Phylogenetics

BIOL 7711
Computational Bioscience

Biochemistry and Molecular Genetics
Computational Bioscience Program
Consortium for Comparative Genomics
University of Colorado School of Medicine

David.Pollock@uchsc.edu
www.EvolutionaryGenomics.com
Reconstructing Phylogenies

- All species are related by descent
- Splitting and divergence
- Homologous genes are also related by descent
- The goal of phylogenetic reconstruction is to determine the order and timing of splitting events
- Phylogenetic analysis includes inferences of substitution (mutation, selection) processes, ancestral states and functions
Why Phylogenetics?

- Resolve evolutionary history
  - Important for comparative analysis to account for correlations due to relatedness
- Disease origins, paths of infection
  - Influenza, HIV
- Origin of genes, systems, functions
Data Types and Issues

- Morphology
  - Continuous or discrete
  - Polarized?
- DNA and Protein Sequence
  - Probabilistic
  - Saturation
- Protein and RNA structure
- Transposable element insertion
- The question of homology
Morphology vs. Molecular

“In analyzing phenotypic features, we do not know which part and how much of the genetic base is being analyzed, and hence cannot know about independence”

Pleiotropy
Multifactoriality
Epigenetic effects (environment)
Homology
  Underlying genes may change!
Parts of a Tree

- Branch, Node
- Edge, vertex
- Tips (sequence)
- Root
- Topology, $T_i$
- Geneology, $G_i$

\[( (A, B), (C, D) ) ; ( (A:0.23, B:0.35):1.28, (C:0.13, D:0.19) :1.52 ) ; \]
The Phylogeny Problem

- Data, Trees, Model of evolution
- Assumptions
  - Data tree reflects species tree
  - Branch lengths reflect time?
  - Homologous data (common ancestor)
The Problem
The Problem

- Oldest group (rooting) is the hardest
- If oldest split is known, use as outgroup to define the rest of the tree
  - Hard to be sure
  - If sure, may not be very informative
Unrooted Tree

- Ignore the biggest problem
- Reversible models are common anyway (root doesn’t matter)
Topology Space

\[
\prod_{i=3}^{T} (2i - 5)
\]

Times branch lengths

Need tree search algorithm
Topology Space

- 11 taxa => 34 million
  - More money than most will make in lifetime
- 13-14 taxa
  - Bill Gates’ wealth
  - Federal deficit
- 24 taxa
  - ~ Number of atoms in a mole
- 30 taxa ~10^{37}
  - # atoms in the solar system?
- 50 taxa ~10^{74} a very big number
Heuristic, Optimal, Posterior

- NP hard, so need tricks
- Distance
  - Estimate of amount of change separating two sequences (species)
  - Calculate analytically (limited), or ML
  - Requires a reversible model of evolution
- Parsimony
  - Minimal number of changes
  - Poorly specified model (but there is one)
  - Easy to calculate
- ML, Bayes
  - Model based, don’t toss the data
Distance Reconstruction

- **UPGMA**
  - Pair closest sequences first
  - Doesn’t account for rate variation (ultrametric)

- **Neighbor Joining (NJ)**
  - Closest but least far from everything else
  - Deals with rate variation

- **BIONJ, WEIGHBOR**
  - Least squares (inverse square) weighting of neighbor information
Parsimony and Cladism

- Find the tree that minimizes the number of changes along the tree
- Can be calculated in polynomial time
- Exact or Branch and Bounds practical up to about 30 species
  - Not very many these days
  - After that, it is heuristic
- Slower than distance, faster than ML
Parsimony and Cladism

- Note: seen as a model of change, it is a complex and awkward hypothesis
- Poor reconstruction of ancestral states
- Has ancestral and derived states, polarity
- Monophyletic, paraphyletic, polyphyletic
  - Highly dependent on time directionality
- Can an organism start a new group?
  - Consider enzyme function
Paraphyly, Rooted Tree

Major Adaptive Shift
Functional Divergence?

Birds
- Penguin
- Wren

Reptiles
- Crocodile
- Alligator
- Iguana
- Snapping Turtle
- Corn Snake
Tree Search

- Exhaustive
- Branch and bounds
  - Abandon routes that are less likely than a route we know is pretty good
- Stepwise addition
  - Best addition point is found as taxa are added
- Star decomposition
Star Decomposition

- Separate out all sets of two taxon groups,
- choose the one that shows the greatest score improvement
Hill Climbing and Simulated Annealing

- Use with any measure of the “goodness” of a solution
- Accept if

\[ z(t + 1) \geq z(t) \]

- Or

\[ P(\text{accept}) = e^{-k[z(t+1)-z(t)]} \]

- Where \( k \) varies over time, usually larger
- Use “tabu list” of recently attempted solutions
Branch Swapping, etc

- Exchange two branches for each other
- Nearest Neighbor interchange
- Subtree pruning and regrafting
- Tree bisection and reconnection

These work surprisingly well
Likelihood Calculation

- Conditional likelihood for each site, j, for a node, A

  - Likelihood of each state (e.g., nucleotide g) given the two descendent nodes, e.g. B and C and connecting branches \( b_{AB} \) and \( b_{AC} \)

\[
L(x_{Aj} = g) = \left[ \sum_k P_{gk}(b_{AB})(L(x_{Bj} = k)) \right] \times \left[ \sum_l P_{gl}(b_{AC})(L(x_{Cj} = l)) \right]
\]

Left node \hspace{2cm} Right node
Likelihood Calculation

- Infinite stack of turtles, except that tips are data, and terminate stack
- At root, multiply each state by its prior (the equilibrium frequencies) and sum over all ancestral states

$$L(j) = \sum_{m=1}^{\text{#states}} \pi_m \times L_{\text{Root}, j}(x = m)$$

- If reversible, root can be anywhere, including tips
Trees and HMMs

- Can combine the two
- Goldman, Thorne, and Jones
  - Secondary structure
  - Buried versus exposed
  - Transition probabilities between structural types
  - Training set of proteins with known structures
  - Use to predict secondary structure, protein features
Newer Tricks

- Bayesian
  - Problem isn’t multiplicative with every parameter
  - Augmented data at roots or specified changes
Modeling Evolutionary Processes

- State frequencies, transitions versus transversions, rates, context dependency

- Model complexity
  - Takes longer time to calculate with every parameter
  - Check if complex model is justified given the amount of data (and computer time available)
  - Compare ML or BF of models, model testing
<table>
<thead>
<tr>
<th>( \text{To} )</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{From} )</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\text{To}</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>----------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>\text{From}</td>
<td>1-\Sigma</td>
<td></td>
<td>1-\Sigma</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1-\Sigma</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>1-\Sigma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td>1-\Sigma</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td>1-\Sigma</td>
</tr>
</tbody>
</table>
Jukes and Cantor, 1969

<table>
<thead>
<tr>
<th>( \text{To} )</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1-( \Sigma )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
</tr>
<tr>
<td>C</td>
<td>( \lambda )</td>
<td>1-( \Sigma )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
</tr>
<tr>
<td>G</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>1-( \Sigma )</td>
<td>( \lambda )</td>
</tr>
<tr>
<td>T</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>1-( \Sigma )</td>
</tr>
<tr>
<td>\text{To}</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>A</td>
<td>1-\Sigma</td>
<td>\beta</td>
<td>\alpha</td>
<td>\beta</td>
</tr>
<tr>
<td>C</td>
<td>\beta</td>
<td>1-\Sigma</td>
<td>\beta</td>
<td>\alpha</td>
</tr>
<tr>
<td>G</td>
<td>\alpha</td>
<td>\beta</td>
<td>1-\Sigma</td>
<td>\beta</td>
</tr>
<tr>
<td>T</td>
<td>\beta</td>
<td>\alpha</td>
<td>\beta</td>
<td>1-\Sigma</td>
</tr>
</tbody>
</table>

From Kimura, 1980
<table>
<thead>
<tr>
<th>( \text{To} )</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1-(\Sigma)</td>
<td>(\pi_C)</td>
<td>(\pi_G)</td>
<td>(\pi_T)</td>
</tr>
<tr>
<td>C</td>
<td>(\pi_A)</td>
<td>1-(\Sigma)</td>
<td>(\pi_G)</td>
<td>(\pi_T)</td>
</tr>
<tr>
<td>G</td>
<td>(\pi_A)</td>
<td>(\pi_C)</td>
<td>1-(\Sigma)</td>
<td>(\pi_T)</td>
</tr>
<tr>
<td>T</td>
<td>(\pi_A)</td>
<td>(\pi_C)</td>
<td>(\pi_G)</td>
<td>1-(\Sigma)</td>
</tr>
</tbody>
</table>

From Felsenstein, 1981
<table>
<thead>
<tr>
<th>( \text{From} )</th>
<th>( \text{A} )</th>
<th>( \text{C} )</th>
<th>( \text{G} )</th>
<th>( \text{T} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{A} )</td>
<td>1-( \Sigma )</td>
<td>( \lambda_{AC} \pi_C )</td>
<td>( \lambda_{AG} \pi_G )</td>
<td>( \lambda_{AT} \pi_T )</td>
</tr>
<tr>
<td>( \text{C} )</td>
<td>( \lambda_{CA} \pi_A )</td>
<td>1-( \Sigma )</td>
<td>( \lambda_{CG} \pi_G )</td>
<td>( \lambda_{CT} \pi_T )</td>
</tr>
<tr>
<td>( \text{G} )</td>
<td>( \lambda_{GA} \pi_A )</td>
<td>( \lambda_{GC} \pi_C )</td>
<td>1-( \Sigma )</td>
<td>( \lambda_{GT} \pi_T )</td>
</tr>
<tr>
<td>( \text{T} )</td>
<td>( \lambda_{TA} \pi_A )</td>
<td>( \lambda_{GC} \pi_C )</td>
<td>( \lambda_{TG} \pi_G )</td>
<td>1-( \Sigma )</td>
</tr>
</tbody>
</table>
Progression of Complexity

- Combining models
- Rate variation
  - Gamma on average rate, Yang 1995
    - Freqs, relative rate parameters the same
- Codon models
  - DNA model underneath amino acid model
  - AAs change slower: constraint, selection
  - Physicochemical, PAM & JTT, context
  - Computationally expensive
- IID => dependent
- Mixture models
Statistical Desiridata

- **Identifiability**
  - A set of trees are identifiable under a model if the distributions of characters are disjoint for different tree topologies (with trivial exceptions, like 0 branch lengths)

- **Consistency**
  - A method is consistent if it will recover the true tree with arbitrarily high probability given enough data (i.e. a long enough sequence)
Dating Times

- Need to assume (or validate) an approximate molecular clock
  - Underlying rate constant over time
- Need paleontological data to link branching patterns to geological time
- Can have continuum from perfect clock to complete freedom of rate on every branch
Gene Tree versus Species Tree

- Not necessarily the same thing
- Sampling from coalescent at speciation
- Recombination
  - Different genes, gene regions sample the coalescent differently
- Horizontal gene transfer
- Non-neutral processes, e.g., convergence
- Do species always bifurcate?
  - Slow decrease in gene flow
  - Inversions prevent recombination
  - Species may form slowly