Maximum Likelihood in Phylogenetics

23 January 2013
Workshop on Molecular Evolution
Český Krumlov

Paul O. Lewis
Department of Ecology & Evolutionary Biology
University of Connecticut, Storrs, CT
Goals

Explain jargon
Increase comfort level
Provide background
In other words...give a hand up
The Plan

- Probability review
- Likelihood
- Substitution models

- Markov model basics
  - Transition probabilities
  - Survey of models
  - Rate heterogeneity
  - Codon models

- The AND and OR rules
- Independence of events

- What does it mean?
  - Likelihood of a single sequence
  - Maximum likelihood distances
  - Likelihoods of trees
Combining probabilities

• *Multiply* probabilities if the component events must happen *simultaneously* (i.e. where you would naturally use the word AND when describing the problem)

Using 2 dice, what is the probability of

\[ \begin{array}{c}
\blacklozenge \\
\blacklozenge \\
\end{array} \text{ AND } \begin{array}{c}
\blacklozenge \\
\blacklozenge \\
\end{array} \ ? \\
\end{array} \]

\[ (1/6) \times (1/6) = 1/36 \]
One use of the AND rule in phylogenetics is to combine probabilities associated with individual branches to produce the overall probability of the data for one site.

\[ \Pr(AAC, AA) = p_0 \cdot p_1 \cdot p_2 \cdot p_3 \cdot p_4 \]
Combining probabilities

- *Add* probabilities if the component events are **mutually exclusive** (i.e. where you would naturally use the word OR in describing the problem)

Using one die, what is the probability of

\[
\begin{array}{c}
\text{●} \\
\text{●●}
\end{array}
\text{ OR } \begin{array}{c}
\text{●●} \\
\text{●●}
\end{array}
\] ?

\[\frac{1}{6} + \frac{1}{6} = \frac{1}{3}\]
Combining AND and OR

What is the probability that the sum of two dice is 7?

\[
(\frac{1}{36}) + (\frac{1}{36}) + (\frac{1}{36}) + (\frac{1}{36}) + (\frac{1}{36}) + (\frac{1}{36}) = \frac{1}{6}
\]
Using both AND and OR in phylogenetics

AND rule used to compute probability of the observed data for each combination of ancestral states.

OR rule used to combine different combinations of ancestral states.
Independence

This is always true...

\[ \Pr(A \text{ and } B) = \Pr(A) \Pr(B|A) \]

joint probability \hspace{1cm} \text{conditional probability}

If we can say this...

\[ \Pr(B|A) = \Pr(B) \]

...then events \( A \) and \( B \) are independent and we can express the joint probability as the product of \( \Pr(A) \) and \( \Pr(B) \)

\[ \Pr(A \text{ and } B) = \Pr(A) \Pr(B) \]
Likelihood
The Likelihood Criterion

The probability of the observations computed using a model tells us how surprised we should be.

*The preferred model is the one that surprises us least.*

Suppose I threw 20 dice down on the table and this was the result...
The Fair Dice model

\[
Pr(\text{obs.} | \text{fair dice model}) = \left( \frac{1}{6} \right)^{20} = \frac{1}{3,656,158,440,062,976}
\]

You should have been very surprised at this result because the probability of this event is very small: only 1 in 3.6 quadrillion!
The Trick Dice model
(assumes dice each have 5 on every side)

\[ \text{Pr(} \text{obs.}|\text{trick dice model}) = 1^{20} = 1 \]

You should not be surprised at all at this result because the observed outcome is certain under this model.
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Likelihood</th>
<th>Surprise level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair Dice</td>
<td>(\frac{1}{3,656,158,440,062,976})</td>
<td>Very, very, very surprised</td>
</tr>
<tr>
<td>Trick Dice</td>
<td>1.0</td>
<td>Not surprised at all</td>
</tr>
</tbody>
</table>

The winning model maximizes likelihood (and thus minimizes surprise).
Likelihood and model comparison

• Analyses using likelihoods ultimately involve **model comparison**

• The models compared can be **discrete** (as in the fair vs. trick dice example)

• More often the models compared differ continuously:
  – Model 1: branch length is 0.05
  – Model 2: branch length is 0.06

Rather than having an infinity of models, we instead think of the branch length as a **parameter** within one model.
Probabilities \((x)\) lie between 0 and 1, which means \(\log(x)\) will always be negative.

Note that \(\log(x)\) will always be highest where \(x\) is highest, so finding the maximum likelihood is equivalent to finding the maximum log-likelihood.

In this talk (and in phylogenetics in general), \(\ln(x) = \log(x)\).
Tree jargon

- **A**
- **B**
- **C**
- **D**
- **E**

- Interior node (or vertex, degree 3+)
- Terminal or tip node (or leaf, degree 1)
- Branch (edge)
- Root node of tree (degree 2)

Bipartition (split) also written AB|CDE or portrayed **---**
Likelihood of a single tip node

First 32 nucleotides of the ζη-globin gene of gorilla:

\[
\text{GAAGTCCCTTGAGAAATAAACTGCACACACTGG}
\]

\[
L = \pi_G \pi_A \pi_A \pi_G \pi_T \pi_C \pi_C \pi_T \pi_G \pi_A \pi_A \pi_G \pi_A \pi_A \pi_T \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G
\]
\[
= \pi_A^{12} \pi_C^7 \pi_G^7 \pi_T^6
\]

Note that we are assuming independence among sites here

\[
\log L = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)
\]

We can already see by eye-balling this that a model allowing \textbf{unequal} base frequencies will \textbf{fit better} than a model that assumes \textbf{equal} base frequencies because there are about twice as many As as there are Cs, Gs and Ts.
Model ranking using AIC

The Akaike Information Criterion (AIC) can be used to evaluate whether an *unconstrained* model ("free") fits the data significantly better than a *constrained* version ("equal") of the same model.

Find *maximum* logL under the *unconstrained* model:

\[
\log L_{\text{free}} = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)
\]

\[
= 12 \log(0.375) + 7 \log(0.219) + 7 \log(0.219) + 6 \log(0.188)
\]

\[
= -43.1
\]

Find *maximum* logL under the *constrained* model:

\[
\log L_{\text{equal}} = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)
\]

\[
= 12 \log(0.25) + 7 \log(0.25) + 7 \log(0.25) + 6 \log(0.25)
\]

\[
= -44.4
\]
Model ranking using AIC

Calculate AIC for each model:

\[ AIC = 2k - 2 \log(L_{\text{max}}) \]
\[ AIC_{\text{free}} = 2(3) - 2(-43.1) = 92.2 \]
\[ AIC_{\text{equal}} = 2(0) - 2(-44.4) = 88.8 \]

The constrained model ("equal") is a better choice than the unconstrained model ("free") according to AIC (K-L stands for Kullback-Leibler)

92.2 = twice expected (relative) K-L divergence from equal model to true model

88.8 = twice expected (relative) K-L divergence from free model to true model

(K-L stands for Kullback-Leibler)
Likelihood of the simplest tree

sequence 1 ──── sequence 2

To keep things simple, assume that the sequences are only 2 nucleotides long:

\[ L = L_1 \cdot L_2 \]

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

Note that we are NOT assuming independence here.
Maximum likelihood estimation

First 32 nucleotides of the ψη-globin gene of gorilla and orangutan:

gorilla   GAA<sup>G</sup>TCTTTGAGAAATAAAACTGCACACACTGG
orangutan GGA<sup>CT</sup>TCTTTGAGAAATAAAACTGCACACACTGG

\[
L = \left[ \left( \frac{1}{4} \right) \left( \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \right) \right]^{30} \left[ \left( \frac{1}{4} \right) \left( \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right) \right]^2
\]

Plot of log-likelihood as a function of the quantity \( \alpha t \)

Maximum likelihood estimate (MLE) of \( \alpha t \) is 0.021753
number of substitutions = rate × time

Overall substitution rate is $3\alpha$, so the expected number of substitutions $(v)$ is

$$v = 3\alpha t$$
Rate and time are confounded

Later this week you will be introduced to models in which constraints on times can be used to infer rates (and vice versa), but without some extra information or constraints, sequence data allow only estimation of the **number** of substitutions.
A convenient convention

Because rate and time are confounded, it is convenient to arbitrarily standardize things by setting the rate to a value such that one substitution is expected to occur in one unit of time for each site.

This results in “time” (the length of a branch) being measured in units of evolutionary distance (expected number of substitutions per site) rather than years (or some other calendar unit).

\[ \nu = 3\alpha t \]

Setting \( \alpha = 1/3 \) results in \( \nu \) equalling \( t \)
Evolutionary distances for several common models

<table>
<thead>
<tr>
<th>Model</th>
<th>Expected no. substitutions: $v = {r}t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC69</td>
<td>$v = {3\alpha} t$</td>
</tr>
<tr>
<td>F81</td>
<td>$v = {2\mu(\pi_R\pi_Y + \pi_A\pi_G + \pi_C\pi_T)} t$</td>
</tr>
<tr>
<td>K80</td>
<td>$v = {\beta(\kappa + 2)} t$</td>
</tr>
<tr>
<td>HKY85</td>
<td>$v = {2\mu [\pi_R\pi_Y + \kappa(\pi_A\pi_G + \pi_C\pi_T)]} t$</td>
</tr>
</tbody>
</table>

In the formulas above, the overall rate $r$ (in curly brackets) is a function of all parameters in the substitution model.

Note that one of the parameters of the substitution model can always be determined from the branch length (using our convention that $v = t$).

Typically, all other model parameters are estimated for the entire tree (for example, each branch uses the same value of $\kappa$)
Likelihood of an unrooted tree
(data shown for only one site)

States at the tips are observed.

Ancestral states like this are not really known - we will address this in a minute.

Arbitrarily chosen to serve as the root node
Likelihood for site $k$

$$L_k = \frac{1}{4} \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_1/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_2/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4\nu_3/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4\nu_4/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4\nu_5/3} \right]$$

$\nu_5$ is the expected number of substitutions for just this one branch.

Note use of the AND probability rule
Brute force approach would be to calculate $L_k$ for all 16 combinations of ancestral states and sum them.

Note use of the OR probability rule.
Pruning algorithm
(same result, less time)

Many calculations can be done just once and then reused several times

Substitution Models
The four bases (A, C, G, T) are expected to be equally frequent in sequences ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$).

Assumes same rate for all types of substitution ($r_{A\to C} = r_{A\to G} = r_{A\to T} = r_{C\to G} = r_{C\to T} = r_{G\to T} = \alpha$).

Usually described as a 1-parameter model (the parameter being the branch length).

- Remember, however, that each branch in a tree can have its own length, so there are really as many parameters in the model as there are edges in the tree!

Assumes substitution is a Markov process...

What is a Markov model?

A substitution occurs, changing T to C

Lineage starts with base T at some site

To predict which base will be present after some time $t$ we need know only which base was present at time 0 (C in this case).

If it is irrelevant that there was a T present at this site before time 0, then this is a Markov model.
Transition Probabilities

A substitution occurs, changing T to C

Lineage starts with base T at some site

The transition probability $P_{CC}(t)$ gives the probability that there is a C present at a site after time t given that there was a C present at time 0.

Note: the term transition here comes from the terminology of stochastic processes and refers to any change of state (and even non-changes!). If this kind of transition represents a change from one nucleotide state to a different nucleotide state, it could thus be either a transition-type or a transversion-type substitution.
Jukes-Cantor transition probabilities

Here is the probability that a site starting in state T will end up in state G after time $t$ when the individual substitution rates are all $\alpha$:

$$P_{TG}(t) = \frac{1}{4} \left( 1 - e^{-4\alpha t} \right)$$

The JC69 model has only one unknown quantity: $\alpha t$

(The symbol $e$ represents the base of the natural logarithms: its value is 2.718281828459045...)

Where does a transition probability formula such as this come from?
"ACHNyons" vs. substitutions

When an *achnyon* occurs, any base can appear in a sequence.

Note: *achnyon* is *my term* for this make-believe event. You will not see this term in the literature.

If the base that appears is *different* from the base that was already there, then a *substitution* event has occurred.

The rate ($\alpha$) at which any *particular* substitution occurs will be $1/4$ the *achnyon* rate ($\mu$).

That is, $\alpha = \mu/4$

(or $\mu = 4\alpha$)
Deriving a transition probability

Calculate the probability that a site currently T will change to G over time $t$ when the rate of this particular substitution is $\alpha$:

\[
\Pr(\text{zero achnyons}) = e^{-\mu t} \quad \text{(Poisson probability of zero events)}
\]

\[
\Pr(\text{at least 1 achnyon}) = 1 - e^{-\mu t}
\]

\[
\Pr(\text{last achnyon results in base G}) = \frac{1}{4}
\]

\[
\Pr(\text{end in G | start in T}) = \frac{1}{4} \left( 1 - e^{-\mu t} \right)
\]

Remember that the rate ($\alpha$) of any particular substitution is one fourth the achnyon rate ($\mu$):

\[
P_{GT}(t) = \frac{1}{4} \left( 1 - e^{-4\alpha t} \right)
\]
Expected number of substitutions

If the base that appears is different from the base that was already there, then a substitution event has occurred.

The overall substitution rate will be $3/4$ the achnyon rate.

$$\nu = \frac{3}{4} \mu t = 3\alpha t$$
Transition Probabilities: Remarks

\[ P_{TA}(t) = \frac{1}{4} \left( 1 - e^{-4\alpha t} \right) \]
\[ P_{TC}(t) = \frac{1}{4} \left( 1 - e^{-4\alpha t} \right) \]
\[ P_{TG}(t) = \frac{1}{4} \left( 1 - e^{-4\alpha t} \right) \]
\[ P_{TT}(t) = \frac{1}{4} \left( 1 - e^{-4\alpha t} \right) \]

These should add to 1.0 because T must change to something!

Doh! Something must be wrong here...
Transition Probabilities: Remarks

\begin{align*}
P_{TA}(t) &= \frac{1}{4} (1 - e^{-4\alpha t}) \\
P_{TC}(t) &= \frac{1}{4} (1 - e^{-4\alpha t}) \\
P_{TG}(t) &= \frac{1}{4} (1 - e^{-4\alpha t}) \\
P_{TT}(t) &= \frac{1}{4} (1 - e^{-4\alpha t}) + e^{-4\alpha t}
\end{align*}

Forgot to account for the possibility of no actions over time $t$
Equilibrium frequencies

• The JC69 model assumes that the frequencies of the four bases (A, C, G, T) are equal
• The equilibrium relative frequency of each base is thus 0.25
• Why are they called equilibrium frequencies?
Imagine a bottle of perfume has been spilled in room A.

The doors to the other rooms are closed, so the perfume has, thus far, not been able to spread.

What would happen if we opened all the doors?
If the doors are suddenly opened, the perfume would begin diffusing from the area of highest concentration to lowest.

Molecules of perfume go both ways through open doors, but more pass one way than another, leading to a net flow from room A to rooms B and C.

In the instant that the doors are opened, A is losing perfume molecules at *twice the rate* each of the other rooms is gaining molecules. As diffusion progresses, however, the rate of loss from A drops, approaching an equilibrium.
Eventually, all four rooms have essentially the same concentration of perfume.

Molecules still move through doors, but now the rates are the same in all directions.

Back to sequence evolution: assume a sequence began with only A nucleotides (a poly-A sequence). Over time, substitution would begin converting some of these As to Cs, Gs, and Ts, just as the perfume diffused into adjacent rooms.
Pr(A|A) and Pr(A|T) as a function of time

Upper curve assumes we started with A at time 0. Over time, the probability of still seeing an A at this site **drops** because rate of changing to one of the other three bases is $3\alpha$ (so rate of staying the same is $-3\alpha$).

The equilibrium relative frequency of A is 0.25

Lower curve assumes we started with some state **other** than A (T is used here). Over time, the probability of seeing an A at this site **grows** because the rate at which the current base will change into an A is $\alpha$. 

$P_{AA}(t)$

$P_{TA}(t)$
**JC69 rate matrix**

<table>
<thead>
<tr>
<th>From</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$-3\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>C</td>
<td>$\alpha$</td>
<td>$-3\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>G</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$-3\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>T</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$-3\alpha$</td>
</tr>
</tbody>
</table>

1 parameter: \( \alpha \)

K80 (or K2P) rate matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>From</strong></td>
<td><strong>To</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-(\alpha - 2\beta)</td>
<td>(\beta)</td>
<td>(\alpha)</td>
<td>(\beta)</td>
</tr>
<tr>
<td>C</td>
<td>(\beta)</td>
<td>-(\alpha - 2\beta)</td>
<td>(\beta)</td>
<td>(\alpha)</td>
</tr>
<tr>
<td>G</td>
<td>(\alpha)</td>
<td>(\beta)</td>
<td>-(\alpha - 2\beta)</td>
<td>(\beta)</td>
</tr>
<tr>
<td>T</td>
<td>(\beta)</td>
<td>(\alpha)</td>
<td>(\beta)</td>
<td>-(\alpha - 2\beta)</td>
</tr>
</tbody>
</table>

2 parameters:
- \(\alpha\)
- \(\beta\)

Transition rate

Transversion rate

K80 rate matrix
(looks different, but actually the same)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$-\beta(\kappa + 2)$</td>
<td>$\beta$</td>
<td>$\kappa \beta$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>C</td>
<td>$\beta$</td>
<td>$-\beta(\kappa + 2)$</td>
<td>$\beta$</td>
<td>$\kappa \beta$</td>
</tr>
<tr>
<td>G</td>
<td>$\kappa \beta$</td>
<td>$\beta$</td>
<td>$-\beta(\kappa + 2)$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>T</td>
<td>$\beta$</td>
<td>$\kappa \beta$</td>
<td>$\beta$</td>
<td>$-\beta(\kappa + 2)$</td>
</tr>
</tbody>
</table>

All I’ve done is re-parameterize the rate matrix, letting $\kappa$ equal the transition/transversion rate ratio

$k = \frac{\alpha}{\beta}$

Note: the K80 model is identical to the JC69 model if $\kappa = 1$ ($\alpha = \beta$)
F81 rate matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$-\mu(1 - \pi_A)$</td>
<td>$\pi_C \mu$</td>
<td>$\pi_G \mu$</td>
<td>$\pi_T \mu$</td>
</tr>
<tr>
<td>C</td>
<td>$\pi_A \mu$</td>
<td>$-\mu(1 - \pi_C)$</td>
<td>$\pi_G \mu$</td>
<td>$\pi_T \mu$</td>
</tr>
<tr>
<td>G</td>
<td>$\pi_A \mu$</td>
<td>$\pi_C \mu$</td>
<td>$-\mu(1 - \pi_G)$</td>
<td>$\pi_T \mu$</td>
</tr>
<tr>
<td>T</td>
<td>$\pi_A \mu$</td>
<td>$\pi_C \mu$</td>
<td>$\pi_G \mu$</td>
<td>$-\mu(1 - \pi_T)$</td>
</tr>
</tbody>
</table>

Note: the F81 model is identical to the JC69 model if all base frequencies are equal

5 parameters:
\[ \kappa \]
\[ \beta \]
\[ \pi_A \]
\[ \pi_C \]
\[ \pi_G \]

A dash means equal to negative sum of other elements on the same row.

Note: the HKY85 model is identical to the F81 model if \( \kappa = 1 \). If, in addition, all base frequencies are equal, it is identical to JC69.

F84 vs. HKY85

F84 model:

$$\mu$$ rate of process generating *all types of substitutions*

$$k\mu$$ rate of process generating *only transitions*

Becomes F81 model if $$k = 0$$

HKY85 model:

$$\beta$$ rate of process generating *only transversions*

$$k\beta$$ rate of process generating *only transitions*

Becomes F81 model if $$k = 1$$

**GTR rate matrix**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>—</td>
<td>(\pi_C a \mu)</td>
<td>(\pi_G b \mu)</td>
<td>(\pi_T c \mu)</td>
</tr>
<tr>
<td>C</td>
<td>(\pi_A a \mu)</td>
<td>—</td>
<td>(\pi_G d \mu)</td>
<td>(\pi_T e \mu)</td>
</tr>
<tr>
<td>G</td>
<td>(\pi_A b \mu)</td>
<td>(\pi_C d \mu)</td>
<td>—</td>
<td>(\pi_T f \mu)</td>
</tr>
<tr>
<td>T</td>
<td>(\pi_A c \mu)</td>
<td>(\pi_C e \mu)</td>
<td>(\pi_G f \mu)</td>
<td>—</td>
</tr>
</tbody>
</table>

Identical to the F81 model if \(a = b = c = d = e = f = 1\). If, in addition, all the base frequencies are equal, GTR is identical to JC69. If \(a = c = d = f = \beta\) and \(b = e = \kappa \beta\), GTR becomes the HKY85 model.

Rate Heterogeneity
# Green Plant rbcL

First 88 amino acids (translation is for *Zea mays*).

<table>
<thead>
<tr>
<th>Organism</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chara</td>
<td>AAAGATTACAGATTAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCGTGTAACTCCA</td>
</tr>
<tr>
<td>Chlorella</td>
<td>GAGATTACAGATTAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
<tr>
<td>Volvox</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCGTG</td>
</tr>
<tr>
<td>Conocephalum</td>
<td>CATACTGACATTTAGCTGCATTTCGTGTAACTCCA</td>
</tr>
<tr>
<td>Bazzania</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
<tr>
<td>Anthoceros</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
<tr>
<td>Osmunda</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
<tr>
<td>Lycopodium</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
<tr>
<td>Ginkgo</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
<tr>
<td>Picea</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
<tr>
<td>Iris</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
<tr>
<td>Asplenium</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
<tr>
<td>Nicotiana</td>
<td>CTGAACTTACTATACTCCCTGAGTATAAAACTAAAGATACTGACATTTAGCTGCATTTCG</td>
</tr>
</tbody>
</table>

All four bases are observed at some sites... while at other sites, only one base is observed.
Site-specific rates

Each defined subset (e.g. gene, codon position) has its own relative rate

$r_1$ applies to subset 1
(e.g. sites 1 - 1000)

$r_2$ applies to subset 2
(e.g. sites 1001-2000)
Site-specific rates

\[ L = \text{Pr}(D_1 | r_1) \cdots \text{Pr}(D_{1000} | r_1) \quad \text{Pr}(D_{1001} | r_2) \cdots \text{Pr}(D_{2000} | r_2) \]

\[ r_1 = 1.2 \quad r_2 = 0.8 \]
Site-specific rates

JC69 transition probabilities that would be used for every site if rate homogeneity were assumed:

\[ P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \]

\[ P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \]
Site specific rates

JC69 transition probabilities that would be used for sites in gene 1:

\[ P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4r_1 \alpha t} \]

\[ P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4r_1 \alpha t} \]

JC69 transition probabilities that would be used for sites in gene 2:

\[ P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4r_2 \alpha t} \]

\[ P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4r_2 \alpha t} \]
Ok, I am going to divide you into 2 groups based on the color of your head, and everyone in each group will get a coat of the average size for their group. Very sorry if this does not work well for some people who are unusually large or small compared to their group.
Site-specific Approach

Pro: costs less: need to buy just one coat for every person
Con: every person in a group has to wear the same size coat, so the fit will be poor for some people if they are much bigger or smaller than the average size for the group in which they have been placed
Mixture Models

All relative rates applied to every site

\[ L_i = \Pr(D_i \mid r_1) \Pr(r_1) + \Pr(D_i \mid r_2) \Pr(r_2) \]

Common examples

- Invariable sites (I) model
- Discrete Gamma (G) model
Ok, I am going to give each of you 2 coats: use the one that fits you best and throw away the other one. This costs twice as much for me, but on average leads to better fit for you. I have determined the two sizes of coats based on the distribution of your sizes.

Mixture Model Approach
Mixture Model Approach

Pro: every person experiences better fit because they can choose the size coat that fits best
Con: costs more because two coats much be provided for each person
Invariable Sites Model

A fraction $p_{invar}$ of sites are assumed to be invariable (i.e. rate = 0.0)

$$L_i = \Pr(D_i | r_1)p_{invar} + \Pr(D_i | r_2)(1 - p_{invar})$$

$r_1 = 0.0$

$r_2 = \frac{1}{1 - p_{invar}}$

Allows for the possibility that any given site could be variable or invariable

Invariable sites model

If site $i$ is a constant site, both terms will contribute to the site likelihood:

$$L_i = \Pr(D_i|0.0)p_{\text{invar}} + \Pr(D_i|r_2)(1 - p_{\text{invar}})$$

If site $i$ is a variable site, there is no way to explain the data with a zero rate, so the first term is zero:

$$L_i = \Pr(D_i|0.0)p_{\text{invar}} + \Pr(D_i|r_2)(1 - p_{\text{invar}})$$
Discrete Gamma Model

No relative rates are exactly 0.0, and all are equally probable

\[ L = \left( \frac{1}{4} \right) \Pr(D_i | r_1) + \left( \frac{1}{4} \right) \Pr(D_i | r_2) + \left( \frac{1}{4} \right) \Pr(D_i | r_3) + \left( \frac{1}{4} \right) \Pr(D_i | r_4) \]

Relative rates are constrained to a discrete gamma distribution

Number of rate categories can vary (4 used here)


Relative rates in 4-category case

Boundaries are placed so that each category represents 1/4 of the distribution (i.e. 1/4 of the area under the curve).

Boundary between 1st and 2nd categories

Boundary between 2nd and 3rd categories

Boundary between 3rd and 4th categories

\[ r_1 = 0.137 \]
\[ r_2 = 0.477 \]
\[ r_3 = 1.000 \]
\[ r_4 = 2.386 \]
Gamma distributions

The mean equals 1.0 for all three of these distributions.

- Larger $\alpha$ means less heterogeneity.

Note: the symbol $\alpha$ here is not the same as the $\alpha$ used in JC transition probability formulas!
Codon models
The Genetic Code

First 12 nucleotides at the 5' end of the \textit{rbcL} gene in corn:

\begin{align*}
5' - & \text{ATG|TCA|CCA|CAA} - 3' \\
3' - & \text{TAC|AGT|GGT|GTT} - 5' \\
\end{align*}

\[ \text{DNA double helix} \]

\begin{itemize}
  \item \textbf{transcription} \\
  \begin{align*}
  5' - & \text{AUG|UCA|CCA|CAA} - 3' \\
  & \text{mRNA}
  \end{align*}
  \item \textbf{translation} \\
  \begin{align*}
  N- & \text{Met|Ser|Pro|Gln-C} \\
  \end{align*}
\end{itemize}

Codon models treat codons as the independent units, not individual nucleotide sites.

\[ \text{http://www.langara.bc.ca/biology/mario/Assets/Geneticcode.jpg} \]
First codon models

• Muse and Gaut model (MG94) is simplest
  \[ \alpha = \text{synonymous substitution rate} \]
  \[ \beta = \text{nonsynonymous substitution rate} \]
  \[ \pi_A, \pi_C, \pi_G, \pi_T = \text{base frequencies} \]

• Goldman and Yang model (GY94) similar
  – accounts for synon./nonsynon. and trs/trv bias and amino acid properties (later simplified, see Yang et al.)


Table 1. Part of Muse and Gaut’s $61 \times 61$ instantaneous rate matrix

<table>
<thead>
<tr>
<th>Codon before substitution (the ‘from’ state)</th>
<th>Codon after substitution (the ‘to’ state)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TTT (Phe)</td>
</tr>
<tr>
<td>TTT (Phe)</td>
<td>- - -</td>
</tr>
<tr>
<td>TTC (Phe)</td>
<td>$\alpha\pi_T$</td>
</tr>
<tr>
<td>TTA (Leu)</td>
<td>$\beta\pi_T$</td>
</tr>
<tr>
<td>TTG (Leu)</td>
<td>$\beta\pi_T$</td>
</tr>
<tr>
<td>CTT (Leu)</td>
<td>$\beta\pi_T$</td>
</tr>
<tr>
<td>CTC (Leu)</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>GGG (Gly)</td>
<td>0</td>
</tr>
</tbody>
</table>

Note that it is still easy for the change CTT → TTA to occur, it just requires more than one instant of time.

Instantaneous rate is 0.0 if two or more nucleotides must change during the codon transition.

Interpreting codon model results

$\omega = \beta / \alpha$ is the nonsynonymous/synonymous rate ratio

<table>
<thead>
<tr>
<th>Omega</th>
<th>Mode of Selection</th>
<th>Example(s)</th>
</tr>
</thead>
</table>
| $\omega < 1$ | stabilizing selection  
(nucleotide substitutions rarely change the amino acid) | functional protein coding genes                             |
| $\omega = 1$ | neutral evolution  
(synonymous and nonsynonymous substitutions occur at the same rate) | pseudogenes                                                  |
| $\omega > 1$ | positive selection  
(nucleotide substitutions often change the amino acid) | envelope proteins in viruses under active positive selection |