Maximum Likelihood in Phylogenetics

21 July 2015

Workshop on Molecular Evolution
Woods Hole, Mass.

Paul O. Lewis
Department of Ecology & Evolutionary Biology
Goals

* Explain jargon
* Increase comfort level
* Provide background

In other words...give a hand up
Tree jargon

- **split (bipartition)**
  - also written AB|CDE
  - or portrayed **---**

- **terminal or tip node**
  - (or leaf, degree 1)

- **interior node**
  - (or vertex, degree 3+)

- **edge (branch)**

- **root node of tree**
  - (degree 2)
The Plan

- Probability review
- Likelihood
- Substitution models
  - Markov model basics
  - Transition probabilities
  - Survey of models
  - Rate heterogeneity
  - Codon models
  - Amino acid 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

\[ \text{AND} \begin{array}{c}
\square
\end{array} \text{ AND } \begin{array}{c}
\square
\end{array} \]

\[ \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} \]
AND rule in phylogenetics

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.
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

\[
\text{OR} \quad \begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc
\end{array} \\
\begin{array}{c}
\bigcirc \\
\bigcirc \\
\bigcirc
\end{array}
\]

\[
(1/6) + (1/6) = 1/3
\]
Combining AND and OR

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

(1/36) + (1/36) + (1/36) + (1/36) + (1/36) + (1/36) = 1/6

1 and 6
2 and 5
3 and 4
4 and 3
5 and 2
6 and 1
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...

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

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 \( \Pr(B) \)

\[ \Pr(A \text{ and } B) = \Pr(A) \Pr(B) \]
A possible example of non-independence

A = you saw this at 4pm

B = you will swim with us at 5pm

$$\Pr(\mathcal{B} | \mathcal{A}) \neq \Pr(\mathcal{B})$$

You might be more willing to go swimming if you had no knowledge of Katherine’s cruising habits.

Track of the Great White Shark named Katherine (from ocearch.org)
Not to worry: Katherine was off the coast of North Carolina 6 days ago.
Non-independence in molecular evolution

A = state in descendant
B = state in ancestor

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

That is, the state in the descendant depends on the state in the ancestor

The state in the descendant also depends on the amount time \( t \) separating ancestor and descendant and the rate of substitution \( r \)

\[ \text{Pr}(A|B, t, r) \]
Conditional Independence

Assume both A and B depend on C:

\[
\Pr(A|C) \neq \Pr(A) \quad \Pr(B|C) \neq \Pr(B)
\]

If we can say this...

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

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

\[
\Pr(A \text{ and } B|C) = \Pr(A|C) \Pr(B|C)
\]
Conditional independence in molecular evolution

The site data patterns AGG and TCC are assumed by most models to be conditionally independent.

The patterns both depend on the underlying tree (including edge lengths) and the substitution model.

\[
\Pr(\text{AGG and TCC}|\text{tree, model}) = \Pr(\text{AGG}|\text{tree, model}) \cdot \Pr(\text{TCC}|\text{tree, model})
\]
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)

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</td>
<td>Not surprised at all</td>
</tr>
</tbody>
</table>

winning model maximizes likelihood (and thus minimizes surprise)
### Likelihood: why a new term?

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Fair coin model</th>
<th>Two-heads model</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

**Probabilities** of data outcomes given one particular model sum to 1.0

**Likelihoods** of models given one particular data outcome are not expected to sum to 1.0
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 lie between 0 and 1, which means log(x) will always be negative if x represents a probability.

In this talk (and in phylogenetics in general), ln(x) = log(x)
Likelihood of a single sequence

First 32 nucleotides of the \( \psi \eta \)-globin gene of gorilla:

\[
\text{GAAGTCCTTGAGAGAAATAAACTGCACACACTGG}
\]

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

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 \textbf{As} as there are \textbf{Cs, Gs and Ts}.
Model ranking using LRT or AIC

Likelihood Ratio Tests (LRT) and the Akaike Information Criterion (AIC) provide two ways to evaluate whether an **unconstrained** model fits the data significantly better than a **constrained** version of the same model.

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

$$\log L_{\text{unconstrained}} = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)$$

$$= 12 \log(0.375) + 7 \log(0.219) + 7 \log(0.219) + 6 \log(0.187)$$

$$= -43.1$$

*This model has 3 estimated parameters*

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

$$\log L_{\text{constrained}} = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)$$

$$= 12 \log(0.25) + 7 \log(0.25) + 7 \log(0.25) + 6 \log(0.25)$$

$$= -44.4$$

*This model has 0 estimated parameters*
Likelihood Ratio Test (LRT)

Calculate the likelihood ratio test statistic:

\[
R = -2 \left[ \log(L_{\text{constrained}}) - \log(L_{\text{unconstrained}}) \right]
\]
\[
= -2 \left[ -44.4 - (-43.1) \right]
\]
\[
= 2.6
\]

(Note that the log-likelihoods used in the test statistic have been maximized under each model separately)

“unconstrained” does fit better than “constrained” (-43.1 > -44.4), but not significantly better (P = 0.457, chi-squared with 3 d.f.*

*The number of degrees of freedom equals the difference between the two models in the number of estimated parameters. In this case, unconstrained has 3 parameters and constrained has 0, so d.f. = 3 - 0 = 3
Akaike Information Criterion (AIC)

Calculate AIC for each model:

\[ AIC = 2k - 2 \log(\max(L)) \]

\[ AIC_{\text{unconstrained}} = 2(3) - 2(-43.1) = 92.2 \]

\[ AIC_{\text{constrained}} = 2(0) - 2(-44.4) = 88.8 \]

The constrained model is a better choice than the unconstrained model according to AIC.

Dave Swofford will give you a much more complete explanation of LRT and AIC this afternoon.
Likelihood of the simplest tree

sequence 1 —— sequence 2

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

\[ L = L_1 L_2 \]

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

\[ \Pr(G) \quad \Pr(G|G, \alpha t) \quad \Pr(A) \quad \Pr(G|A, \alpha t) \]

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

First 32 nucleotides of the \( \psi\eta \)-globin gene of gorilla and orangutan:

- **gorilla**: G\textcolor{red}{A}AGTCCTTTGAGAAATAAAACTGCACACACACTGG
- **orangutan**: G\textcolor{red}{C}GT\textcolor{red}{C}CTTTGAGAAATAAAACTGCACACACACTGG

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

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 \times \text{time}

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

\[ v = 3\alpha t \]
On Friday, Alexei Drummond will introduce models that allow separate estimation of rates and times, but without extra information/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).

\[ \nu = 3 \alpha t \]

\[ \nu = 3 \left( \frac{1}{3} \right) t \]

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.

One substitution model parameter is always determined from the edge length (using our convention that $v = t$); the others are usually global (i.e. same value applies to all edges).
Likelihood of an unrooted tree
(data shown for only one site)

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

States at the tips are observed.

Arbitrarily chosen to serve as the root node.

Ancestral states like this are not really known - we will address this in a minute.
Likelihood for site $k$

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

Note use of the AND probability rule

$\nu_5$ is the expected number of substitutions for just this one branch.
Brute force approach would be to calculate $L_k$ for all 16 combinations of ancestral states and sum them.
Pruning algorithm
(same result, less time)

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

Substitution Models
Jukes-Cantor (JC69) model

- 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 ($r_{A \rightarrow C} = r_{A \rightarrow G} = r_{A \rightarrow T} = r_{C \rightarrow G} = r_{C \rightarrow T} = r_{G \rightarrow T} = \alpha$)
- Usually described as a 1-parameter model (the parameter being the edge length)
  - Remember, however, that each edge 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 process?

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.
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.

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

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

(or $\mu = 4\alpha$)

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

$\text{ACHN} = "$Anything Can Happen Now"$
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 \( \alpha \):

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

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

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

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

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 $\frac{3}{4}$ the achnyon rate.

$\nu = \frac{3}{4} \mu t = 3\alpha t$

$\mu$ = mutation rate

$\nu$ = substitution rate

$\alpha$ = selection coefficient
Transition Probabilities: Remarks

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

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

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

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

Forgot to account for the possibility of *no* acnyons 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?
Equilibrium Frequencies

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.
Equilibrium Frequencies

Eventually, all 3 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 $-3\alpha$).

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 $\alpha$.

The equilibrium relative frequency of A is 0.25
### 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

\[
\begin{array}{cccc}
A & C & G & T \\
\hline
A & -\alpha - 2\beta & \beta & \alpha & \beta \\
C & \beta & -\alpha - 2\beta & \beta & \alpha \\
G & \alpha & \beta & -\alpha - 2\beta & \beta \\
T & \beta & \alpha & \beta & -\alpha - 2\beta \\
\end{array}
\]

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*

\[
\kappa = \frac{\alpha}{\beta}
\]

Note: the K80 model is identical to the JC69 model if $\kappa = 1$ ($\alpha = \beta$)
Transition/transversion ratio (tratio) versus Transition/transversion rate ratio (kappa)

Cobbler analogy:

- 4 cobblers in a factory make loafers
- 8 cobblers in the factory make work boots
- all cobblers produce the same number of shoes per unit time, regardless of shoe type
- what is the loafer/boot rate ratio and how does that compare to the loafer/boot ratio?

The loafer/boot rate ratio is 1.0 because each cobbler cranks out shoes at the same rate.

The loafer/boot ratio, however, is 0.5 because there are twice as many cobblers making boots as there are cobblers making loafers.

There are 8 possible transversion-type substitutions and only 4 possible transition-type substitutions: the transition/transversion ratio is thus 0.5 when the transition/transversion rate ratio is 1.
### 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

---

### HKY85 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>$\pi_C \beta$</td>
<td>$\pi_G \beta \kappa$</td>
<td>$\pi_T \beta$</td>
</tr>
<tr>
<td>C</td>
<td>$\pi_A \beta$</td>
<td>−</td>
<td>$\pi_G \beta$</td>
<td>$\pi_T \beta \kappa$</td>
</tr>
<tr>
<td>G</td>
<td>$\pi_A \beta \kappa$</td>
<td>$\pi_C \beta$</td>
<td>−</td>
<td>$\pi_T \beta$</td>
</tr>
<tr>
<td>T</td>
<td>$\pi_A \beta$</td>
<td>$\pi_C \beta \kappa$</td>
<td>$\pi_G \beta$</td>
<td>−</td>
</tr>
</tbody>
</table>

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.

5 parameters: $\kappa$, $\beta$, $\pi_A$, $\pi_C$, $\pi_G$

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

---

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
\[ \kappa\beta \] rate of process generating only transitions

Becomes F81 model if \( \kappa = 1 \)

### 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>$\pi_Ca\mu$</td>
<td>$\pi_Gb\mu$</td>
<td>$\pi_Tc\mu$</td>
</tr>
<tr>
<td>C</td>
<td>$\pi_Aa\mu$</td>
<td>—</td>
<td>$\pi_Gd\mu$</td>
<td>$\pi_Te\mu$</td>
</tr>
<tr>
<td>G</td>
<td>$\pi_Ab\mu$</td>
<td>$\pi_Cd\mu$</td>
<td>—</td>
<td>$\pi_Tf\mu$</td>
</tr>
<tr>
<td>T</td>
<td>$\pi_Ac\mu$</td>
<td>$\pi_Ce\mu$</td>
<td>$\pi_Gf\mu$</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 \textit{rbcL}

First 88 amino acids (translation is for \textit{Zea mays})

\begin{verbatim}
Chara (green alga; land plant lineage) AAAGATTACAGATTACATTACATTACCTGAGTATAAAAACTAAAGATACTGAGCATTTAGCTGCATTTCTCTGAGCTAACCTCCA
Volvox (green alga) T.C.T.A.C.A...C.G.TA.GC.A........C.A........A.G....
Conceophalum (liverwort) T.C.T...G.T.G........A...T........A.A.A.G...T
Bazzania (moss) T.........C.T.........G.A.G.G.C.G.A.T.G.........A.A.G....
Osmunda (fern) T.C.G.C.C.T.G.C.G.G.T.G.........A.A.G...C
Ginkgo (gymnosperm; Ginkgo biloba) ............G...........T........G.A.....C...A.G.....C
Picea (gymnosperm; spruce) .....................T........G.A.....C...A.G.....C
Nicotiana (flowering plant; tobacco) ........G........C........T..T.G.A.G.T........C.C...G...C.A........A.C........T
\end{verbatim}

CAACCTGGCGTTCCACCTGAAGAAGCAGGGGCTGCAGTAGCTGCAGAATCTTCTACTGGTACATGGACTACTGTTTGGACTGACGGATTAACTAGTTTGGACCGATACAAAGGAAGATGCTACGATATTGAA

...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)

Relative rates have mean 1:

$$\frac{r_1 + r_2}{2} = 1$$

More generally:

$$r_1 \, p(r_1) + r_2 \, p(r_2) = 1$$
Site-specific rates

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

Gene 1

Gene 2

\[ r_1 = 1.2 \]

\[ r_2 = 0.8 \]
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:

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

\[
P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4r_2 \alpha t}
\]
Sit-specific Approach

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

**Good:** costs less: need to buy just one coat for every person
**Bad:** 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

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

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**

*Good:* every person experiences better fit because they can choose the size coat that fits best

*Bad:* costs more because two coats much be provided for each person
Invariable Sites Model

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

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

\[ r_1 = 0.0 \]
\[ r_2 = \frac{1}{1 - p_{\text{invar}}} \]

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:

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

No relative rate is exactly 0.0, and all are equally probable

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 relative rate equals 1.0 for all three of these distributions. Larger shape means less heterogeneity.
Codon models

Joe Bielawski will discuss codon models in *much* greater detail tomorrow.
The Genetic Code

First 12 nucleotides at the 5' end of the *rbcL* gene in corn:

\[
5' - ATG | TCA | CCA | CAA - 3' \quad \text{coding strand}
\]

\[
3' - TAC | AGT | GGT | GTT - 5' \quad \text{template strand}
\]

DNA double helix

\[
5' - AUG | UCA | CCA | CAA - 3' \quad \text{mRNA}
\]

transcription

N-Met | Ser | Pro | Gln - C polypeptide

translation

Codon models treat codons as the independent units, not individual nucleotide sites.
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. 1998)


Table 1. Part of Muse and Gaut’s 61 × 61 instantaneous rate matrix\textsuperscript{a}

<table>
<thead>
<tr>
<th>Codon before substitution (the ‘from’ state)</th>
<th>Codon after substitution (the ‘to’ state)</th>
<th>TTT (Phe)</th>
<th>TTC (Phe)</th>
<th>TTA (Leu)</th>
<th>TTG (Leu)</th>
<th>CTT (Leu)</th>
<th>CTC (Leu)</th>
<th>GGG (Gly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT (Phe)</td>
<td></td>
<td>−−−</td>
<td>$\alpha\pi_C$</td>
<td>$\beta\pi_A$</td>
<td>$\beta\pi_G$</td>
<td>$\beta\pi_C$</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>TTC (Phe)</td>
<td>$\alpha\pi_T$</td>
<td>−−−</td>
<td>$\beta\pi_A$</td>
<td>$\beta\pi_G$</td>
<td>0</td>
<td>$\beta\pi_C$</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>TTA (Leu)</td>
<td>$\beta\pi_T$</td>
<td>$\beta\pi_C$</td>
<td>−−−</td>
<td>$\alpha\pi_G$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>TTG (Leu)</td>
<td>$\beta\pi_T$</td>
<td>$\beta\pi_C$</td>
<td>$\alpha\pi_A$</td>
<td>−−−</td>
<td>0</td>
<td>$\alpha\pi_C$</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>CTT (Leu)</td>
<td>$\beta\pi_T$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\alpha\pi_T$</td>
<td>−−−</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>CTC (Leu)</td>
<td></td>
<td>$\beta\pi_T$</td>
<td>0</td>
<td>0</td>
<td>$\alpha\pi_T$</td>
<td>−−−</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>GGG (Gly)</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>−−−</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 |
Amino acid models
### JC69 Flashback

**Q matrix** (instantaneous rates)

<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>(-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>

**P matrix** (transition probabilities)

\[
\begin{align*}
\frac{1}{4} + \frac{3}{4} e^{-4\alpha t} & & \frac{1}{4} (1 - e^{-4\alpha t}) & & \frac{1}{4} (1 - e^{-4\alpha t}) & & \frac{1}{4} (1 - e^{-4\alpha t}) \\
\frac{1}{4} (1 - e^{-4\alpha t}) & & \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} & & \frac{1}{4} (1 - e^{-4\alpha t}) & & \frac{1}{4} (1 - e^{-4\alpha t}) \\
\frac{1}{4} (1 - e^{-4\alpha t}) & & \frac{1}{4} (1 - e^{-4\alpha t}) & & \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} & & \frac{1}{4} (1 - e^{-4\alpha t}) \\
\frac{1}{4} (1 - e^{-4\alpha t}) & & \frac{1}{4} (1 - e^{-4\alpha t}) & & \frac{1}{4} (1 - e^{-4\alpha t}) & & \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}
\end{align*}
\]
A different path from Q to P

For many interesting models, it is not possible to obtain a formula for the transition probability.

We can, however, obtain transition probabilities numerically (i.e. obtain the value of the transition probability without plugging values into a formula)

\[ P(t) = e^{Qt} \]

\( \lambda_1, \lambda_2, \lambda_3, \) and \( \lambda_4 \) are the eigenvalues of \( Q \)

\[ P(t) = U \begin{pmatrix} e^{\lambda_1 t} & 0 & 0 & 0 \\ 0 & e^{\lambda_2 t} & 0 & 0 \\ 0 & 0 & e^{\lambda_3 t} & 0 \\ 0 & 0 & 0 & e^{\lambda_4 t} \end{pmatrix} U^{-1} \]

Matrix of eigenvectors of Q

Separation of Q into eigenvectors and a diagonal matrix of eigenvalues is known as diagonalization.
A different path from Q to P

Once freed from having to derive formulas for transition probabilities, we can use a great variety of Q matrices.

Dayhoff, Schwartz and Orcutt (1978; DSO78) identified 1572 accepted point mutations using closely-related sequences (<15% pairwise divergence), producing this matrix.

Kosiol and Goldman (2005) discussed ways of estimating a Q matrix from these numbers.


The elements of Q

Once the elements of the Q matrix are estimated, the Q matrix can be separated into a symmetric matrix R of exchangeabilities and a set of state frequencies.
Similarly, the GTR matrix can be separated into a symmetric R matrix and a diagonal matrix of frequencies.

\[
\begin{pmatrix}
- & a & b & c \\
- & a & d & e \\
- & b & d & f \\
- & c & e & f \\
\end{pmatrix}
\begin{pmatrix}
\pi_A & 0 & 0 & 0 & 0 \\
0 & \pi_C & 0 & 0 \\
0 & 0 & \pi_G & 0 \\
0 & 0 & 0 & \pi_T \\
\end{pmatrix}
\]

R matrix

Frequencies
What does all this accomplish?

• Empirical Q matrix derived from many closely-related pairwise comparisons
• Fixed Q matrix can be extrapolated using diagonalization to generate a P matrix for any desired value of $t$
• This model has 0 parameters!
• Models generic features of protein evolution; does not necessarily reflect your particular sequences
• Fixed frequencies can be swapped with more appropriate set (locally estimated)
Ways to improve

• Base everything on a much larger protein database (JTT model)

• Avoid need to use closely-related sequence pairs by obtaining ML estimate of Q matrix (WAG model)

• Add rate heterogeneity to ML estimation of Q matrix (LG model)


The End