Phylogeny

- We do not have a time machine to see evolution unfold
- We cannot do experiments on evolutionary history
- Inferences are indirect (except in bacteria!!)
- Make phylogenetic hypotheses that are provisionally accepted (as is any scientific statement)
What is the truth?

- We cannot know the truth
- Is there any truth in science?
- Can only make hypotheses.
- Phylogenies represent our best hypothesis for what happened in the past based on the evidence that we have on hand.
- Our hypothesis may be true for false
- We hope that the accumulation of evidence will lead us closer to the truth.

Ernst Haeckel published one of the earliest phylogenetic trees in 1866
Phylogenetics is a component of systematics

Aim to understand the evolutionary relationships among units of inheritance:

• Genes
• Species
• Higher taxonomic groups

Modified from Darwin 1859
Hemoglobin

- **β Subfamily** (± 50 kb)
  - Early in fetus: ε
  - Fetal liver spleen & bone: Gγ, Aγ, ψβ1, δ

- **α Subfamily** (± 30 kb)
  - Psuedogene: ςζ, ψζ1, ψα1
  - Adult: α2, α1

Li 1997
Hartwell et al. 2000

---

**α-like family** includes three functional loci:
- α₁, α₂, ζ (zeta)

**β-like family** includes five functional loci:
- β, ε (epsilon), δ (delta),
  - Gγ (G-gamma)
  - Aγ (A-gamma)

---

### Graph

- **Percentage of total globin synthesis**
- **Postconceptual age (weeks)**
- **Birth**
- **Postnatal age (weeks)**
• Phylogeny of paralogous genes within a genome

Hartwell et al. 2000
Phylogeny and systematics

Aim to understand the evolutionary relationships among units of inheritance:
- Genes
- Species
- Higher taxa

This lecture
- What does a phylogeny show you?
- What are the major philosophical approaches to phylogeny construction?
- How can we use trees to test other evolutionary processes?

Modified from Darwin 1859
What does a phylogeny represent?
Different tree topologies convey the same information

Branches can rotate
Branch lengths have different meanings

**Scaled branches**
- Length proportional to the number of changes

**Unscaled branches**
- Position of nodes calibrated by a model of base pair change

**Unrooted trees**
- Shows relationships without identifying a common ancestor
What can you do with a phylogeny?

- Resolve historical relationships among groups
- Improve our understanding of the process of evolution
- Test evolutionary hypotheses
  - Sequence of events
  - Timing of events
  - Pattern of change

Modified from Darwin 1859
Before 1950’s, taxonomists constructed phylogenies

- Some characters deemed “important”
- But definitions seemed somewhat arbitrary
  1. Based on common ancestry
  2. Indicative of adaptive divergence
- Opinion
Numerical taxonomy introduced in 50’s

- Look at the overall degree of similarity using as many all available data
- Rationale: any traits that are heritable and that vary among the taxa should be useful
  - Presence or absence of skeletal traits
  - Presence or absence of flower parts
  - Mode of larval development
- Seem fairly straightforward
<table>
<thead>
<tr>
<th>Character</th>
<th>Taxon</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>G</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>H</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>I</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(B) Number of shared character states

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(C) Phenogram

```
A       C       B       D
A - 6 7 3
B - 4 0
C - 5
D -
```
Phenetic trees can be built using trait or molecular data by counting the number of similarities between taxa.

<table>
<thead>
<tr>
<th>Amino acid sequence</th>
<th>DNA sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGAGC</td>
<td>TGAAAG</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>ATT</td>
<td>ATT</td>
</tr>
</tbody>
</table>

Cytochrome c
Phenetic tree based on the number of similarities and differences between the amino acid constituents of the protein cytochrome c of different organisms.
Problems with phenetics

• Two reasons why organisms may share characteristics?
  1. Character evolved in the ancestor and was inherited by descendants (homologous)
  2. Convergent evolution – same character state evolved multiple times in different lineages (analogous)

**Tetrapods** (Greek tetrapoda, Latin quadruped, "four-legged") are vertebrate animals having four feet, legs. Ancestral state is five toes.

Human  Iguana
Analogous features that look similar

All are sharp, stiff and nasty

Spines are modified leaves

Thorns are modified branches

Prickles are modified outgrowth of the epidermis
So what is the real thang?
Principle of Parsimony: evolution will take the simplest route. Or the best tree is the one with the fewest steps.
<table>
<thead>
<tr>
<th>Character</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Outgroup – ancestral condition

(D) Number of shared derived character states

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>–</td>
<td>6</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>–</td>
<td>4</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>C</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>D</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

(F) Phylogeny inferred from (D)

Steps = 9
Problems with cladistics

<table>
<thead>
<tr>
<th>Number of Taxa</th>
<th>Number of possible trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>34,459,425</td>
</tr>
<tr>
<td>20</td>
<td>800,000,000,000,000,000,000</td>
</tr>
</tbody>
</table>

\[
\text{# of rooted trees} = \frac{(2n - 3)!}{2^{n-2}(n - 2)!}, \text{ for } n \geq 2
\]

Number of data points??
Cladistics reduces bias

- Two reasons why organisms may share characteristics?
  1. Character evolved in the ancestor and was inherited by descendents
  2. Convergent evolution – same character state evolved independently in different lineages

- If lineages evolved at the same constant rate
- If there traits have evolved independently,
- The phenogram would be the same as a cladogram
Data

• Validity of phylogenetic inference depends on quality and quantity of data

Morphological data

• Difficulty in defining characters
• Two features so closely correlated that they should be considered a single character
• Some characters missing for some taxa
  • Can’t be interpreted from fossils
  • Character lacking (e.g. number of wing veins in a wingless insect)
  • Continuous characters are more difficult than discrete (short, medium, long)
• What characters are homologous?
• Should characters be weighted differently?
Data

- Validity of phylogenetic inference depends on quality and quantity of data
  - Molecular data
    - What characters are homologous?
      - Gene duplication can lead to paralogous genes
      - Want to compare orthogonal genes – may be difficult if paralogous genes retain a similar sequence
    - Aligning sequence made difficult because of insertions and deletions
    - Should characters be weighted differently?
      - Equal weighting assumes equally informative
      - Condon weighting?
    - Transitions outnumber transversions 3:1 – use this weighting scheme?

  **Transition** is a mutation changing a purine to another purine nucleotide (A <-> G) or a pyrimidine to another pyrimidine nucleotide (C <-> T).

  **Transversion** is a mutation changing a purine for a pyrimidine or vice versa.
Problems with molecular data

Multiple hits
- Ancient divisions may have genetic signal erased by multiple hits
Problems with molecular data

Rapid divergence
• If changes occur rapidly, as is often the case with adaptive radiations, there might not be enough signal in the data to determine the order of events.
Genes evolve at different rates

**Slower**
- rbcL
- Lycopersicon
- Datura
- Physalis
- Nicandra
- Atropa
- Nolana
- Capsicum
- Lycium
- Juanulloa
- Solandra
- Mandragora
- Nicotiana tabacum
- Nicotiana acuminata
- Anthocercis
- Petunia
- Salpiglossis
- Schizanthus
- Ipomoea

**Faster**
- ndhF
- Lycopersicon
- Mandragora
- Juanulloa
- Solandra
- Nicandra
- Capsicum
- Datura
- Physalis
- Atropa
- Lycium
- Nolana
- Nicotiana tabacum
- Nicotiana acuminata
- Anthocercis
- Petunia
- Salpiglossis
- Schizanthus
- Ipomoea
Diverse approaches to tree building

Similarity/Difference Methods that use clustering algorithms

- Repeatable – always results in the same “best” tree
- Less computationally intensive – suitable for large # of taxa

**UPGMA**
- Assumes equal rate of evolution – simple clustering algorithm (average distances)

**Neighbor-joining method**
- Doesn’t assume equal rate of evolution – slightly more complex clustering algorithm

**Parsimony (Swofford et al. 1996)**
- Best tree is the one with the smallest number of character state changes

**Maximum likelihood (Felsenstein 1981)**
- Start with the rate at which you expect each nucleotide to be replaced
- Evaluates the likelihood for each possible phylogeny that the evolutionary model will yield the observed characters (sequences)
- The tree with the highest likelihood is the best estimate of true phylogeny

**Bayesian methods**
- Start with an a priori hypothesis (instead of testing against randomness)
- Calculates a probability that the hypothesis is true given the evidence that has been observed.
Methods

• In all cases, you have to make decisions
  • Codon weighting
  • Transition – transversion weighting
  • Allow reversals or not?
• Best result is if you obtain the same answer using multiple methods
Close correspondence between morphological and molecular tree

- Genera of sea urchins
- Parsimony analysis
- (A) 81 morphological characters
- (B) 380 bp sequence of 28S rRNA
- Same phylogenetic structure
- Molecular tree does not resolve two areas (unresolved bush – consensus tree)

**Morphology**

```
(A) 81 morphological characters

1. Cidaris

11(50)

2. Echinocardium

16

3. Spatangus

11(50)

4. Brissus

5

17

6. Arbacia

13

7. Echinus

12

Psammechinus

8

Paracentrotus

9

Lytechinus

10

Sphaerechinus

```

**DNA sequence**

```
(B) 380 bp sequence of 28S rRNA

99(96)

Echinocardium

Spatangus

Brissus


83(84)

Arbacia

Echinus

Psammechinus

Paracentrotus

Lytechinus

Sphaerechinus

```

Smith et al. 1992

Number of character state changes (Proportion of bootstrap trees that supported this clade)
Bootstrapping

- Remove one base-pair position and substitute in another so you have the same sample size (sampling with replacement)
- Do this 100 or 1000 or 10000 times
- Build the best phylogenetic tree for each of these bootstrap data sets
- Calculate the % of trees that support a particular branch point
- This is the bootstrap index of confidence

**Aligned nucleotide sequences for the malaria parasite Plasmodium**

<table>
<thead>
<tr>
<th>Site:</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>6</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Pre (Chimp)</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>Pme (Lizard)</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Pma (Human)</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Pfa (Human)</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>Pbe (Rodent)</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>Plo (Bird)</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Pfr (Monkey)</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>Pkn (Monkey)</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Pcy (Monkey)</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>Pv (Human)</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Pga (Bird)</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
</tr>
</tbody>
</table>

Efron B et al. PNAS 1996;93:13429-13429
Bootstrapping

- Remove one base-pair position and substitute in another so you have the same sample size (sampling with replacement)
- Do this 100 or 1000 or 10000 times
- Build the best phylogenetic tree for each of these bootstrap data sets
- Calculate the % of trees that support a particular branch point
- This is the bootstrap index of confidence

Efron B et al. PNAS 1996;93:13429-13429
Close correspondence between morphological and molecular tree

- Genera of sea urchins
- Parsimony analysis
- (A) 81 morphological characters
- (B) 380 bp sequence of 28s rRNA
- Same phylogenetic structure
- Molecular tree does not resolve two areas (unresolved bush)

Number of character state changes (Proportion of bootstrap trees that supported this clade)

Smith et al. 1992
Experimental validation of phylogenetic tree

- Compare to trees based on molecular and morphological data
- Compare to trees based on different genes
- Use computer simulation to test patterns of evolution
  - Allow computer-generated lineages to branch and their characters to change according to various models of evolutionary processes
    - Random change
    - Same average rate
    - Different rates
- Make tree
- Does it fit known phylogeny?
Experimental validation of phylogenetic tree

- Create a known phylogeny
- T7 bacteriophage culture exposed to mutagen
- Divided cultures (bifercate)
- Repeated three times over 300 generations
- Know the phylogenetic relationship because it was experimentally generated
- Used changes in restriction enzyme sites to build phylogeny
- Found the correct tree out of 135,350 possible bifurcating trees

Hillis et al. 1992
How can you use phylogenies?

• Do interacting species coevolve over time?
• Natural selection occurs during interactions
• Produces adaptation in both species
• Can result in cospeciation – simultaneous speciation
• Bacteria live inside specialized aphid cells called bacteriocytes
• Get nutrients and housing
• Bacteria produce amino acids aphids can’t produce
• Bacteria are passed to the next generation in eggs cell
• Aphids also attacked by parasitic wasps that lay their eggs in the host body
How can you use phylogenies?

- When aphid populations diverge to form distinct species, do their symbiotic bacteria speciate along with them?
- Alternative hypothesis is that parasitic wasps that insert their ovipositor into the bodies of their victims may move bacteria around from one host to another.
- In all cases but two, the branching pattern in the bacterial tree is the same as the aphid tree.
- Aphids and bacteria cospeciate.
How can you use phylogenies?

- Mapping characters onto phylogenies
- Does evolution proceed from generalist to a specialist or vise-versa?
- Seven known species of cowbirds that parasitize other species nests
- Some parasitize only a single species whereas others parasitize > 200 species
- What came first?
How can you use phylogenies?

- Mapping characters onto phylogenies
- Does evolution proceed from generalist to a specialist or vise-versa?
- Seven known species of cowbirds that parasitize other species nests
- Some parasitize only a single species whereas others parasitize > 200 species
- What came first?
Phylogenetics

- Resolve historical relationships among groups
- Improve our understanding of the process of evolution
- Test evolutionary hypotheses
  - Sequence of events
  - Timing of events
  - Pattern of change