Phylogenetic and comparative methods

Nicolas Salamin
nicolas.salamin@unil.ch; tel: 4154; office: 3212

Martha Serrano (martha.serrano@unil.ch)
Guangpeng Ren (guangpeng.ren@unil.ch)
Oriane Loiseau (oriane.loiseau@unil.ch)
Jonathan Rolland (jonathan.rolland@unil.ch)

April, 2015
Tree of Life

Phylogeny is the evolutionary history of living and extinct organisms.

Some key dates:

1964 Cavalli-Sforza and Edwards introduced parsimony and likelihood

1966 Hennig and the theory of cladistics

1977 Fitch used parsimony on DNA sequences

1978 Felsenstein worked out maximum likelihood method

1996 Rannala and Yang proposed Bayesian inference

since 2000 Computational phylogenetics
New uses for phylogenetics

Beside their use in systematics, trees are tools to

- study macroevolutionary processes
- detect evolution of genes and their function
- assess conservation issues
- drug design
- epidemiology
- evolution of languages
Genes evolution

Conservation issues

Extinction risks between plant lineages

Davies et al. (2011) PLoS Biol

A=Cypereae, B=Disa, C=Indiogofera, D=Lachnæa, E=Muraltia, F=Podalyrieae, H=Restionaceae, I=Zygophyllum, J=Protea, K=Moraea
Drug design

Cardoso et al. (2009)

Distribution of HIV-1 subtypes by age, gender, route of infection, year of diagnosis and level of primary resistance in patients from Central West Brazil:

<table>
<thead>
<tr>
<th></th>
<th>Subtype B</th>
<th>Subtype F1</th>
<th>Subtype C</th>
<th>Mosaic F1/F2, F2/F3, F3/CBI</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n %</td>
<td>n %</td>
<td>n %</td>
<td>n %</td>
<td>n</td>
</tr>
<tr>
<td><strong>Age (year)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2–18</td>
<td>80  82.5</td>
<td>6  6.2</td>
<td>3  3.1</td>
<td>8  8.2</td>
<td>97</td>
</tr>
<tr>
<td>20–40</td>
<td>2  2.1</td>
<td>– –</td>
<td>1  1</td>
<td>– –</td>
<td>3</td>
</tr>
<tr>
<td>&gt;40</td>
<td>60 61.8</td>
<td>6 6.2</td>
<td>2 2.1</td>
<td>5 5.1</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>18 18.6</td>
<td>– –</td>
<td>– –</td>
<td>3 3.1</td>
<td>21</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>57 58.8</td>
<td>4 4.1</td>
<td>1 1</td>
<td>4 4.1</td>
<td>66</td>
</tr>
<tr>
<td>Female</td>
<td>23 23.7</td>
<td>2 2.1</td>
<td>2 2.1</td>
<td>4 4.1</td>
<td>31</td>
</tr>
<tr>
<td><strong>Route of infection</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterosexual</td>
<td>37 38.1</td>
<td>3 3.1</td>
<td>2 2.1</td>
<td>5 5.1</td>
<td>47</td>
</tr>
<tr>
<td>IDU</td>
<td>21 21.6</td>
<td>2 2.1</td>
<td>1 1</td>
<td>2 2</td>
<td>25</td>
</tr>
<tr>
<td>Vertical</td>
<td>1 1</td>
<td>– –</td>
<td>– –</td>
<td>– –</td>
<td>2</td>
</tr>
<tr>
<td>MSM</td>
<td>2 2.1</td>
<td>– –</td>
<td>– –</td>
<td>– –</td>
<td>2</td>
</tr>
<tr>
<td>Unknown</td>
<td>18 18.6</td>
<td>1 1</td>
<td>– –</td>
<td>1 1</td>
<td>20</td>
</tr>
<tr>
<td><strong>HIV-1 diagnosis (year)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997–2006</td>
<td>22 22.7</td>
<td>4 4.1</td>
<td>1 1</td>
<td>4 4.1</td>
<td>31</td>
</tr>
<tr>
<td>&gt;2007</td>
<td>58 59.8</td>
<td>2 2.1</td>
<td>2 2.1</td>
<td>6 6.1</td>
<td>66</td>
</tr>
<tr>
<td><strong>Primary resistance (SDRM)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9  9.3</td>
<td>– –</td>
<td>– –</td>
<td>1 1</td>
<td>10</td>
</tr>
<tr>
<td><strong>Primary resistance (IAS-USA)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7  7.2</td>
<td>– –</td>
<td>– –</td>
<td>1 1</td>
<td>8</td>
</tr>
</tbody>
</table>

Epidemiology

H5N1 influenza epidemics

Phylogenetic trees: definitions.
Phylogeny and phylogenetic tree

The true and unobservable evolutionary history of a group of organisms is called a **phylogeny**.

A **phylogenetic tree** is an estimate of the true unobservable evolutionary history derived from morphology, molecular data, etc.

A **cladogram** or **topology** is the hierarchical structure of a phylogenetic tree, while a **phylogram** is a cladogram with explicit branch lengths and a **chronogram** has branch length as unit of time.
Mono-, para-, and polyphyly

A monophyletic group contains an ancestral taxa and \textbf{all} its descendant.

A paraphyletic group contains an ancestral taxa and \textbf{some} of its descendant.

A polyphyletic group contains some terminal taxa but \textbf{not} their common ancestor.
Congruence and incongruence

Two trees are **congruent** if their topology is identical. They are said to be **incongruent** if their topology is not identical.

An example:
A clade on a phylogenetic tree is **unresolved** if there is a node with three or more direct descendants.

This creates **polytomies**, which are called

- **soft** if it’s due to lack of data or a conflict between two plausible trees
- **hard** if they represent real two or more consecutive events of speciation
Orthologs, paralogs, xenologs

Two homologous genes are

**ortholog** if derived from a gene present in their common ancestor

**paralog** if resulted from the duplication of the same ancestral gene

**xenolog** if arose by lateral gene transfer

When inferring species phylogenetic trees, it is **essential** to have orthologous DNA sequences!
Branches of a tree

A branch can have a length that represent the evolutionary time elapsed between its endpoints.

- If the edge set of a tree have no lengths assigned, the shape of the tree is called **topology**.
- Suppose that $d_{xy}$ is the length of the path separating $x$ and $y$.
- Suppose that all leaves $x_i$ of a tree $T$ are equally distant from the root $r$, that is, there exist a constant $c$ such that $d_{x_ir} = c$.
- Then $T$ is called **ultrametric**
The root of a tree

In evolutionary studies, phylogenetic trees are drawn as branching trees deriving from a single ancestral species. This species is known as the **root** of the tree.

- A rooted tree is a tree to which a special internal node $r$ is added with degrees $\leq 2$.

**Labeled histories**: rooted trees with interior nodes ordered according to their age.
A tree that summarized the common features of a set of trees is called a consensus tree.

- **strict consensus**: shows only the groups that are shared among all trees.
- **semi-strict consensus**: shows only the resolved groups that are shared among all trees.
- **$n\%$ majority-rule consensus**: shows the groups that are shared by $n\%$ of the trees.
How do we measure evolution?
Minimum net amount of evolution

Parsimony methods are the easiest ones to explain and the idea behind the method was first expressed by Cavalli-Sforza and Edwards (1963):

The preferred evolutionary tree is the one involving “the minimum net amount of evolution”.

To measure this amount of evolution, we must be able to make a reconstruction of events, involving as few events as possible, on any phylogenetic tree.
Parsimony on DNA

Fitch algorithm counts the number of evolutionary changes for nucleotide data, in which any one of the four bases A, C, G, and T can change to any other.

Algorithm

- at each internal node, create a set of possible character that is the intersection of sets of the two descendant nodes
- if the set is empty, use the union of the descendant nodes instead
- everytime we need to use the union, count one change of state

Very quick algorithm: computation proportional to the number of species in the tree.
Sankoff algorithm

This is the general algorithm for parsimony. All other proposed algorithms are special cases of this one.

Let $c_{ij}$ be the **cost of change** from character state $i$ to $j$.

For a given character, compute for each node $k$, the **minimal cost** $S_k(i)$, given that node $k$ is assigned the state $i$, of all the changes in the upward portion of the tree.

For each node, this quantity is a recursive function

$$S_a(i) = min_j[c_{ij} + S_l(j)] + min_k[c_{ik} + S_r(k)]$$

This quantity for the bottom node of a tree is then the length of the tree

$$S = min_i S_0(i)$$
Sankoff algorithm

\[
\begin{array}{c|c|c|c}
\{C\} & \{A\} & \{C\} & \{A\} & \{G\} \\
\infty & 0 & \infty & \infty & \infty \\
2.5 & 2.5 & 3.5 & 3.5 & \\
1.5 & 1.5 & \\
3.5 & 3.5 & 3.5 & 4.5 & \\
6 & 6 & 7 & 8 & \\
\end{array}
\]

Cost matrix:

\[
\begin{array}{cccc}
\text{from} & A & C & G & T \\
A & 0 & 2.5 & 1 & 2.5 \\
C & 2.5 & 0 & 2.5 & 1 \\
G & 1 & 2.5 & 0 & 2.5 \\
T & 2.5 & 1 & 2.5 & 0 \\
\end{array}
\]
Statistical properties of parsimony

As the amount of data approach infinity, an estimator is **consistent** if it convergences to the true value of the parameter with probability 1. **inconsistent** if it converges to something else.

Explicit assumptions of parsimony:
- independence of the site along the sequence

Implicit assumptions of parsimony:
- no single rate of changes applying across characters
- no single rate of change for a branch that applies across all characters
- if $n$ species and $p$ characters, $(2n - 3)p$ parameters involved...

Tuffley and Steel (1997)
Consistency and parsimony

Long branch attraction:
- probability of parallel changes along long branches is $p$
- probability of single change in any other branches is $q$

If the tree is short enough, even large ratios of the length of the long to the short branches do not cause inconsistency.

Common “assumption” of parsimony: low rate of evolution of your gene(s).

Felsenstein (1978)
Generalisation

The proof of inconsistency of parsimony has been generalized to DNA data, and larger trees.

However, parsimony can be “rescued” if long branches in the tree are broken by adding more taxa:

Problem: we don’t know in advance which branch to break, but a good taxon sampling should take care of that

Graybeal (1995)
So why still using parsimony?

Because of
- its simplicity (you can check it by hand)
- its rapidity
- overall, the topology obtained is not that bad. . .

Several areas of phylogenetics use parsimony
- reconstruction of gene genealogies (Median Joining Networks, TCS, . . .)
- ancestral character state reconstruction
- phylogeography (haplotype networks, nested clade analysis, . . .)
- starting trees for more complex optimality criteria (e.g. RAxML)
- any areas where building complex statistical models is impossible
Modeling evolution through time and lineages?
Likelihood and models

Maximum likelihood relies on explicit probabilistic models of evolution.

But, the process of evolution is so complex and multifaceted that basic models involve assumption built upon assumption.

This reliance is often seen as a weakness of the likelihood framework, but

- the need to make explicit assumptions is a strength
- enables both inferences about evolutionary history and assessments of the accuracy of the assumptions made
- this leads to a better understanding of evolution

“The purpose of models is not to fit the data, but to sharpen the questions” (S. Karlin)
Description of Maximum Likelihood

Given an hypothesis $H$ and some data $D$, the likelihood of $H$ is

$$L(H) = \text{Prob}(D|H) = \text{Prob}(D_1|H)\text{Prob}(D_2|H) \cdots \text{Prob}(D_n|H)$$

if the $D$ can be split in $n$ independent parts.

Note that $L(H)$ is not the probability of the hypothesis, but the probability of the data, given the hypothesis.

Maximum likelihood properties (Fisher, 1922)

- consistency – converge to correct value of the parameter
- efficiency – has the smallest possible variance around true parameter value
Let say we toss a coin 11 times and obtain 5 heads and 6 tails. All tosses are independent and all have the same unknown head probability $p$. What is the probability of this data?

$$L(p) = \text{Prob}(D|p) = p^5(1 - p)^6$$

The maximum likelihood is $p = 0.454545$, which can be found by equating the derivative of $L(p)$ with respect to $p$ to zero and solving:

$$\frac{dL(p)}{dp} = 5p^4(1 - p)^6 - 6p^5(1 - p)^5 = 0$$

which yields $\hat{p} = 5/11 = 0.454545$. 
Measuring evolution

The problem that we are dealing with here is how to calculate the probability of evolution in sequences (whether DNA, AA at the inter- or intra-specific level).

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strepsiptera</td>
<td>AAGCTCATTAATCGCTTTGTTCCATTAGATAGTTGGAT...</td>
</tr>
<tr>
<td>Aedes</td>
<td>AGGCTCAGTATAACATATAATTTACAAAGATCATTTGGAT...</td>
</tr>
<tr>
<td>Drosophila</td>
<td>AGGCTCATATATATAATGTTCATTTAGATCGTTGGAT...</td>
</tr>
<tr>
<td>Flea</td>
<td>TGGCTCATTATATCTTTAGTTCATTAGATCGTTGGAT...</td>
</tr>
<tr>
<td>Meloe</td>
<td>AGGCTCATTAAATGATTAGTGTTCTTAGATCGTTGGAT...</td>
</tr>
<tr>
<td>Tenebrio</td>
<td>AGGCTCATTAATGATTAGTGTTCTTAGATCGTTGGAT...</td>
</tr>
</tbody>
</table>

\[ H^{0.01} \]
What do we want to infer

The data at hand are observed pattern frequencies in aligned sequences:

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strepsiptera</td>
<td>AAGCTCATTAATCGCTTTGGTTCTTCTAAGTAGTTGGAT...</td>
</tr>
<tr>
<td>Aedes</td>
<td>AGGCTCATTAACACTATAATTTACAAAGATCATTGGAT...</td>
</tr>
<tr>
<td>Drosophila</td>
<td>AGGCTCATATATCATTATGGTTCTTCTAAGATCGTTGGAT...</td>
</tr>
<tr>
<td>Flea</td>
<td>TGGCTCATATATCATTATGGTTCTAAGATCGTTGGAT...</td>
</tr>
<tr>
<td>Meloe</td>
<td>AGGCTCATAAATCATATATGGTTCTTCTAAGATCGTTGGAT...</td>
</tr>
<tr>
<td>Tenebrio</td>
<td>AGGCTCATAAATCATATATGGTTCTTCTAAGATCGTTGGAT...</td>
</tr>
</tbody>
</table>

\[
\hat{p}_{AAAAAAA} = \frac{\# \text{ observations of } AAAAAAA}{\text{sequence length}}, \text{ etc.}
\]

which, assuming a model of molecular evolution along a tree, are estimators for the true joint distribution \( p_{AAAAAAA} \), etc.
Derivation of the process

We assume that the distribution of the number of substitutions $s$ is a Poisson random variable with mean $\lambda t$.

Then,
- rate of substitutions relative to a unit of time at a given site is $\lambda t$
- probability of $s \geq 0$ at a site in a time period $t$ is
  \[ Pr(s) = \frac{\exp^{-\lambda t} (\lambda t)^s}{s!} \]
- mean number of substitutions during $t$ units of time is $\lambda t$
Probabilities of changes (and no change)

The probability of no changes occurring at a site

$$Pr(s = 0) = \exp(-\lambda t)$$

and the probability for at least one substitution

$$Pr(s \neq 0) = 1 - \exp(-\lambda t)$$

When $t$ is small (so that multiple and back substitutions are rare), these probabilities can be approximated by

$$Pr(s = 0) \approx 1 - \lambda t$$

and

$$Pr(s \neq 0) \approx \lambda t$$
Infinitesimal probabilities

These probabilities can be seen as the infinitesimal probabilities of a Markov process.

For DNA sequences,

- this process is a 4-state Markov chain
- $p_{ij}$ represents the transition probability that the next state is $j$ given that the current state is $i$

\[ p_{ij} = Pr(\text{next state } S_j | \text{current state } S_i) \]

Let $P = \{P\}_{ij}$ denote the matrix of transition probabilities, then

\[ P(t + h) = P(t)P(h) \]
Extending to 4 states

We can extend the results from the simpler Poisson process to a 4-state Markov process for DNA data.

For $h$ small, the transition probabilities are approximated by

$$P(h) \approx I + Qh$$

where $Q$ is the infinitesimal or instantaneous rate matrix for a continuous Markov process

$$Q = \begin{bmatrix}
-r_{AA} & r_{AC} & r_{AG} & r_{AT} \\
r_{CA} & -r_{CC} & r_{CG} & r_{CT} \\
r_{GA} & r_{GC} & -r_{GG} & r_{GT} \\
r_{TA} & r_{TC} & r_{TG} & -r_{TT}
\end{bmatrix}$$
Replacing \( P(h) \) by its approximation in \( P(t + h) = P(t)P(h) \) gives

\[
P(t + h) = P(t)(I + Qh) = P(t) + P(t)Qh
\]

Rearranging, we get

\[
\frac{P(t + h) - P(t)}{h} = P(t)Q
\]

Taking the limit as \( h \to 0 \), leads to the differential equation \( P' = PQ \)

with initial condition \( P(0) = I \), which leads to

\[
P(t) = \exp(Qt)
\]
Instantaneous rate matrix

Three conditions are added
  • from theory of finite Markov process,
    \[ q_{ii} = - \sum_{j=0, j \neq i}^{4} q_{ij}, \]
  • thus
    \[ \sum_{j=0}^{4} q_{ij} = 0. \]
  • rate of change should be 1 per unit time, so that we are scaling branches in expected nucleotide substitutions per site. Therefore,
    \[ \sum_{j=0, j \neq i}^{4} q_{ij} = 1. \]
Ergodicity

The continuous time Markov chain is further assumed (see conditions on previous slide) to be **ergodic**.

As $t \to \infty$, the probability that the site is in some state $j$ is non-zero and independent of the starting state. That is, there are positive values $\pi_1, \ldots, \pi_c$ such that, for all $i, j$

$$\lim_{t \to \infty} P_{ij}(t) = \pi_j$$

The values $\pi_1, \ldots, \pi_c$ comprise a **stationary distribution** (also called equilibrium distribution, equilibrium frequencies or invariant measure) for the states.

For all $t \leq 0$ these values satisfies $\pi_j = \sum_{i \in E} \pi_i P_{ij}(t)$. 
Time reversibility

For DNA evolution, the next common, but not necessary, assumption is time reversibility.

We do **NOT** assume that the probability of going from state $i$ to state $j$ is the same as probability of going from $j$ to $i$.

Instead, we assume that probability of sampling $i$ from the stationary distribution and going to $j$ is the same as sampling $j$ from the equilibrium distribution and going to $i$:

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$$
Rate of substitutions

As $Q$ and $t$ occur only in the form of a product in $P(t) = e^{Qt}$, it is conventional to scale $Q$ so that the average rate is 1.

The branch length is therefore measured in expected substitutions per site.

A long branch can therefore either by due to

- long evolutionary time
- a rapid rate of substitution
- a combination of both
In summary

The Markov chain is characterized by its generator matrix $Q = \{q_{ij}\}$, where $q_{ij}$ is the **instantaneous** rate of change from $i$ to $j$ when $\Delta t \to 0$, that is

$$Pr\{X(t + \Delta t) = j|X(t) = i\} = q_{ij}\Delta t$$

The $Q$ matrix fully determines the dynamics of the Markov chain.

It specifies, in particular, the transition-probability matrix over any time $t > 0$, $P(t) = \{p_{ij}(t)\}$ where

$$p_{ij}(t + s) = Pr\{X(t + s) = j|X(t) = i\} = \sum_{k} p_{ik}(t)p_{kj}(s)$$

with the relationship

$$P(t) = e^{Qt}$$
Jukes-Cantor, 1969

\[
Q = \begin{bmatrix}
-1 & 1 & 1 & 1 \\
1 & -1 & 1 & 1 \\
1 & 1 & -1 & 1 \\
1 & 1 & 1 & -1 \\
\end{bmatrix}
\]

Scaling and setting the correct value for \( q_{ii} \) gives:

\[
Q = \begin{bmatrix}
-1 & 1/3 & 1/3 & 1/3 \\
1/3 & -1 & 1/3 & 1/3 \\
1/3 & 1/3 & -1 & 1/3 \\
1/3 & 1/3 & 1/3 & -1 \\
\end{bmatrix}
\]

Transition probabilities are \( P(t) = \exp^{Qt} \)

\[
p_{ij}(t) = \begin{cases} 
3/4 + \exp(-4t/3)/4, & \text{if } i = j, \\
1/4 + \exp(-4t/3)/4, & \text{if } i \neq j, 
\end{cases}
\]
Kimura 2-parameters, 1981

\[
Q = \begin{bmatrix}
-1 & 1 & \kappa & 1 \\
1 & -1 & 1 & \kappa \\
\kappa & 1 & -1 & \\
1 & \kappa & 1 & -1 \\
\end{bmatrix}
\]

Here, \( \kappa = \frac{\alpha}{\beta} \), where
- \( \alpha \) is the transition rate
- \( \beta \) is the transversion rate.

It simplifies to
- Jukes-Cantor, 1969 if \( \alpha = \beta \).
Felsenstein, 1981

\[ Q = \begin{bmatrix}
- & \pi_C & \pi_G & \pi_T \\
\pi_A & - & \pi_G & \pi_T \\
\pi_A & \pi_C & - & \pi_T \\
\pi_A & \pi_C & \pi_G & - \\
\end{bmatrix} \]

where \( \pi_i \) is the equilibrium frequency of nucleotide \( i \).

It simplifies to

- Jukes-Cantor, 1969 if \( \pi_A = \pi_C = \pi_G = \pi_T \).
Felsenstein, 1984

\[
Q = \begin{bmatrix}
\pi_A & \pi_C & (1 + \kappa/\pi_R)\pi_G & \pi_T \\
\pi_A & \pi_C & \pi_G & (1 + \kappa/\pi_Y)\pi_T \\
(1 + \kappa/\pi_R)\pi_A & (1 + \kappa/\pi_Y)\pi_C & \pi_G & \pi_T \\
\pi_A & (1 + \kappa/\pi_Y)\pi_C & \pi_G & - \\
\end{bmatrix}
\]

Here, \( \pi_i \) is the equilibrium frequency of nucleotide \( i \) and \( \kappa = \alpha/\beta \), where

- \( \alpha \) is the transition rate
- \( \beta \) is the transversion rate.

It simplifies to

- Kimura, 1980 if \( \pi_A = \pi_C = \pi_G = \pi_T \).
- Felsenstein, 1981 if \( \kappa = 0 \).
Hasegawa-Kishino-Yano, 1985

\[
Q = \begin{bmatrix}
- & \pi_C & \kappa \pi_G & \pi_T \\
\pi_A & - & \pi_G & \kappa \pi_T \\
\kappa \pi_A & \pi_C & - & \pi_T \\
\pi_A & \kappa \pi_C & \pi_G & -
\end{bmatrix}
\]

Here, \( \pi_i \) is the equilibrium frequency of nucleotide \( i \) and \( \kappa = \alpha/\beta \), where

- \( \alpha \) is the transition rate
- \( \beta \) is the transversion rate.

It simplifies to
- Kimura, 1980 if \( \pi_A = \pi_C = \pi_G = \pi_T \).
- Felsenstein, 1981 if \( \kappa = 1 \).
Here, $\pi_i$ is the equilibrium frequency of nucleotide $i$, $\kappa_1 = \alpha_Y/\beta$ and $\kappa_2 = \alpha_R/\beta$ where

- $\alpha_R$ is purine transition rate (A+G).
- $\alpha_Y$ is pyrimidine transition rate (C+T).
- $\beta$ is the transversion rate.

It simplifies to

- Hasegawa, Kishino, Yano, 1985 if $\alpha_R/\alpha_Y = \pi_R/\pi_Y$.
- Felsenstein, 1984 if $\alpha_R = \alpha_Y$. 

\[
Q = \begin{bmatrix}
- & \pi_C & \kappa_2 \pi_G & \pi_T \\
\pi_A & - & \pi_G & \kappa_1 \pi_T \\
\kappa_2 \pi_A & \pi_C & - & \pi_T \\
\pi_A & \kappa_1 \pi_C & \pi_G & -
\end{bmatrix}
\]
General-time reversible

$$Q = \begin{bmatrix}
- & \alpha \pi_C & \beta \pi_G & \gamma \pi_T \\
\alpha \pi_A & - & \delta \pi_G & \epsilon \pi_T \\
\beta \pi_A & \delta \pi_C & - & \eta \pi_T \\
\gamma \pi_A & \epsilon \pi_C & \eta \pi_G & - 
\end{bmatrix}$$

where

- $\alpha \cdots \eta$ are rates of changes from one nucleotide to another
- $\pi_i$ are frequencies of nucleotides

It simplifies to

- all other models if parameters are correctly constrained
Variation of substitution rates

How to model rate variation among sites?

The idea is to use a probability distribution to model changes in rates of substitution among sites, e.g. Gamma distribution

- mean of distribution is $\alpha \beta$, and variance $\alpha \beta^2$
- set the mean rate of substitution to 1, so assume $\beta = 1/\alpha$
- $\alpha$ parameter allows to change characteristics of distribution

Felsenstein, 2004
Fig. 4. Likelihood values and estimates of the $\alpha$ parameter as functions of $k$, the number of categories in the discrete gamma model. The $\alpha$ and $\beta$ globin genes for the five mammalian orders (570 bp) are analyzed, assuming the best tree (Fig. 3) and the F84 + dG model. The average nucleotide frequencies are $\pi_T = 0.2200$, $\pi_C = 0.2449$, $\pi_A = 0.2761$, and $\pi_G = 0.2590$, with $\ell_{\text{max}} = -1,579.76$. When $k = \infty$, that is, with the F84 + $\Gamma$ model, $\ell = -1,761.17$ and $\hat{\alpha} = 0.360$.  

Yang, 1994
How to calculate the likelihood of a tree?
Suppose we have a data set $D$ of DNA sequences with $m$ sites.

We are given a topology $T$ with branch lengths and a model of evolution, $Q$ that allow us to compute $P_{ij}(t)$.

Assumptions made to compute likelihood $L(T, Q) = \text{Prob}(D|T, Q)$

- evolution in different sites is independent
- evolution in different lineages is independent
- the rate of evolution is the same between site and along lineages
The assumption of independence of sites along a sequence allow us to decompose the likelihood in a product of likelihood for each site

\[ L(T, Q) = \text{Prob}(D | T, Q) = \prod_{i=1}^{m} \text{Prob}(D_i | T, Q) \]

The likelihood for this tree for site i is

\[ \text{Prob}(D_i | T, Q) = \sum_{x} \sum_{y} \sum_{z} \sum_{w} \text{Prob}(A, C, C, C, G, x, y, z, w | T, Q) \]
With the independence of lineages assumptions, we can decompose the right hand side of the equation a bit further.

\[
\text{Prob}(A, C, C, C, G, x, y, z, w | T, Q) = \sum_x \sum_y \sum_z \sum_w \text{ACGT} \text{ACGT} \text{ACGT} \text{ACGT} \\
\text{Prob}(x) P_{xy}(t_6) P_{yA}(t_1) P_{yC}(t_2) \\
P_{xz}(t_8) P_{zC}(t_3) \\
P_{zw}(t_7) P_{wC}(t_4) P_{wG}(t_3)
\]
Pruning algorithm

- **Goal**: render the likelihood computation practicable using “dynamic programming”
- **Idea**: move summation signs as far right as possible and enclose them in parentheses where possible

\[
Prob(A, C, C, G, x, y, z, w | T) = \\
\sum_{x} \left( \sum_{y} Prob(x) \left( \sum_{z} P_{xy}(t_6)P_{yA}(t_1)P_{yC}(t_2) \right) \right) \\
\times \left( \sum_{z} P_{xz}(t_8)P_{zC}(t_3) \right) \\
\times \left( \sum_{w} P_{zw}(t_7)P_{wC}(t_4)P_{wG}(t_5) \right)
\]

The parentheses pattern and terms for tips has an exact correspondence to the structure of the tree.
Numerical integration

Up to now we assumed that both branch lengths and model parameters were fixed at a given value.

Of course, we don’t know these values so
- to get these values, start from a random guess
- make small changes and compare the improvement in likelihood
- do so until the likelihood do not change any more
- every time the topology is changed, reestimate these parameters

Drawback: very very very computationally intensive, but there are solutions to that.
Consistency and overparameterisation

Maximum likelihood can be proved to be consistent (see Felsenstein, 2004 pp. 271)

- true if we use the correct model of evolution
- but what happen if model in not correct? No guarantee can be given
- less problematic if what we want to infer is just the topology

Beware of overparameterisation problems

- adding more and more parameters to the model will result in a better fit to the data
- but this will lead to inconsistency
- a few parameters are more important than others, in particular the gamma distribution
How to choose the best tree?
Number of rooted trees?

\[ n = 2 \]
\[
\begin{array}{c}
  a \\
  b \\
/ \\
\end{array}
\]

\[ n = 3 \]
\[
\begin{array}{c}
  a \\
  b \\
  c \\
/ \\
/ \\
/ \\
\end{array}
\]

\[ n = 4 \]
\[
\begin{array}{c}
  a \\
  b \\
  c \\
  d \\
/ \\
/ \\
/ \\
/ \\
/ \\
/ \\
\end{array}
\]

\[
\begin{array}{c}
  a \\
  b \\
  c \\
  d \\
/ \\
/ \\
/ \\
/ \\
/ \\
/ \\
\end{array}
\]

\[
\begin{array}{c}
  a \\
  b \\
  c \\
  d \\
/ \\
/ \\
/ \\
/ \\
/ \\
/ \\
\end{array}
\]

\[
\begin{array}{c}
  a \\
  b \\
  c \\
  d \\
/ \\
/ \\
/ \\
/ \\
/ \\
/ \\
\end{array}
\]

\[
\begin{array}{c}
  a \\
  b \\
  c \\
  d \\
/ \\
/ \\
/ \\
/ \\
/ \\
/ \\
\end{array}
\]
More trees than electrons in Universe

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<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>6,190,283,353,629,375</td>
</tr>
<tr>
<td>17</td>
<td>191,898,783,962,510,625</td>
</tr>
<tr>
<td>18</td>
<td>6,332,659,870,762,850,625</td>
</tr>
<tr>
<td>19</td>
<td>221,643,095,476,699,771,875</td>
</tr>
<tr>
<td>20</td>
<td>8,200,794,532,637,891,559,375</td>
</tr>
<tr>
<td>30</td>
<td>$4.9518 \times 10^{38}$</td>
</tr>
<tr>
<td>40</td>
<td>$1.00985 \times 10^{57}$</td>
</tr>
<tr>
<td>50</td>
<td>$2.75292 \times 10^{76}$</td>
</tr>
</tbody>
</table>

Branching process

- $n$-th taxa can be added to $2n – 3$ places
- $3 \times 5 \times 7 \times \cdots \times (2n – 3)$
- $\frac{(2n-3)!}{2^{n-1}(n-1)!}$

- with 50 taxa, approaching Eddington’s number of electrons in the visible universe
Exhaustive search

To find the best tree, enumerate all the possible trees from the tree space:

But you can quickly see that it is impossible for trees larger than 10 taxa...
Taking shortcuts

To find the best tree, say with parsimony, we do not need to estimate all possible trees

- estimate the length of a random tree $L(T_r)$
- iterate over all possible trees
- if length of incomplete tree $T_i$
  $L(T_i) > L(T_r)$
  stop and forget about this path, it cannot find shorter trees

- more refined strategies to find starting length exists
Phylogenetics: Despair and hope

Branch and bound speed up the process greatly, and allow to compute best tree for data sets of 20 to 30 taxa.

But, computational time still increase exponentially with number of taxa!
- phylogenetic reconstruction is an NP-hard problem
- finding the length of a particular tree is an NP-complete problem
- no solution can be found in polynomial time
Even more shortcuts

Algorithms to speed up even more the search through the tree space

- take an initial tree $T_i$
- make small rearrangements (i.e. swap branches) to reach neighboring trees
- if you find a better, keep it and start again

- strategy will find local optimum in the tree space

They are called **greedy** algorithm because they seize the first improvement they see
Greedy algorithms

The problem with greedy algorithms:
- you can miss the global optimum!
- but we don’t have any other choice...
Nearest-Neighbour Interchange

A subtree

is rearranged by dissolving the connections to an interior branch

and reforming them in one of the two possible alternative ways:
Subtree Pruning and Regrafting

Break a branch, remove a subtree

Add it in, attaching it to one (*) of the other branches

Here is the result:
Tree Bisection and Reconnection

Break a branch, separate the subtrees

Connect a branch of one to a branch of the other

Here is the result:
Swapping is not all

Number of “neighbor” searched

**NNI** \(2(n - 3)\)

**SPR** \(4(n - 3)(n - 2)\), although some may be the same neighbour

**TBR** \((2n_1 - 3)(2n_2 - 3)\) possible ways to reconnect the two trees

Previous methods just make rearrangements to a particular tree by swapping branches, which is not enough to fully cover the tree space. Additional techniques have to be used

- stepwise taxon addition: start several heuristic searches by starting from a different initial tree each time
- allow suboptimal trees to be swapped on in order to cross “valleys” in the tree space
Other heuristic searches

You can “invent” all kinds of heuristic algorithm, and many have been proposed

- tree-fusing: exchange subtrees from two equally optimal trees
- genetic algorithms: let populations of trees evolve by setting fitness functions and mutation, migration, etc...
- tree windowing: do extensive rearrangement but on a local region of the tree
- search by reweighting: change the landscape of the tree space by reweighting characters
- simulated annealing: iteratively change the tree scoring function to lower the difference between optimal and suboptimal trees
Bootstrap on phylogenetic tree

Allow us to infer the variability of parameters in models that are too complex for easy calculation of their variance. This is the case of topologies!

Procedure

- sample whole columns of data with replacement
- recreate $n$ pseudomatrices with the same number of species and sites than original one
- build $n$ phylogenetic trees from these $n$ pseudoreplicates
- weigh each tree in replicate $i$ by the number of trees obtained
Summarizing results

The end result of a bootstrap is a cloud of trees.

What is the best way to summarize this given that trees have discrete topologies and continuous branch lengths?

We could make a histogram of the length of a particular branches

- it will give a lower limit on the branch length
- then check if 0 is in the 95% interval, we would assert the existence of the branche

A simpler solution is to count how many times a particular branch appears in the list of trees estimated by bootstrap.

A majority-rule consensus tree containing clades appearing in more than 50% of them can then be built
Simple example

Trees:

```
Trees: E A C F B D E C A B D F
      E A F D B C E A D F B C E C A D F B
```

Number of times each partition of species is found:

- AE | BCDF 3
- ACE | BDF 3
- ACEF | BD 1
- AC | BDEF 1
- AEF | BCD 1
- ADEF | BC 2
- ABDF | EC 1
- ABCE | DF 3

Majority–rule consensus tree of the unrooted trees:

```
Majority–rule consensus tree of the unrooted trees:
```

```
```

```
```
Multiple tests

We don’t (or rarely) know in advance which group interest us.

If we look for the most supported group on the tree and report its p-value, we have a “multiple-tests” problem

- if no significant evidence for existence of any groups on a tree
- 5% of branches are expected to be above 0.95
- so one out of every 20 branches of a tree would be significant
- furthermore, branches are not independent...

The p-value cannot be interpreted as statistical test.
Biases in bootstrap

Estimated p-value is conservative:

One source of conservatism

- with bootstrap, we make statement about branch lengths $\mu$
- then we reduce that to statements about tree topology, i.e. $\mu > 0$ or $\mu < 0$
- generalisation of Hillis and Bull (1993) results is not clear
Uncertainty assessment

Likelihood does not only allow to make point estimate of the topology and branch length, it also gives information about the uncertainty of our estimate.

It is possible to use the likelihood curve to test hypothesis and to make interval estimates.

Asymptotically (i.e. when the number of data point tend towards $\infty$), the ML estimate $\hat{\theta}$ is normally distributed around its true value $\theta_0$. 
Likelihood ratio test

At the maximum $\hat{\theta}$

$$S(\hat{\theta}) = \frac{\partial \ln L(\hat{\theta})}{\partial \theta} = 0$$

$$I(\hat{\theta}) = \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta^2} < 0$$

Then the log likelihood of the true parameter value can be linked with the maximum value estimated by a Taylor series:

$$\ln L(\theta_0) = \ln L(\hat{\theta}) + S(\hat{\theta})(\theta_0 - \hat{\theta}) - \frac{1}{2} I(\hat{\theta})(\theta_0 - \hat{\theta})^2$$

Subtracting $\ln L(\theta_0)$ from both sides, and rearranging leads to

$$2[\ln L(\hat{\theta}) - \ln L(\theta_0)] \approx \frac{(\theta_0 - \hat{\theta})^2}{\sigma} = \chi^2_1$$

For $p$ parameters, twice the difference in likelihood is $\chi^2_p$ distributed.
Nested models

Assumptions of the likelihood ratio test

• null hypothesis should be in the interior space that contains the alternative hypotheses

• if $q$ parameters have been constrained, they must be able to vary in both sense
  • if $L_0$ restricts one parameter to the end of its range, distribution of twice log likelihood ratio has half its mass at 0 and the other half in the usual $\chi^2$ distribution
  • halve the tail probability obtained with usual $\chi^2$

Valid only asymptotically

• should be close enough to the true values

• true with very large amounts of data
Akaike Information Criterion

More general model will always have higher likelihood than restricted models.
So, choosing model with highest likelihood will lead to one that is unnecessarily complex.

We should therefore compromise goodness of fit with complexity of model.

**AIC** for hypothesis $i$ with $p_i$ parameters:

$$AIC_i = -2\ln L_i + 2p_i$$

Hypothesis with the lowest AIC is preferred.
Adding more complexity

All the different models seen so far are special cases of the GTR+$\Gamma$+I model

- setting $\Gamma$ and I to 0 leads to GTR
- setting transversion rates to $\beta$ and transition rates to $\alpha$ leads to F84
- setting $\beta = \alpha$ leads to K2P
- setting all nucleotide frequencies to $\frac{1}{4}$ lead to JC69
Examples

Possible comparisons

• GTR+$\Gamma$ vs GTR
  • $2 \times [lnL_{GTR+\Gamma} - lnL_{GTR}]$
  • difference in df = 1

• GTR+$\Gamma$ vs F84+$\Gamma$
  • $2 \times [lnL_{GTR+\Gamma} - lnL_{F84+\Gamma}]$
  • difference in df = 4

• F84+$\Gamma$ vs JC69
  • $2 \times [lnL_{F84+\Gamma} - lnL_{JC69}]$
  • difference in df = 5

But not JC69+$\Gamma$ vs GTR because they are not nested!
Are two topologies different?

LRT or AIC not possible on topologies because they are discrete entities, and not quantitative parameters.

We therefore have to rely on **paired-sites** tests

- expected $lnL$ is average $lnL$ per site as number of sites grows without limit
- if sites are independent and two trees have equal expected $lnL$, then differences in $lnL$ at each site is drawn independently with expectation zero
- statistical test of mean of these differences is zero
- valid for likelihood and parsimony
Example

232-sites mtDNA for 7 mammals

Tree I
- Bovine
  - Gibbon
    - Orang
      - Gorilla
        - Chimp
          - Human

Tree II
- Bovine
  - Gibbon
    - Orang
      - Gorilla
        - Chimp
          - Human

<table>
<thead>
<tr>
<th>Tree</th>
<th>Site</th>
<th>Site</th>
<th>Site</th>
<th>Site</th>
<th>Site</th>
<th>Site</th>
<th>231</th>
<th>232</th>
<th>ln L</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
<td>-2.971</td>
<td>-4.483</td>
<td>5.673</td>
<td>-5.883</td>
<td>-2.691</td>
<td>-8.003</td>
<td>...</td>
<td>-2.971</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-2.983</td>
<td>-4.494</td>
<td>5.685</td>
<td>-5.898</td>
<td>-2.700</td>
<td>-7.572</td>
<td>...</td>
<td>-2.987</td>
</tr>
<tr>
<td>II</td>
<td>3</td>
<td>+0.012</td>
<td>+0.111</td>
<td>+0.013</td>
<td>+0.015</td>
<td>+0.010</td>
<td>-0.431</td>
<td>...</td>
<td>+0.012</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>231</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>232</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diff</td>
<td>+0.012</td>
<td>+0.111</td>
<td>+0.013</td>
<td>+0.015</td>
<td>+0.010</td>
<td>-0.431</td>
<td>...</td>
<td>+0.012</td>
</tr>
</tbody>
</table>

Felsenstein, 2004
Possible tests

Different form of tests possible

$z$ assumes the differences at each site are normally distributed and estimate the variance of differences of the scores

KH use bootstrap sampling to infer distribution of sum of differences of scores and see whether 0 lay in the tails of distribution

Results

$z$ sum of $\ln L = 3.18$, $\sigma^2 = 0.04$, so variance of sum of differences is 11.31; $z$ statistic is $3.18/3.36 = 0.94$ with $p$-value of 0.34

KH 10,000 bootstrap samples of sites; 8,326 favored tree II, which yields a one-tailed probability of 0.16, and a 0.33 two-tailed probability
Multiple tests, again

If we want to test more than two trees

• compare each tree to best tree
• accept all trees that cannot be rejected by KH test

• multiple tests setting, but no reduction of nominal rejection level possible
• need to correct for all different ways the data can vary, ways that support different trees

When two trees are compared, but one of them is the actual best tree, we should do a one-tailed test.
Resampling technique that approximately corrects for testing multiple trees

1. make $R$ bootstrap of the $N$ sites
2. for each tree, normalize resampled $lnL$ so they have same expectation
3. for $j$th bootstrap, calculate $\tilde{S}_{ij}$ for $i$th tree how far normalized value is below maximum across all trees for that replicate
4. for each tree $i$, the tail probability is proportion of bootstrap replicates in which $\tilde{S}_{ij}$ is less than the actual difference between ML and $lnL$ of that tree

Resampling build a “least-favorable” case in which the trees show some patterns of covariation of site as in actual data but do not differ in overall $lnL$.

**One limitation**: assume that all proposed trees are possibly equal in likelihood
Parametric bootstrap

Use computer simulations to create pseudorandom data sets

Advantages and disadvantages

- can sample from the desired distribution, even with small data sets
- flexible hypothesis testing framework
- close reliance on the correctness of the model of evolution
Measuring, searching and assessing trees all in one!
Bayesian methods

The likelihood calculation is used as well by Bayesian methods. However, another component is added to the method: the prior distributions.

Before observing any data, each parameter will be assigned a prior distribution
- topologies
- branch lengths
- each parameter of the model of evolution

The prior distributions are then combined with the likelihood of the data to give the posterior distribution.

This is a highly attractive quantity because it computes what we most need: the probabilities of different hypotheses in the light of the data.
Bayes theorem

To combine all this together, we use the Bayes theorem

\[ \text{Prob}(T|D) = \frac{\text{Prob}(T \cup D)}{\text{Prob}(D)} \]

where \( \text{Prob}(T \cup D) = \text{Prob}(T)\text{Prob}(D|T) \)

so that

\[ \text{Prob}(T|D) = \frac{\text{Prob}(T)\text{Prob}(D|T)}{\text{Prob}(D)} \]
Normalizing constant

The denominator $\text{Prob}(D)$ is the sum of the numerator $\text{Prob}(T)\text{Prob}(D|T)$ over all possible hypotheses $T$.

This quantity is needed to normalize the probabilities of all $T$ so that they add up to 1.

This leads to

$$\text{Prob}(T|D) = \frac{\text{Prob}(T)\text{Prob}(D|T)}{\sum_T \text{Prob}(T)\text{Prob}(D|T)}$$

In words:

$$\text{posterior probability} = \frac{\text{prior probability} \times \text{likelihood}}{\text{normalizing constant}}$$
We want to estimate $p$, the probability of obtaining head, by tossing a coin $n$ times, which results in $n_h$ heads and $n_t$ tails

- binomial distribution to calculate the likelihood of $p$

$$B(n, n_h, p) = \binom{n}{n_h} p^{n_h} (1 - p)^{n-n_h}$$

- we make two trials of 10 and 1000 draws resulting in
  - 3 heads and 7 tails
  - 300 heads and 700 tails
Exponential prior

Likelihood 10 coins

Posterior 10 coins
Exponential prior

Likelihood 1000 coins

Posterior 1000 coins
Estimate normalizing constant

Posterior distribution expression has a denominator, i.e. 
$$\sum_T Prob(T)Prob(D|T)$$, that is often impossible to compute.

Fortunately, samples from the posterior distribution can be drawn using a Markov chain that does not need to know the denominator

- draw a random sample from posterior distribution of trees
- becomes possible to make probability statements about true tree
- e.g. if 96% of the samples from posterior distribution have (human, chimp) as monophyletic group, probability of this group is 96%
Makov chain Monte Carlo

**Idea:** to wander randomly through tree space by sampling trees until we settle down into an equilibrium distribution of trees that has the desired distribution, i.e. posterior distribution.

- Markov chain: the new proposed tree will depend only on the previous one
- to reach equilibrium distribution, the Markov chain must be
  - aperiodic – no cycles should be present in the Markov chain
  - irreducible – every trees must be accessible from any other tree
  - probability of proposing $T_j$ when we are at $T_i$ is the same as probability of proposing $T_i$ when we are at $T_j$
- the Markov chain has no end
MCMC in practice

Metropolis algorithm

- start with a random tree $T_i$
- select a new tree $T_j$ by modifying $T_i$ in some way
- compute

$$R = \frac{\text{Prob}(T_j|D)}{\text{Prob}(T_i|D)}$$

**The beauty of MCMC:** the normalizing constant being the same, it simplifies

$$R = \frac{\text{Prob}(T_j)\text{Prob}(D|T_j)}{\text{Prob}(T_i)\text{Prob}(D|T_i)}$$

- if $R \geq 1$, accept $T_j$
- if $R < 1$, draw a random number $n$ between $[0, 1]$ and accept $T_j$ if $R > n$, otherwise keep $T_i$
The robot takes a step if it draws a random number (uniform on 0.0 to 1.0), and that number is less than or equal to R.
Putting it all together

• Start with random tree and arbitrary initial values for branch lengths and model parameters
• Each generations, chose at random to propose either:
  • a new tree
  • new branch length
  • new model parameter values
• Either accept or reject the move
• Every $k$ generations, save tree topology, branch lengths and model parameters (i.e. sample the chain)
• After $n$ generations, summarize the sample using histograms, means, credibility intervals, etc.
How to propose a new tree

We could invent any type of proposal distribution to wander through the tree space

- e.g. NNI by selecting a node at random
- should be able to reach all trees from any starting tree
- at least after “sufficient” running, but impossible to know how much running is enough

Should be careful because

- if trees proposed are too different $\Rightarrow$ these trees will be rejected too often
- if trees proposed are too similar $\Rightarrow$ tree space won’t be sampled well enough
Metropolis-coupled Markov-chain Monte Carlo or MC$^3$

- involves running several MCMC chains simultaneously
- the cold chain is the one that counts, i.e. trees are sampled from that one
- the rest are called heated chains and are there to help the convergence
- chain is heated by raising densities to a power less than 1.0
MCMC

- Short steps fall short
- Longer step suggested by scout

- Small drop
- Big drop

P. Lewis, 2006
MCMC trace plot

"White noise" appearance is a good sign

Burn-in is over right about here

We started off at a very low point
Slow mixing

Chain is spending long periods of time "stuck" in one place

Indicates step size too large, and most proposed steps would take the robot "off the cliff"
Convergence in MCMC

Greatest practical problem: how long to run a chain to obtain a good approximation of the posterior probabilities

The most important is to check that independent runs lead to the same posterior distribution.
Type of prior distributions

Prior distributions for topologies have been proposed
  • stochastic process of random speciation and extinction
  • uniform distribution of all possible rooted trees

Prior distributions on branch lengths
  • exponential distribution
  • uniform distribution

Prior on model parameters
  • Dirichlet distribution on nucleotide frequency
  • uniform or exponential distribution on shape of $\gamma$ distribution
  • uniform distribution on proportion of invariant sites
Problems with priors

As shown before, we have to be careful with prior because they can exclude possible estimated values.

Other problematic aspects:

- universality of priors
- use of “uninformative” flat priors
  - issues of scale
  - unbounded quantities
Jukes-Cantor model

Sequence disimilarity and branch length under this model:

\[ D = \frac{3}{4} \left( 1 - e^{-\frac{4}{3}t} \right) \]

- **Different scale**
  - Graphs showing differences per site against branch length.
  - Scale variation indicated.

- **Unbounded priors**
  - Graphs showing different distributions or priors against time and probability.
  - Scale variation and priors demonstrated.

- **General observations**
  - Graphs and equations present.
  - Scale and priors varied across different axes.

Effect of priors

---

Yang and Rannala, 2006

---

(a)

(b)
What can we do with Bayesian

Beside their use in building phylogenetic trees, Bayesian methods are useful to deal with complex biological problems

- testing hypotheses about rates of host switching and cospeciation in host/parasite systems
- dating phylogenetic trees using autocorrelated prior distribution on rates of evolution
- infer rate of change of states of a character and the bias in the rate of gain of the character
- infer accuracy of inference of ancestral states
- infer position of the root of the tree
- testing rates of speciation and assess key innovations associated with changes in rates