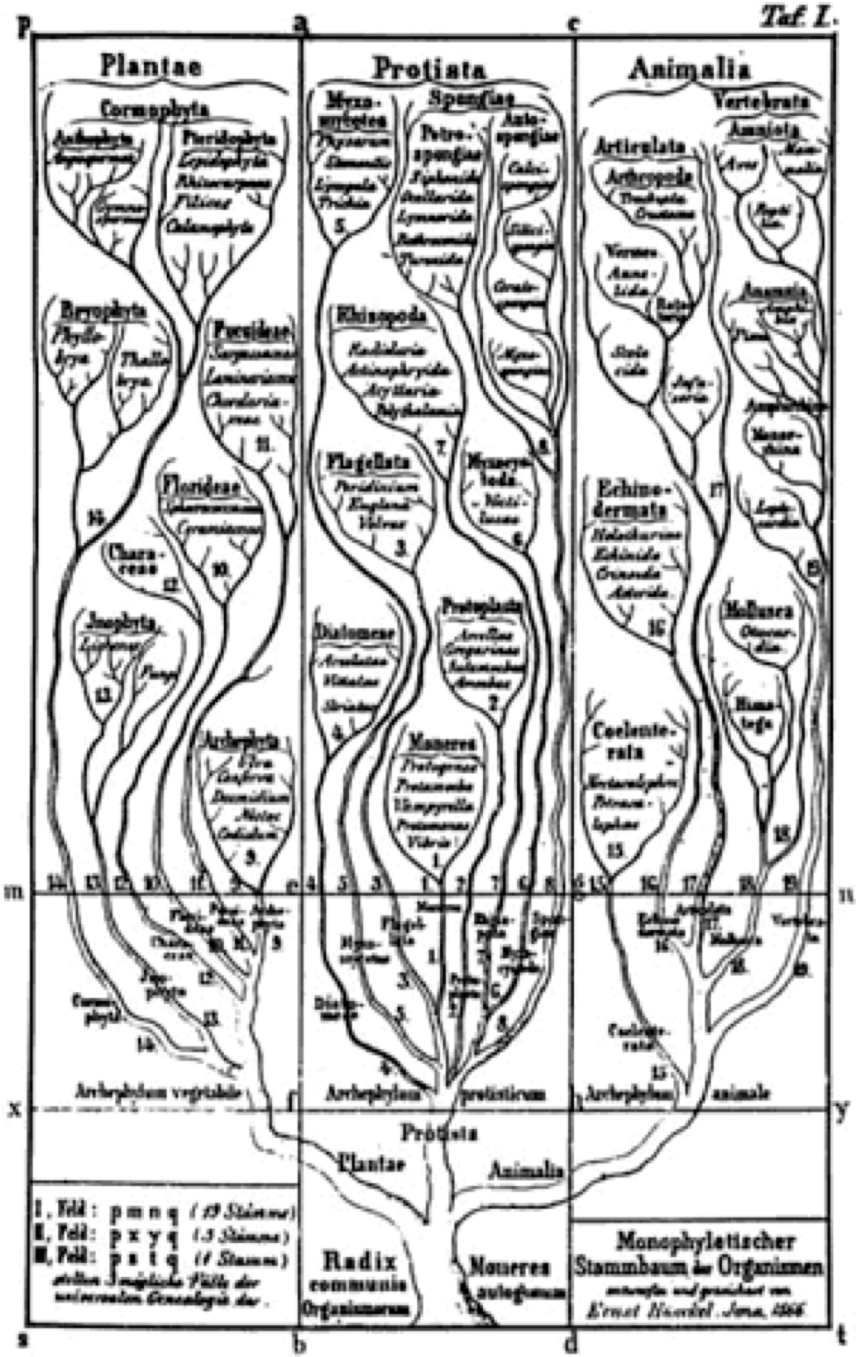
Early 16S rRNA tree



Woese's new tree of life. The position and length of each branch is determined by comparisons of ribosomal RNA

Tree of life (mid 19th century)

http://www.ucmp.berkeley.edu/education/events/eukevol.html



Background

- Phylogenetics comes from phylogeny (evolutionary history)
- Phylogenetics can be morphological or molecular
  - Paper clips is morphological
  - Sequence alignments are molecular
- Two areas of research:
  - Molecular systematics: infer “tree of life”
  - Molecular evolution: understand molecules in the context of related species (e.g., Ka/Ks)

Inference

- We infer trees because we don't really know all the species, esp. ancestors represented by internal nodes.

- Today, we'll start to discuss simple approaches for phylogenetic tree inference based on distance.

Why phylogenetics is hard

- Number of unrooted trees for more than 2 taxa is:

$$\frac{(2n-5)!}{(2n-3(n-3))!}$$

- # of rooted trees for more than 1 taxa:

$$\frac{(2n-3)!}{(2n-2(n-2))!}$$

- Example: 34,459,425 unrooted trees for only ten taxa

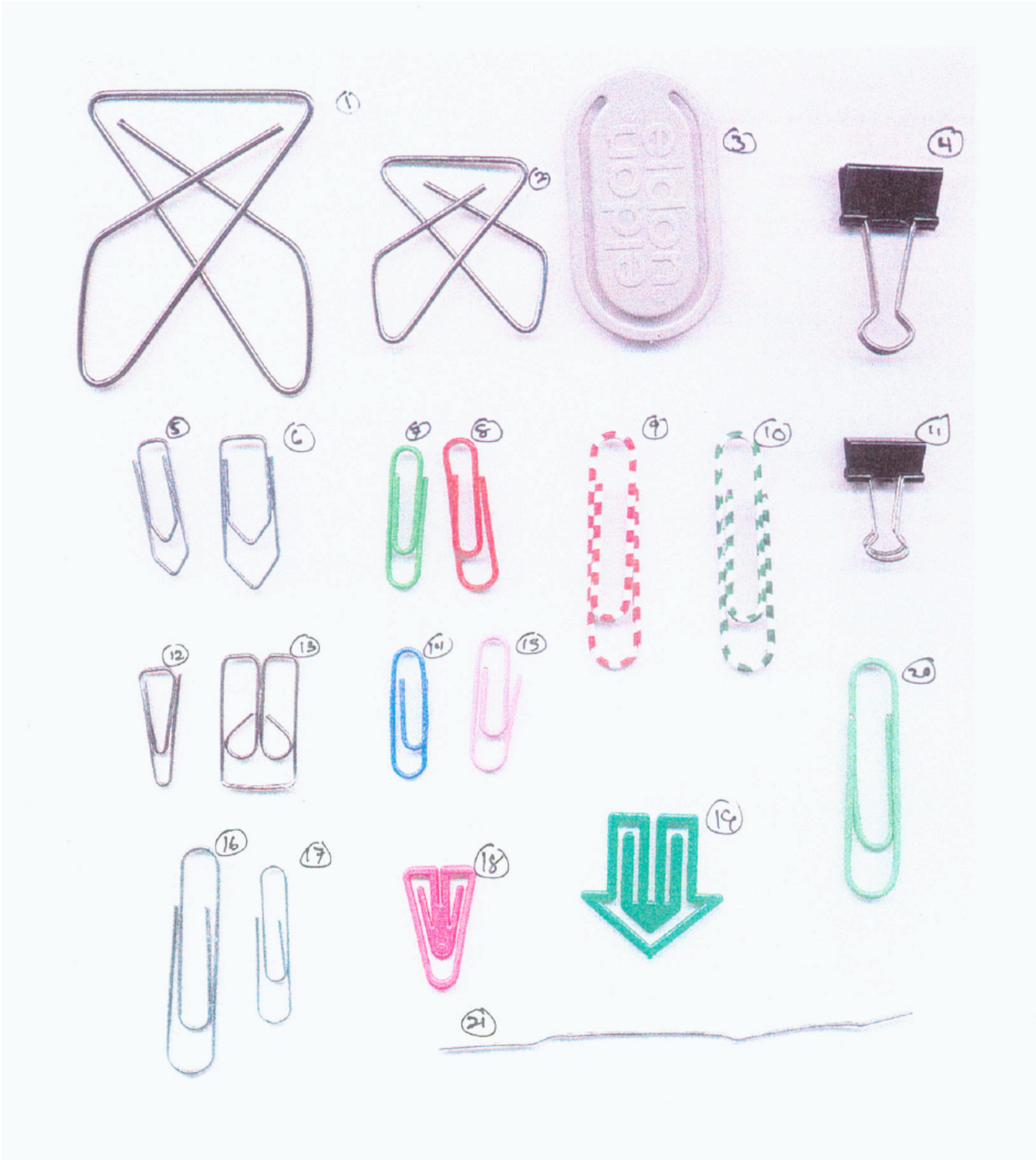
Approach

- Three main ways to build a tree:
  - Discrete (per site)
  - Distance (convert into pairwise distance)
  - Combination (make a tree from a bunch)
  - Optimal (looks at all possible trees)
  - Statistical (e.g., maximum likelihood)

Basic construction approaches

- Distance
  - Tree accounts for evolutionary distances estimated from data
- Parsimony
  - Tree that requires minimum about of change to explain the data
- Maximum likelihood
  - Tree that maximizes the likelihood of the data

Try it out again!



Source:
Warren Ewens
U. of Penn

SARS

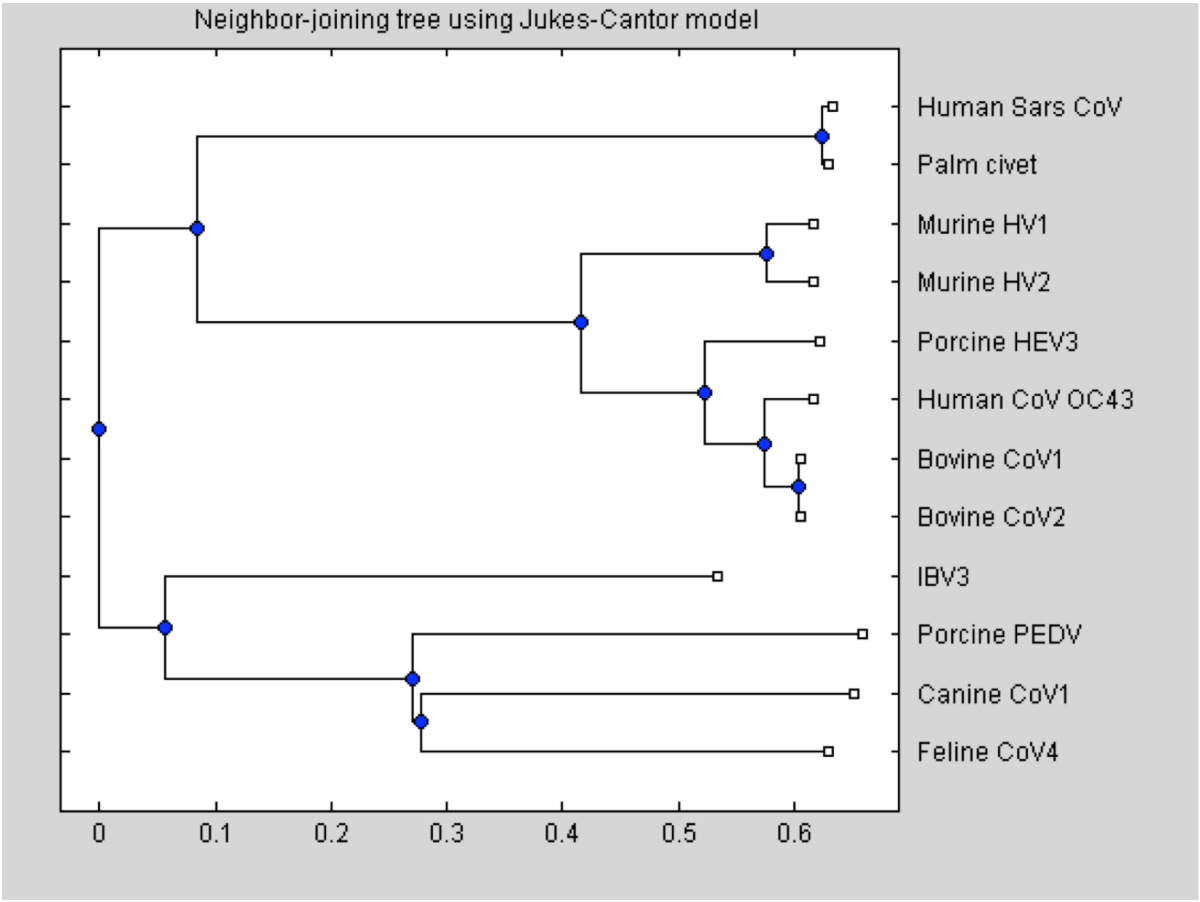
- The genome of SARS was sequenced by a Canadian group in April 2003
- 29,751bp, single stranded RNA sequence
- Has 5-6 genes in the typical structure of a coronavirus
  - One of the causes of the common cold

Genomics

- This is a good example of how epidemology, and virology benefit from the tools and algorithms in this course.

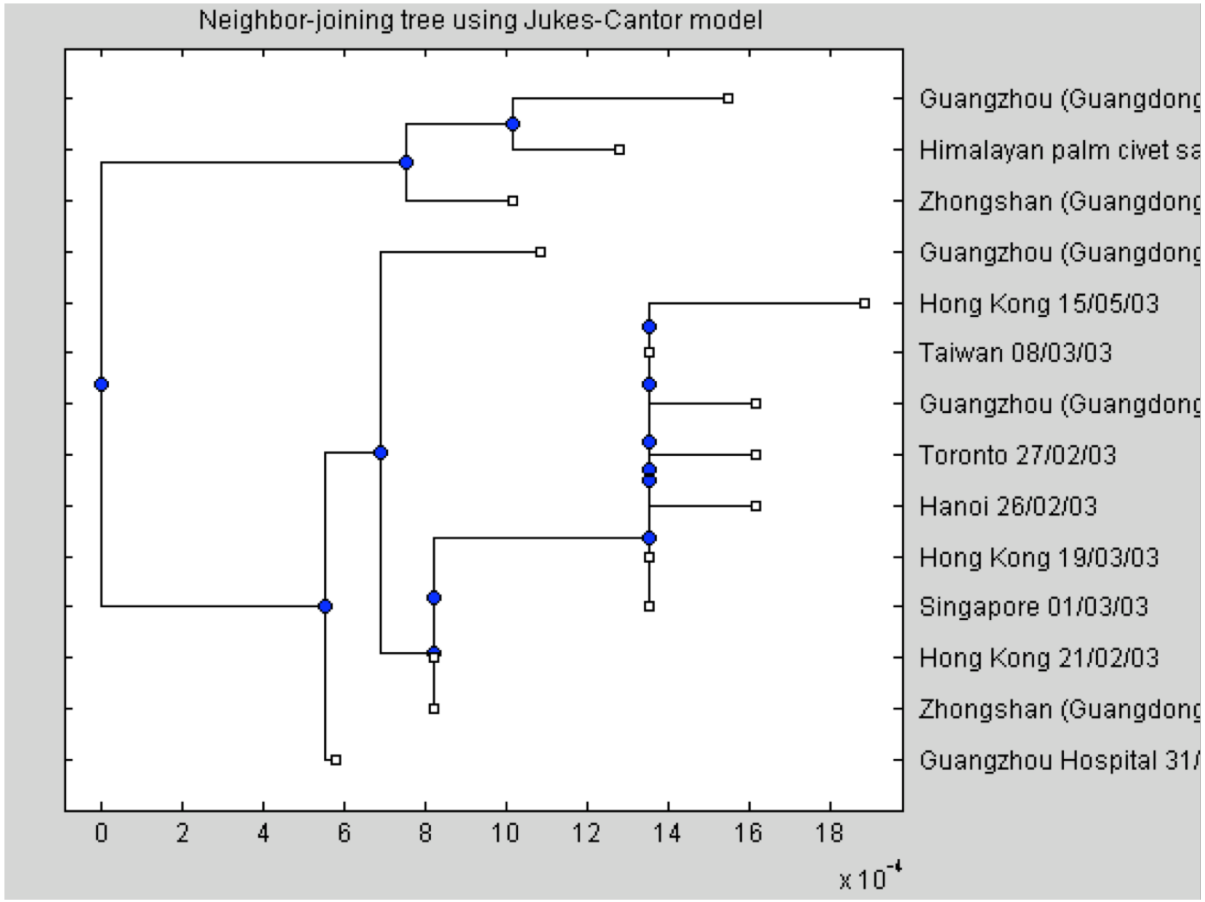
- Soon we will discuss:
  - What kind of virus cause the epidemic?
  - What species is it from?
  - Where did it start?

Where did SARS come from?



Himalayan palm civet

Neighbor joining can also be used to study epidemiology



Date of origin

- Genetic distance scales almost linearly with time

- Authors of the text estimate SARS jumped to humans around Sept 16 2002

Distance-based approaches to inferring phylogenetic trees

Review

- Input:
  - Data from a set of genes/species
- Output:
  - A phylogenetic tree that accurately characterizes the respective lineages

Details about ideal distance metrics

- $D(x_i, x_j) >= 0$
  - Distances must be non-negative
- $D(x_i, x_i) = 0$
- $D(x_i, x_j) = D(x_j, x_i)$
  - symmetric
- $D(x_i, x_j) <= D(x_i, x_a) + D(x_a, x_j)$
  - Additive property

Goal of distance approach

- Given a *m* x *m* matrix, where each value is the distance between two sequences.

- Build a tree such that distances between two leaves *i* and *j* is consistent with the matrix data.

UPGMA Method

- Unweighted Pair Group Method using Arithmetic Averages

- Distance is defined between two clusters C_i and C_j such that:

$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$$

Basic idea

- D_{ij} is the average distance between pairs of taxa from each cluster

- Algorithm:
  - Start with one taxa per cluster
  - Iteratively pick two clusters and merge
  - Create a new node in the tree for the merged cluster

More specifics

- Place each taxon at height 0 in the tree
- While more than two clusters:
  - Determine clusters with smallest d_{ij}
  - Merge clusters into a new one C_k
  - Make a new node k at height $d_{ij}/2$
  - Replace C_i and C_j with C_k
  - Recompute distance of C_k to other clusters
- Hook in the two remaining clusters to the root with height calculated as above.

Updating distances

- Distance between C_k and C_l defined as:

$$d_{kl} = \frac{d_{il} \mid C_i \mid + d_{jl} \mid C_j \mid}{\mid C_i \mid + \mid C_j \mid}$$

In-class example

- Consider the following symmetric matrix:

	A	B	C	D	E
A	0	5	3	8	10
B	5	0	5	8	10
C	3	5	0	8	10
D	8	8	8	0	1
E	10	10	10	1	0

UPGMA visually