Phylogenetics
The basics

- We will assume all life comes from a common ancestor

- Relationships can be illustrated using trees
  - Phylogenetics’ task is to infer these trees
ancestor

1 change

1 change

2 changes

1 change

1 change

evolutionary change
(b) The Galápagos finches

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

Source:
Warren Ewens
U. of Penn
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)
Building trees


• Based on protein sequences
Tree of life (mid 19th century)

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

Early 16S rRNA tree
Phylogenetic tree basics

- Nodes: points that connect branches
- Branches: lines that connect nodes
- Taxa: things being compared
- Rooted tree: one node is the base
- Unrooted tree: no explicit starting point
Rooted vs unrooted

Rooted

Unrooted
Tree styles

- Trees can be thought of as a mobile
- Internal nodes represent common ancestors
Getting it right

• Some but not all distance based methods:
  – UPGMA
  – Neighbor joining
  – Invariant
  – Dollo
  – Wagner
  – Sokal
  – ...
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
Interesting study

• [http://www.pnas.org/content/99/22/14292](http://www.pnas.org/content/99/22/14292)

• First time phylogenetics was used in a criminal court case in the U.S.
Case study: SARS

• The text outlines the story of the origin of SARS and how phylogenetics played a role

• In the beginning, there was an outbreak in Vietnam; in only a few weeks, the WHO official and 5 hospital workers were dead
Back story

- Actually first appeared in Guangdong province, China late 2002
- Most people got sick in a hospital, one doctor in hospital visited Hong Hong
- Travelers staying on the same floor as the doctor, who died, then got sick
- These people brought it to other places in the world
Genomics

• This is a good example of how epidemiology, 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?
But first....

- Lets review basics of infering phylogenies
- All molecular reconstruction methods assume you start with a set of aligned sequences
- This provides the *homology* information we need, and is critical.
Approaches

• 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)
Other species trees

http://members.aol.com/darwinpage/trees.htm
Example gene tree

Lodish et al. (2000)
Basic construction approaches

• Distance
  – Tree accounts for evolutionary distances estimated from data

• Parsimony
  – Tree that requires minimum amount of change to explain the data

• Maximum likelihood
  – Tree that maximizes the likelihood of the data
Try it out again!

(make distance matrix to help)

1: AATAT
2: ATTAT
3: TTTTT
4: CCTTT

Source:
Warren Ewens
U. of Penn
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.
A sample tree

Reed et al. (2004)
Lice support the hypothesis that Homo erectus and Homo sapiens were separate for a period of time.
Preparing for Thurs

• **Input:**
  – Data from a set of genes/species

• **Output:**
  – A phylogenetic tree that accurately characterizes the respective lineages
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
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