Trees imply nested subsets, and vice versa
Rooted binary tree with labeled leaves (directed graph)

Graph with labeled leaves and internal vertices of degree 3 (undirected graph)

An unrooted tree can imply many possible rooted trees

3 unrooted trees and 15 rooted trees for 4 taxa:
Networks

• Network: unrooted tree, often including cycles.
  – Uses:
    • Initial step in formation of rooted tree if polarities are not hypothesized *a priori*.
    • Depiction of non-hierarchical structure within population, in which nodes represent observed individuals:
      – Minimum-spanning trees.
      – Nested clade analysis.
    • Depiction of reticulation:
      – Hybridization.
      – Lateral (horizontal) transfer.
      – Recombination.

South Park

Networks

• Network depicting a taxonomy of networks:

http://www.lirmm.fr/~gambette/RePhylogeneticNetworks.php
1. Evolution is “descent with modification”.
   - Species (living or fossil) are descended from ancestral species.
   - Sets of two (or more) observed taxa are descended from common ancestors.
   - The processes of evolution result in changes from ancestors to descendants.
     - Genetic, morphological, ecological, behavioral, etc.
   - Many process-level explanations have been proposed to account for this pattern.
     - Models, verbal and mathematical.
   - *Translate*: the tree model is adequate to describe patterns of evolutionary relationships.

2. The basic structure of a phylogenetic tree is hierarchical, with dichotomous (=binary, =bifurcated) branching.

Complications:
  - Reticulation (e.g., due to hybridization or horizontal transfer) is biologically possible, but difficult to fit uniquely to data.
    - *Translate*: binary branching is a simplifying assumption of the model.
  - Polychotomous (multifurcated) branching patterns are biologically possible, but can’t be distinguished from lack of resolution.
    - *Translate*: zero branch lengths might be real or artifactual.
Phylogenetic model: basic premises

3. We can use information on living species and fossils to infer the historical relationships among them.
   - *Translate*: we use the states of variables on observations to “optimally” fit the model of dichotomous branching.
   - Involves simultaneous estimation of the branching structure (tree topology) and the branch lengths.
   - *Homoplasy* is the residual variation: lack of fit of the data to the model.

Phylogenetic model: basic premises

4. We should make clear distinctions between:
   a) Trees as summaries of the data.
   b) The uses to which they are put:
      • Hypotheses and explanations about:
         – Evolutionary relationships (phylogenies).
         – Patterns of character evolution.
      
      – Aim is to avoid logical circularity.
This stuff should be easy

- The process of phylogenetic inference seems like it should be a simple, straightforward process.
- However, there are three basic problems:
  1. The biological problem: different characters do not necessarily produce the same trees.
  2. The first methodological problem: it’s unclear as to what optimality criterion and assumptions to use to fit the ‘best’ tree.
  3. The second methodological problem: finding the ‘best’ tree may be impossible.

1. The biological problem: different characters do not necessarily produce the same trees.
   - If all evolution were divergent, all characters would agree.
     - Same tree topology.
     - Not necessarily the same branch lengths.
   - Some characters become more similar in different species over time (convergence) or change in parallel over time (parallelism).
     - Translate: lack of fit of the model to the data (homoplasy).
   - Hybridization or horizontal transfer may produce reticulation rather than dichotomous evolution.
     - Translate: wrong model.
2. The *first methodological* problem: it’s unclear as to what optimality criterion to use to fit the ‘best’ tree.

• Spectrum of opinion about underlying assumptions. The extremes:
   (a) We should make as many realistic assumptions as possible.
      • The “best” phylogeny should incorporate as much as is known or surmised about evolutionary processes (*translate*: strong assumptions).
        – E.g., specific models of nucleotide substitution or developmental change.
      • Favored by molecular systematists.
      • Problem: if the assumptions are wrong, the fitted model is probably wrong.
   (b) We should make as few assumptions as possible.
      • Parsimony criterion (scientific sense).
      • The model is only as good as the assumptions.
      • The most “robust” phylogeny should be relatively independent of specific models of evolution.
        – Especially if the resulting tree is used to test independent biological hypotheses, to avoid logical circularity.
          » Rates or sequences of character evolution.
          » Biogeographic patterns.
          » Phenotypic distributions.
      • “Pattern cladists” (after C. Patterson):
        – Trees should assume nothing except the existence of hierarchy in nature.
          » Nested sets of objects: characters, taxa, etc.
      • *Translate*: weak assumptions.
   (c) Most biologists take a tenuous middle ground.
3. The second methodological problem: finding the ‘best’ tree may be impossible.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Number of rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>10,395</td>
</tr>
<tr>
<td>8</td>
<td>135,135</td>
</tr>
<tr>
<td>9</td>
<td>2,027,025</td>
</tr>
<tr>
<td>10</td>
<td>34,459,425</td>
</tr>
<tr>
<td>11</td>
<td>654,729,075</td>
</tr>
<tr>
<td>12</td>
<td>13,749,310,575</td>
</tr>
<tr>
<td>13</td>
<td>316,234,143,225</td>
</tr>
<tr>
<td>14</td>
<td>7,905,853,580,625</td>
</tr>
<tr>
<td>15</td>
<td>213,458,046,676,875</td>
</tr>
<tr>
<td>16</td>
<td>6,190,283,353,629,370</td>
</tr>
<tr>
<td>17</td>
<td>191,898,783,962,510,000</td>
</tr>
<tr>
<td>18</td>
<td>6,332,659,870,762,850,000</td>
</tr>
<tr>
<td>19</td>
<td>221,643,095,476,699,000,000</td>
</tr>
<tr>
<td>20</td>
<td>8,200,794,532,637,890,000,000</td>
</tr>
</tbody>
</table>

Age of universe ≈ 13.7×10⁹ yrs.
If scan 10⁷ trees/sec, could scan almost all trees for 22 taxa in that time (=1.3×10²⁵).

For 52 taxa, number of trees = 2.8×10⁸⁰.

Optimization
• Minimization of a continuous function of a parameter:
  – E.g., regression function.
Optimization

- Example of 1-parameter model: optimal slope of line through center of data.
  - Objective function: sum of absolute orthogonal residuals (least-absolute regression).

Optimization

- Example of 1-parameter model: optimal slope of line through center of data.
  - Objective function: sum of squared orthogonal residuals (major-axis regression).
Optimization

• Minimization of a continuous function of two parameters:

Optimization

• Analogy with Sewell Wright’s concept of an adaptive landscape:
Optimization

• Example of 2-parameter model: least-squares regression.
  – Objective function: sum of squared vertical residuals.

  \[ Y = b_0 + b_1 X \]

  \[ b_0 = 4.60 \quad b_1 = 0.96 \]

Optimization

• Minimization of a discrete function of a parameter:
  – E.g., tree.
Optimization

- Minimization of a discrete function of two parameters:

![3D plot of objective function value against Parameter 1 and Parameter 2](image)

Optimization

- Distribution of some near-optimal trees in Bayesian tree space:

![Images of trees](image)

Hillis et al. 2005
Optimization

• Algorithmic complexity: how long does it take to solve a problem?
  (1) P: problems can be solved in polynomial time:
    • Can get in the answer in at most $cn^k$ steps, where $n$ is a measure of the ‘size’ of the problem.
    • Large problems are not too difficult.
    • E.g., solving a set of simultaneous equations.
  (2) N-P: problems can be guessed in polynomial time and checked in polynomial time:
    • Can get in the answer in at most $(cn^k)(cn^k)$ steps.
    • Large problems are difficult.
    • E.g., given a set of integers, does some non-empty subset of them sum to zero?
  (3) N-P complete: problems can be solved only in exponential time:
    • Can get the answer in at most $ck^n$ steps.
    • Large problems are intractable, and most can be solved only by complete enumeration.
    • E.g., Steiner (travelling-salesman) problem, and shortest-tree problem.
    • Some heuristics (e.g., branch-and-bound) can be used to speed-up solutions.

Taxonomy of phylogenetic methods

• Two kinds of similarity: phenetic and cladistic.
• Two kinds of methods: algorithmic and combinatoric.
• Five basic numerical approaches:
  (a) Cluster analysis: phenetic, algorithmic.
  (b) Additive trees: phenetic, algorithmic.
  (c) Minimum evolution: cladistic, combinatoric.
  (d) Maximum likelihood: phenetic or cladistic, algorithmic or combinatoric.
  (e) Bayesian ‘likelihood’: phenetic or cladistic, algorithmic or combinatoric.
Two kinds of similarity
Two fundamentally different philosophical approaches, based solely on different concepts of "similarity":

(1) Phenetic: based on overall similarity.
   - Uses matrix of pairwise distances among taxa.
   - Problem: distance measures can't distinguish divergence (model) from convergence (residual variation).

(2) Cladistic: based on derived similarity.
   - Similarity based on derived states since most recent common ancestor.
   - Uses character matrix.
   - Problem: identification of ancestral vs. derived character states presumes knowledge of relationships (the "polarity" problem).
     » Logically circular.
   - Extrinsic criteria must be used:
     » Prior: developmental or paleontological precedence, etc.
     » Posterior: use of outgroup to root the tree.

Two classes of methods

(1) Algorithmic:
   - Formulate an algorithm (set of rules) for generating a tree from a data matrix.
   - Play out the algorithm (exact or iterative).
   - Global optimality criterion: maximum-likelihood estimation.
   - Local optimality criterion:
     • Additive and ultrametric clustering methods.
   - Problems:
     • The rules for generating the tree might not have any biological/evolutionary basis.
     • Even if the optimality criterion is global, any particular solution might be only locally optimal.
Two classes of methods

(2) Combinatoric:
- Formulate a quantitative criterion for choosing the "best" tree (objective function).
- Compare all possible trees (or a reasonable subset), giving each tree a score based on the criterion.
- Choose the tree (or possibly set of trees) having the best score.
- *Problem*: the number of possible trees is enormous.

Five basic numerical approaches applied in the literature

1. Cluster analysis: ultrametric hierarchical agglomerative clustering (phenetic)
- Object: find a unique tree (dendrogram) based on:
  - A measure of distance (dissimilarity):
    - Euclidean, Manhattan, Mahalanobis, correlation complement, etc.
  - A clustering criterion:
    - UPGMA (unweighted pair grouping)
    - WPGMA (weighted pair grouping)
    - Single linkage (Prim network)
    - Complete linkage (farthest neighbors)
    - Nearest neighbor
    - Minimum variance (Ward's method)
• Cluster analysis
  – Advantages:
    • Provides a unique solution, given the distance measure and clustering criterion.
    • Recursive and fast.
  – Disadvantages:
    • Unique solution not globally optimal in any sense.
    • Phenetic: based on pairwise measures of overall difference (overall similarity).
    • No biological rationales for choosing among the infinite numbers of different distance measures and clustering criteria, all combinations of which produce different results.
    • Most methods assume ultrametric pairwise distances (3-point condition).
      – All terminal taxa equidistant from nodes.
      – Assumes equal mean rates of evolution.
    • Most methods are special cases of a general 5-parameter model and are therefore arbitrary.
    • Provide unique trees even for completely random data.
Basic problem with cluster analysis

- Six species of fishes.
- Three different measures of similarity.
- Three different assessments of relationship.

Lance and Williams (1967) ‘flexible’ model
2. Additive trees (phenetic)
   – Object: find a unique tree (dendrogram) based on:
     • A measure of distance (dissimilarity).
       – Euclidean, Mahalanobis, etc. (continuous characters).
       – Manhattan, Hemming, percent dissimilarity, etc. (discrete characters).
     • A clustering criterion; e.g.,
       – Neighbor-joining (Saitou-Nei)
       – Distance-Wagner algorithm (Farris)
       – Fitch-Margoliash algorithm
       – Lake’s criterion
       – Least squares (optimization criterion)

- Advantages:
  - Unique solution, given the distance measure and clustering criterion.
  - Recursive and fairly fast.
  - Relaxes assumption of equal rates of evolution.
  - Can be rerooted for a specified outgroup.

- Disadvantages:
  - Not globally optimal in any sense (except least-squares).
  - Based on pairwise measures of overall difference (phenetic).
  - No biological rationales for choosing among different distance measures and clustering criteria.
  - Produces trees for random data.

**Ultrametric and additive trees for same data**

![Ultrametric and additive trees for same data](image)
3. Minimum evolution (=‘parsimony’) (cladistic)

- Object: find the shortest tree consistent with the data:
  - Cladogram, for discrete characters.
    - Minimize the number of "ad hoc" hypotheses about character-state changes.
  - Steiner minimum tree, for continuous characters.
    - Minimize total tree length.

- Versions of the discrete minimum-evolution model:
  - Fitch parsimony (=strict parsimony): character states are unordered and reversible.
  - Wagner parsimony: character states are ordered but reversible.
  - Dollo parsimony: synapomorphies assumed to be uniquely derived, and can reverse only once (reversals are minimized and convergence is prohibited).
  - Camin-Sokal parsimony: character states are irreversible (because reversals are ‘really’ new apomorphies).
• **Minimum-evolution:**
  - **Advantages:**
    • Assumes only minimal character-state change as a criterion for choosing among trees.
  - **Disadvantages:**
    • For discrete characters, usually produces sets of equally short ("parsimonious") trees.
      - Can differ markedly in topology and branch lengths.
    • Consensus methods.
    • When extended to use as a phylogenetic hypothesis (evolutionary tree), implies minimal evolution.
  • NP-complete combinatorial problem.
    - Only way to find shortest tree is to examine all possible trees.
      » For small numbers of taxa (<20 or so), might find the globally optimal solution by brute force.
    - Some methods examine an intelligent subset of all possible trees:
      » E.g., branch and bound (<30 or so taxa).
    - Heuristics not guaranteed to be globally optimal.

4. **Maximum likelihood (phenetic or cladistic)**
  - **Object:** find the tree (topology + branch lengths) having the greatest probability of giving rise to the observed data.
    • Maximizes $Pr(\text{data | model}) = Pr(\text{data | [tree + assumptions]})$.
    • Based on an explicit model of evolution:
      - Specifies the "rules" and probabilities of character-state change over time.
  - **Advantages:**
    • Provides a unique parametric solution that is globally optimal.
    • Provides standard errors and related measures to assess the statistical goodness of fit.
    • Allows for the statistical comparison of different trees for the same data.
  - **Disadvantages:**
    • Explicit assumptions. E.g., for continuous characters:
      - Random-walk (Brownian) model of evolution.
      - Deviations along branches normally distributed with constant variance (homoscedasticity).
    • Computationally expensive.
Some character-state graphs and corresponding transition matrices

Swofford and Maddison
1992

Some maximum-likelihood trees (rooted and unrooted)
5. Bayesian likelihood (phenetic or cladistic)

- **Object:** find the tree (topology + branch lengths) having the greatest probability of being true, given the observed data.
  - Maximizes $Pr(\text{model} | \text{data}) = Pr([\text{tree} + \text{assumptions}] | \text{data})$.
  - Based on an explicit model of evolution:
    - Specifies distributions of “rules” and probabilities of character-state change over time.
  - Provides a consensus of a large number of highly likely trees.

- **Advantages:**
  - Parametric solution is close to globally optimal.
  - Provides posterior probabilities and related measures to assess the statistical goodness of fit.

- **Disadvantages:**
  - Explicit assumptions.
  - Computationally expensive: Markov-chain Monte Carlo (MCMC) methods.

<table>
<thead>
<tr>
<th>Approach</th>
<th>1. Phenetic</th>
<th>2. Cladistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity concept:</td>
<td>Overall similarity</td>
<td>Derived similarity</td>
</tr>
<tr>
<td>Evidence supporting groups:</td>
<td>Degree of similarity</td>
<td>Character changes leading up to the most recent common ancestor</td>
</tr>
<tr>
<td>Basic data</td>
<td>Pairwise distances among taxa</td>
<td></td>
</tr>
<tr>
<td>Kinds of characters</td>
<td>Discrete (nominal, ordinal)</td>
<td></td>
</tr>
<tr>
<td>Numerical methods</td>
<td>Cluster analysis</td>
<td></td>
</tr>
<tr>
<td>Character-state changes</td>
<td>Secondarily mapped onto tree</td>
<td></td>
</tr>
<tr>
<td>Primarily used by</td>
<td>Morphological systematists (until the past 10–15 years)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Cladistic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian inference</td>
<td></td>
</tr>
<tr>
<td>Combinatorics</td>
<td></td>
</tr>
<tr>
<td>Graph theory</td>
<td></td>
</tr>
</tbody>
</table>

Other characters can be secondarily mapped onto the tree.