An Introduction to Phylogenetics

Bret Larget
bret.larget@wisc.edu

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 7, 2017
Outline I

1. Introduction
2. Example
3. Trees
4. Models of Molecular Evolution
Outline

1 Introduction
   - The Big Question
   - Some Modern Uses of Phylogenies
   - Phylogenetics and Forensics
The Big Questions

- How much of the history of life on Earth can we reconstruct from the traces left in the genomes of all living organisms today?
- What are major features in the Tree of Life?
- How does phylogenetics relate to other problems in science?
Modern Phylogenetics

- Phylogenies are usually estimated from aligned DNA sequence data.
- Phylogenetics is the primary tool for systematics.
- Phylogenetics is used for studying viruses such as HIV and Influenza.
- Phylogenetics has been used in court for forensic purposes.
- Phylogenetics is being used increasingly in comparative genomics and study of gene function.
Phylogenetics and Systematics

- Phylogenetic methods, particularly for molecular sequence data, have become the primary tool for systematicists to determine evolutionary relationships.
- These tools have been used to confirm expected relationships — for example, that chimpanzees are the closest living relative to humans — and have also been key in revealing several more surprising findings, including:
  - birds are descended from dinosaurs;
  - polar bears form a monophyletic group within brown bears;
  - the most closely related land mammal to whales is the hippopotamus.
Phylogenetic Tree of Whales
Phylogenetics and Forensics

- Phylogenetic trees have been used in several instances in the courts to provide evidence about the likely transmission of HIV.

Examples include:
- Confirming that a nurse contracted HIV from mishap with a broken glass blood collection tube from an infected patient and not from an alternative source;
- Providing evidence of deliberate infection in a criminal case;
- Indicated that an infected friend was likely not the direct source of infection in a case.
Forensic Phylogenetic Tree
2 Example
   Phylogeny of Birds
DNA Data from a Sample of Birds

First 24 bases of 1558 from Cox I gene.

<table>
<thead>
<tr>
<th>Animal</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alligator</td>
<td>GTG AAC TTC CAC --- CGT TGA CTC</td>
</tr>
<tr>
<td>Emu</td>
<td>GTG ACA TTC ATT ACT CGA TGA TTT</td>
</tr>
<tr>
<td>Kiwi</td>
<td>GTG ACC TTC ATT ACT CGA TGA CTC</td>
</tr>
<tr>
<td>Ostrich</td>
<td>GTG ACC TTC ATT ACT CGA TGA CTT</td>
</tr>
<tr>
<td>Swan</td>
<td>GTG ACC TTC ATC AAC CGA TGA CTA</td>
</tr>
<tr>
<td>Goose</td>
<td>GTG ACC TTC ATC AAC CGA TGA CTA</td>
</tr>
<tr>
<td>Chicken</td>
<td>GTG ACC TTC ATC AAC CGA TGA TTA</td>
</tr>
<tr>
<td>Woodpecker</td>
<td>GTG ACC TTC ATC AAC CGA TGA TTA</td>
</tr>
<tr>
<td>Finch</td>
<td>ATG ACA TAC ATT AAC CGA TGA TTA</td>
</tr>
<tr>
<td>Ibis</td>
<td>GTG ACC TTC ATC AAC CGA TGA CTA</td>
</tr>
<tr>
<td>Stork</td>
<td>GTG ACC TTC ATT ACC CGA TGA CTA</td>
</tr>
<tr>
<td>Osprey</td>
<td>ATG ACA TTC ATC AAC CGA TGA CTA</td>
</tr>
<tr>
<td>Falcon</td>
<td>GTG ACC TTC ATC AAC CGA TGA CTA</td>
</tr>
<tr>
<td>Vulture</td>
<td>ATG ACA TTC ATC AAT CGA TGA CTA</td>
</tr>
<tr>
<td>Penguin</td>
<td>GTG ACC TTC ATT AAC CGA TGA CTA</td>
</tr>
</tbody>
</table>
An Estimated Phylogeny

- Alligator
- Emu
- Kiwi
- Ostrich
- Swan
- Goose
- Chicken
- Falcon
- Finch
- Osprey
- Woodpecker
- Ibis
- Stork
- Vulture
- Penguin
- Swan
- Goose
- Ostrich
- Kiwi
- Emu
- Alligator
Outline

3 Trees
- Phylogeny Basics
- Unrooted Trees
- Counting Trees
Activity 1: Example Tree

- How many descendent taxa does the common ancestor of taxa A and C have?
- Which taxon is sister to A?
- Which taxa are more closely related, A and C or C and D?
- Which taxa are more closely related, A and E or D and E?
Activity 2: Compare Trees

Which trees have the same tree topology?
Activity 3: Unrooted Trees

- Some methods estimate unrooted trees.
- If $C$ is the outgroup, what is the rooted tree topology?
- If taxon $C$ is the outgroup, which node is sister to $B$?
- If taxon $A$ is the outgroup, which node is sister to $B$?
- How many rooted tree topologies are consistent with this unrooted tree topology?
# How Many Trees?

<table>
<thead>
<tr>
<th># of Taxa</th>
<th># Unrooted Trees</th>
<th># Rooted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
<td>10395</td>
</tr>
<tr>
<td>8</td>
<td>10,395</td>
<td>135,135</td>
</tr>
<tr>
<td>9</td>
<td>135,135</td>
<td>2,027,025</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
<td>34,459,425</td>
</tr>
<tr>
<td>11</td>
<td>34,459,425</td>
<td>654,729,075</td>
</tr>
<tr>
<td>12</td>
<td>654,729,075</td>
<td>13,749,310,575</td>
</tr>
<tr>
<td>13</td>
<td>13,749,310,575</td>
<td>316,234,143,225</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>52</td>
<td>&gt; # of atoms in universe</td>
<td></td>
</tr>
</tbody>
</table>
Formula for Counting Trees

- The number of rooted tree topologies with \( n \) taxa is 
  \[ 1 \times 3 \times \cdots \times (2n - 3) \equiv (2n - 3)!! \text{ for } n \geq 3. \]

- There are more rooted trees with 51 species \( (2.7 \times 10^{78}) \) than estimated # of hydrogen atoms in the universe \( (1.3 \times 10^{77}) \).

- Biologists often estimate trees with more than 100 species.
Outline

4 Models of Molecular Evolution
  • Continuous-time Markov Chains
Probabilistic Framework

Essentially, all models are wrong, but some are useful.

George Box

- Commonly used models of molecular evolution treat sites as independent.
- These common models just need to describe the substitutions among four bases — A, C, G, and T — at a single site over time.
- The substitution process is modeled as a continuous-time Markov chain.
Markov Property

- Use the notation \( X(t) \) to represent the base at time \( t \).
- \( X(t) \in \{ A, C, G, T \} \) for DNA.
- Formal statement:

  \[
  P \{ X(s + t) = j \mid X(s) = i, X(u) = x(u) \text{ for } u < s \} \\
  = P \{ X(s + t) = j \mid X(s) = i \}
  \]

- Informal understanding: given the present, the past is independent of the future.
- If the expression does not depend on the time \( s \), the Markov process is called homogeneous.
A stationary, homogeneous, continuous-time, finite-state-space Markov chain is parameterized by a rate matrix where:

- off-diagonal rates are nonnegative;
- diagonal terms are negative row sums of off-diagonal elements;
- consequently, row sums are zero.

Example:

\[
Q = \{q_{ij}\} = \begin{pmatrix}
-1.1 & 0.3 & 0.6 & 0.2 \\
0.2 & -1.1 & 0.3 & 0.6 \\
0.4 & 0.3 & -0.9 & 0.2 \\
0.2 & 0.9 & 0.3 & -1.4 \\
\end{pmatrix}
\]
How to simulate a continuous-time Markov chain beginning in state \( i \).

- time to the next transition \( \sim \text{Exponential}(q_i) \) where \( q_i \equiv -q_{ii} \).
- transition is to state \( j \) with probability

\[
\frac{q_{ij}}{\sum_{k \neq i} q_{ik}} = \frac{q_{ij}}{q_i}
\]
The transition matrix is \( P(t) = e^{Qt} \) where

\[
e^{A} = \sum_{k=0}^{\infty} \frac{A^k}{k!} = I + A + \frac{A^2}{2} + \frac{A^3}{6} + \cdots
\]

A probability transition matrix has non-negative values and each row sums to one.

Each row contains the probabilities from a probability distribution on the possible states of the Markov process.
Examples

\[ P(0.1) = \begin{pmatrix} 0.897 & 0.029 & 0.055 & 0.019 \\ 0.019 & 0.899 & 0.029 & 0.053 \\ 0.037 & 0.029 & 0.916 & 0.019 \\ 0.019 & 0.080 & 0.029 & 0.872 \end{pmatrix} \]

\[ P(0.5) = \begin{pmatrix} 0.605 & 0.118 & 0.199 & 0.079 \\ 0.079 & 0.629 & 0.118 & 0.174 \\ 0.132 & 0.118 & 0.671 & 0.079 \\ 0.079 & 0.261 & 0.118 & 0.542 \end{pmatrix} \]

\[ P(1) = \begin{pmatrix} 0.407 & 0.190 & 0.276 & 0.126 \\ 0.126 & 0.464 & 0.190 & 0.219 \\ 0.184 & 0.190 & 0.500 & 0.126 \\ 0.126 & 0.329 & 0.190 & 0.355 \end{pmatrix} \]

\[ P(10) = \begin{pmatrix} 0.200 & 0.300 & 0.300 & 0.200 \\ 0.200 & 0.300 & 0.300 & 0.200 \\ 0.200 & 0.300 & 0.300 & 0.200 \\ 0.200 & 0.300 & 0.300 & 0.200 \end{pmatrix} \]
Well-behaved continuous-time Markov chains have a stationary distribution $\pi$. (For finite-state-space chains, irreducibility is sufficient.)

When the time $t$ is large enough, the probability $P_{ij}(t)$ will be close to $\pi_j$ for each $i$. (See $P(10)$ from earlier.)

The stationary distribution can be thought of as a long-run average — the proportion of time the state spends in state $i$ converges to $\pi_i$.

The stationary distribution satisfies $\pi^\top Q = 0^\top$.

Also, $\pi^\top P(t) = \pi^\top$ for any time $t$. 
Numerical Example

\[ \pi^\top Q = 0^\top \]

\[
\begin{pmatrix}
0.2 & 0.3 & 0.3 & 0.2 \\
0.2 & -1.1 & 0.3 & 0.6 \\
0.4 & 0.3 & -0.9 & 0.2 \\
0.2 & 0.9 & 0.3 & -1.4
\end{pmatrix}
\begin{pmatrix}
-1.1 & 0.3 & 0.6 & 0.2 \\
0.2 & -1.1 & 0.3 & 0.6 \\
0.4 & 0.3 & -0.9 & 0.2 \\
0.2 & 0.9 & 0.3 & -1.4
\end{pmatrix}
= \begin{pmatrix}
0 & 0 & 0 & 0
\end{pmatrix}
\]
Usual Parameterization

- The matrix $Q = \{q_{ij}\}$ is typically scaled and parameterized

$$q_{ij} = \frac{r_{ij}\pi_j}{\mu}$$

for $i \neq j$ where

$$\mu = \sum_i \pi_i \sum_{j \neq i} r_{ij}\pi_j$$

which guarantees that $\pi$ will be the stationary distribution when $r_{ij} = r_{ji}$.

- With this scaling, there is one expected transition per unit time.
A continuous-time Markov chain is time-reversible if the probability of a sequence of events is the same going forward as it is going backwards.

The matrix $Q$ is the matrix for a time-reversible Markov chain when $\pi_i q_{ij} = \pi_j q_{ji}$ for all $i$ and $j$.

That is, the overall rate of substitutions from $i$ to $j$ equals the overall rate of substitutions from $j$ to $i$ for every pair of states $i$ and $j$.

The matrix equivalent is $\Pi Q = Q^\top \Pi$ where $\Pi = \text{diag}(\pi)$.
General Time-Reversible Model

- The GTR model is the most general basic time-reversible continuous-time Markov model for nucleotide substitution.

- The model is typically parameterized with 8 free parameters where

$$q_{ij} = \begin{cases} r_{ij} \pi_j / \mu & \text{for } i \neq j \\ - \sum_{j \neq i} q_{ij} & \text{for } i = j \end{cases}$$

with $\mu = \sum_i \pi_i \sum_{j \neq i} r_{ij} \pi_j$.

- The stationary distribution $\pi$ has three free parameters as $\pi$ sums to one;

- The vector $r = (r_{AC}, r_{AG}, \ldots, r_{GT})$ is usually constrained to five degrees of freedom (either by setting $r_{GT} = 1$ or constraining the sum).

- Many other popular models are special cases.

- These models are often named by the initials of the authors and the year in which they were published.
Rate Variation Among Sites

- A common extension to the standard CTMC models is to assume that there is rate variation among sites.
- At these sites, the $Q$ matrix is multiplied by a site-specific rate.
- The two most popular extensions are:
  - Invariant sites: some sites have rate 0
  - Gamma-distributed rates: rates are drawn from a mean 1 gamma distribution
- For computational tractability, the Gamma distribution is typically replaced by a mean 1 discrete distribution with four distinct rates based on quantiles of a Gamma distribution.
Other Extensions

- There are many other model extensions in common use and under development.
- It is common to partition sites (by gene, by codon position, by genomic location) and to use different models for each part.
- The **covarion** model allows different lineages to have different rates at the same site.
- This is typically modeled with a hidden Markov model where the site can turn “off”.
- There are models for amino acid substitution, models for codons, models for RNA pairs, models that incorporate protein structure information, and so on.
- Current models still do not capture much of the important biological processes that affect evolution of molecular sequences.