Some of these slides have been borrowed from Dr. Paul Lewis, Dr. Joe Felsenstein. Thanks!

Paul has many great tools for teaching phylogenetics at his web site:

http://hydrodictyon.eeb.uconn.edu/people/plewis

Balanced Minimum Evolution

Pauplin (2000) showed that you can calculate a tree length from the pairwise distances without calculating branch lengths. The key is weighting the distances:

$$l = \sum_{i}^{N} \sum_{j=i+1}^{N} w_{ij} d_{ij}$$

where:

$$w_{ij} = \frac{1}{2^{n(i,j)}}$$

and $n(i, j)$ is the number of nodes on the path from i to j.

Balanced Minimum Evolution

Desper and Gascuel (2002, 2004) called minimizing this estimate of the tree length Balanced Minimum Evolution and showed that it is equivalent to a form of weighted least squares in which distances are down-weighted by an exponential function of the topological distances between the leaves.

Desper and Gascuel (2005) showed that NJ is star decomposition under BME. See Gascuel and Steel (2006).

NJ review

Neighbor-joining is a quick $O(N^3)$ algorithm for estimating the balanced minimum evolution tree.

Performance (based on simulation studies) is often close to the performance of searches under minimum evolution.

Branch length estimates in neighbor joining do not take into account the higher variance associated with large dissimilarity measurements. Thus the method tends to be sensitive to having long distances in the input data.

Distance-based phylogenetics review

Least squares and minimum evolution try to find trees and branch lengths such that the path length (or "patristic distance") between all taxa is as close as possible to the dissimilarity that is based on your data.

Frequently the observed p-distances must be corrected for multiple hits using models of character evolution.

Even when using corrected distances, distance-based approaches suffer from not adequately enforcing all of the natural constraints on the processes that generate the data.

Weaknesses in distance-based approaches

The fact that different errors in the pairwise distances is (the generalized least-square approach outlined by Bulmer, 1991) is computationally expensive.

Pairwise distance estimates, even when corrected, do not benefit from insights about which positions are evolving fast and which are slow (or other higher level patterns of sequence evolution).

Reconstructed tree is not forced to be compatible with reasonable constraints on sequence evolution.

Farris' example

Taxon	Character State
1	A
2	C
3	G
4	T

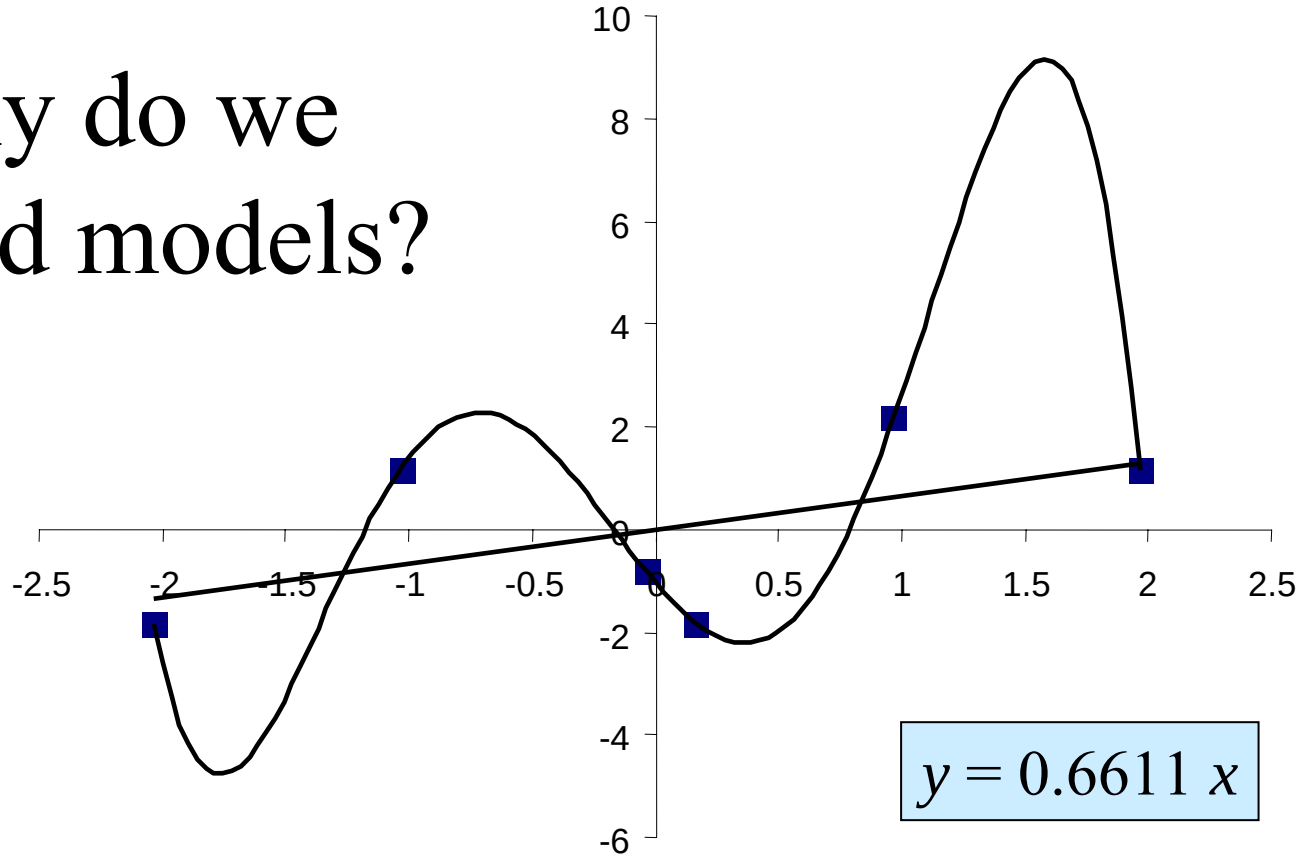
Another example

Taxon			
A	1000000	11111	00000
B	0100000	11111	11111
C	0010000	01111	00000
D	0001000	00011	00000
E	0000100	00001	10000
F	0000010	00000	01100
G	0000001	00000	00011

If we consider just the characters in black and red, then we'd have a homoplasy-free data matrix. The characters in green display character-conflict (they are incompatible with some of the characters in green and all of the read characters). Homework: What tree does ordinary (unweighted) least squares on the uncorrected total distance favor? What tree does parsimony favor? Perform a MP reconstruction of characters on the OLS tree. Which branches look suspicious according to parsimony? (you can use PAUP. I'll post a NEXUS file).

$$y = -1.5972\,x^5 + 23.167\,x^4 - 126.18\,x^3 + 319.17\,x^2 - 369.22\,x + 155.67$$

Why do we need models?



$$y = 0.6611\,x$$

Models

- Models help us intelligently **interpolate between our observations** for purposes of **making predictions**
- **Adding parameters** to a model generally increases its fit to the data
- **Underparameterized** models lead to poor fit to observed data points
- **Overparameterized** models lead to poor prediction of future observations
- Criteria for choosing models include likelihood ratio tests, AIC, BIC, Bayes Factors, etc.
  - all provide a way to choose a model that is neither underparameterized nor overparameterized

The Poisson distribution

Probability distribution on the number of events when:

1. events are assumed to be independent,
2. the *rate* of events some constant, μ, and
3. the process continues for some duration of time, t.

The expectation of the number of events is $\nu = \mu t$.

Note that ν can be any non-negative number, but the Poisson is a discrete distribution – it gives the probabilities of the number of events (and this number will always be a non-negative integer).

The Poisson distribution

$$\Pr(k \text{ events}) = \frac{\nu^k e^{-\nu}}{k!}$$

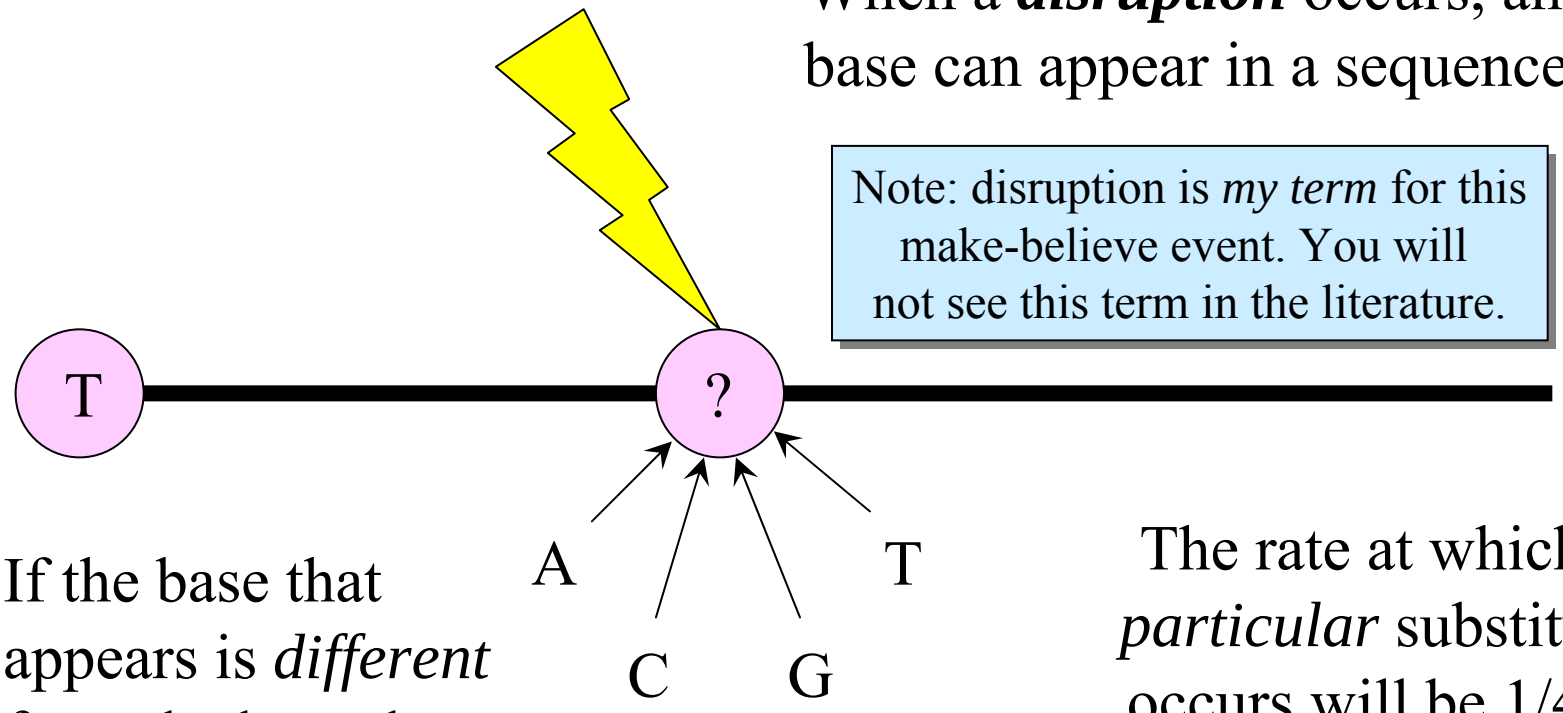
$$\Pr(0 \text{ events}) = \frac{\nu^0 e^{-\nu}}{0!} = e^{-\nu} = e^{-\mu t}$$

$$\Pr(\geq 1 \text{ events}) = 1 - e^{-\nu} = 1 - e^{-\mu t}$$

"Disruptions" vs. substitutions

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

Note: disruption 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 at which any *particular* substitution occurs will be 1/4 the disruption rate (assuming equal base frequencies)

Probability of T$\rightarrow$G over time t

If μ is the rate of disruptions, and a branch is t units of time long then: Let's use θ for the rate of any particular "disruption."

$$\mu_{TA} = \mu_{TC} = \mu_{TG} = \mu_{TT} = \theta$$
$$\mu = 4\theta$$

Furthermore, given that there is a disruption the chance of any particular change is $\frac{1}{4}$

Probability of $T \to G$ over time t

$$
\begin{aligned}
\Pr(0 \text{ disruptions} | t) &= e^{-\mu t} \\
\Pr(\text{at least 1 disruption} | t) &= 1 - e^{-\mu t} \\
\Pr(\text{last disruption leads to } G) &= 0.25 \\
\Pr(T \to G | t) &= 0.25 \left(1 - e^{-\mu t}\right) \\
&= 0.25 \left(1 - e^{-4\theta t}\right)
\end{aligned}
$$

JC69 model

- Bases are assumed to be equally frequent (all 0.25)
- Assumes rate of substitution (α) is the same for all possible substitutions
- Usually described as a 1-parameter model (the parameter being α)
- Remember, however, that each edge in a tree can have its own α, so there are really as many parameters in the model as there are edges in the tree!

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 *in* H. N. Munro (ed.), *Mammalian Protein Metabolism*. Academic Press, New York.

JC transition probabilities

$$
\begin{aligned}
\Pr(T \rightarrow A|t) &= 0.25\left(1 - e^{-4\theta t}\right) \\
\Pr(T \rightarrow C|t) &= 0.25\left(1 - e^{-4\theta t}\right) \\
\Pr(T \rightarrow G|t) &= 0.25\left(1 - e^{-4\theta t}\right) \\
\Pr(T \rightarrow T|t) &= 0.25\left(1 - e^{-4\theta t}\right)
\end{aligned}
$$

but this only adds up to:

$$\left(1 - e^{-4\theta t}\right)$$

instead of 1!

We left out the probability of no disruptions:$e^{-4\theta t}$

So:

$$
\begin{aligned}
\Pr(T \rightarrow A|t) &= 0.25\left(1 - e^{-4\theta t}\right) \\
\Pr(T \rightarrow C|t) &= 0.25\left(1 - e^{-4\theta t}\right) \\
\Pr(T \rightarrow G|t) &= 0.25\left(1 - e^{-4\theta t}\right) \\
\Pr(T \rightarrow T|t) &= e^{-4\theta t} + 0.25\left(1 - e^{-4\theta t}\right) \\
&= 0.25 + 0.75e^{-4\theta t}
\end{aligned}
$$

JC transition probabilities

$$\begin{aligned} \Pr(i \rightarrow j|t) &= 0.25\left(1 - e^{-4\theta t}\right) \\ \Pr(i \rightarrow i|t) &= 0.25 + 0.75e^{-4\theta t} \end{aligned}$$

When $t = 0$, then $e^{-4\theta t} = 1$, and:

$$\begin{aligned} \Pr(i \rightarrow j|t) &= 0 \\ \Pr(i \rightarrow i|t) &= 1 \end{aligned}$$

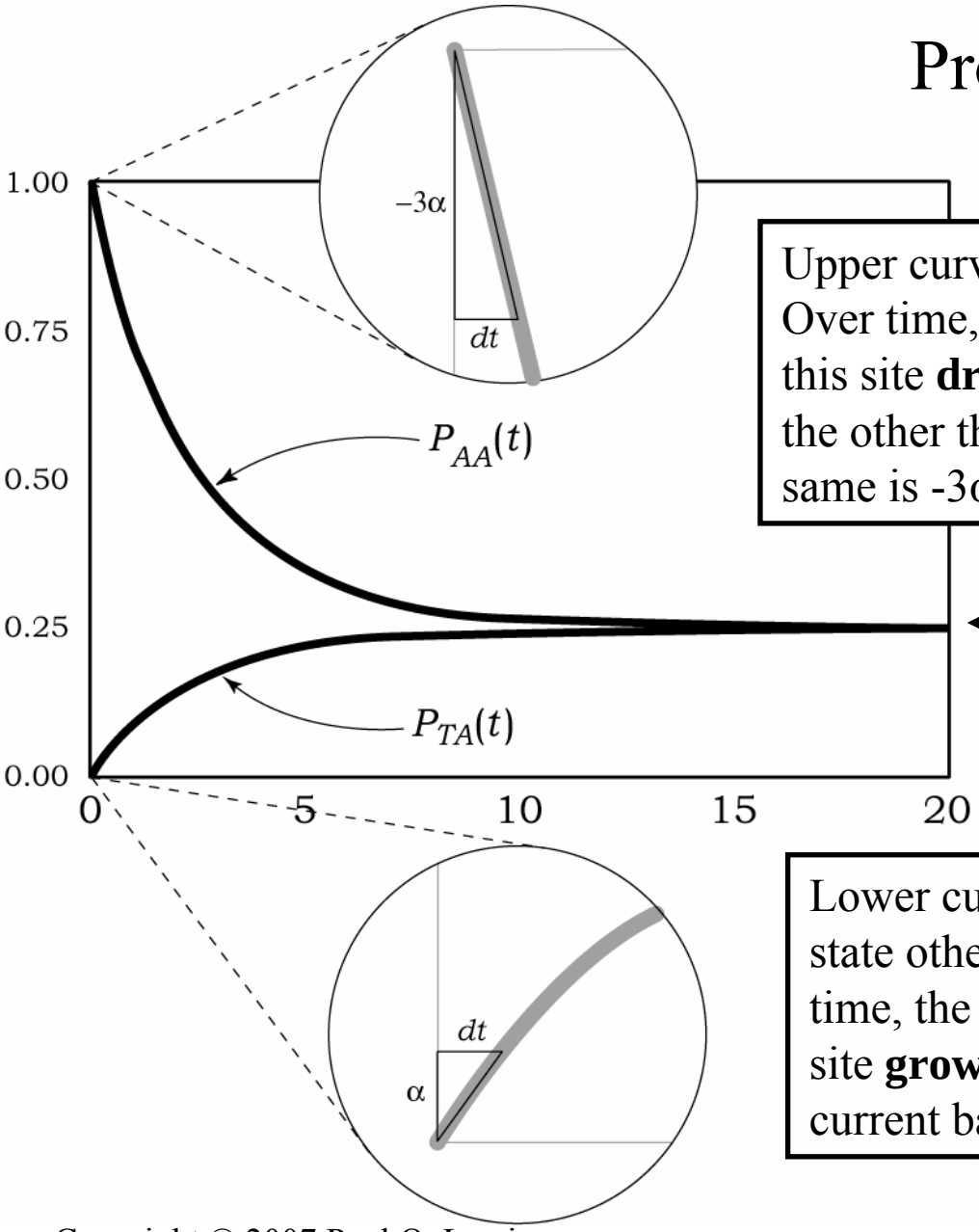
JC transition probabilities

$$\begin{aligned} \Pr(i \rightarrow j|t) &= 0.25\left(1 - e^{-4\theta t}\right) \\ \Pr(i \rightarrow i|t) &= 0.25 + 0.75e^{-4\theta t} \end{aligned}$$

When $t = \infty$, then $e^{-4\theta t} = 0$, and:

$$\begin{aligned} \Pr(i \rightarrow j|t) &= 0.25 \\ \Pr(i \rightarrow i|t) &= 0.25 \end{aligned}$$

Probability of "A present" as a function of time

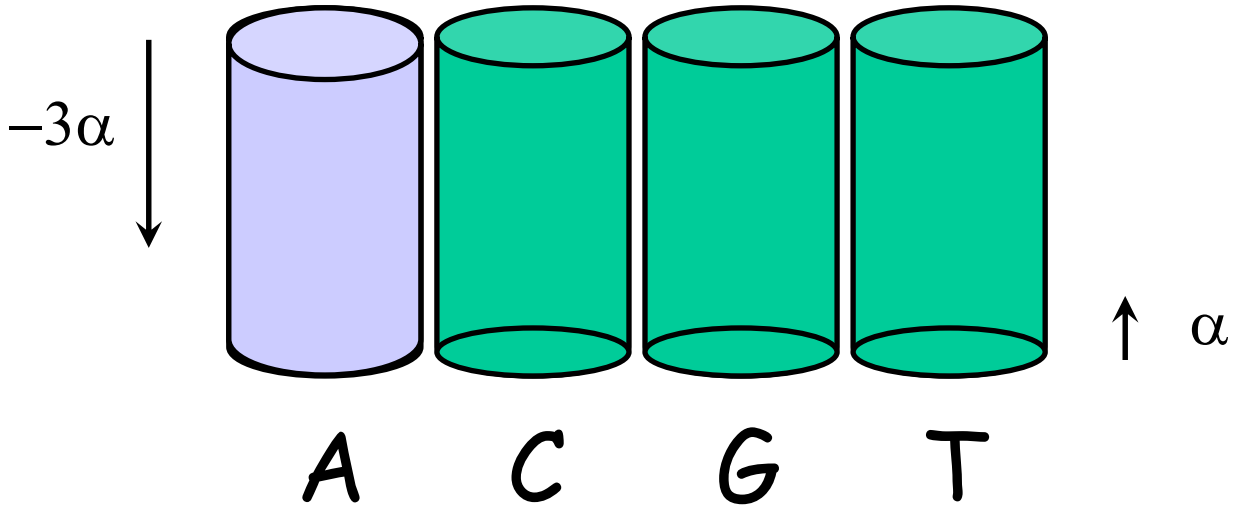


Upper curve assumes we started with A at time 0. Over time, the probability of still seeing an A at this site **drops** because rate of changing to one of the other three bases is 3α (so rate of staying the same is -3α).

The equilibrium relative frequency of A is 0.25

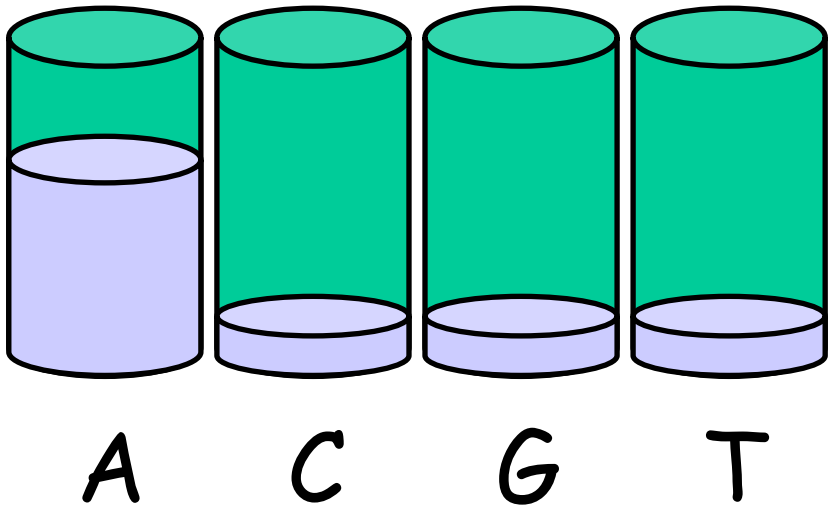
Lower curve assumes we started with some state other than A (T is used here). Over time, the probability of seeing an A at this site **grows** because the rate at which the current base will change into an A is α.

Water analogy (time 0)



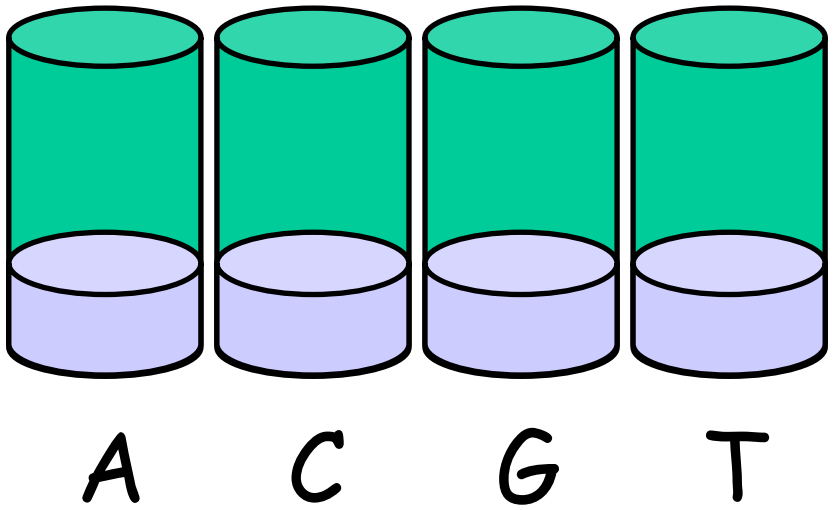
- Start with container A completely full and others empty
- Imagine that all containers are connected by tubes that allow same rate of flow between any two
- Initially, A will be losing water at 3 times the rate that C (or G or T) gains water

Water analogy (after some time)



A's level is not dropping as fast now because it is now also *receiving* water from C, G and T

Water analogy (after a very long time)



Eventually, all containers are one fourth full and there is zero *net* volume change – **stationarity** (equilibrium) has been achieved

(Thanks to Kent Holsinger for this analogy)

JC instantaneous rate matrix - the Q matrix for JC

The 1 parameter is α (sometimes parameterized in terms of μ). This is the rate of replacements ("disruptions" that change the state):

		To State			
		A	C	G	T
From State	A	-3α	α	α	α
	C	α	-3α	α	α
	G	α	α	-3α	α
	T	α	α	α	-3α

Change probabilities

We can calculate a transition probability matrix as a function of time by:

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

The important thing to note is the rates ($\mathbf{Q}$ matrix) is multiplied by the time.

We can't separate rates and times since we always see the effect of their product.

Is a medium level of character divergence:

1. medium rate of change and medium amount of time,
2. high rate, but short time period,
3. low rate, but a long time period?

JC instantaneous rate matrix again

What if you do not know the length of time for a branch in the tree? We estimate branch lengths in terms of character divergence – the product of rate and time. What is important is that we know the relative rates of different types of substitutions, so JC can be expressed:

		To State			
		A	C	G	T
From State	A	-3	1	1	1
	C	1	-3	1	1
	G	1	1	-3	1
	T	1	1	1	-3

JC instantaneous rate matrix yet again

We estimate branch lengths in terms of expected number of changes *per site*. To do this we standardize the total rate of divergence in the Q matrix and estimate $\nu = \mu t = 3\alpha t$ for each branch.

		To State			
		A	C	G	T
From State	A	-1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
	C	$\frac{1}{3}$	-1	$\frac{1}{3}$	$\frac{1}{3}$
	G	$\frac{1}{3}$	$\frac{1}{3}$	-1	$\frac{1}{3}$
	T	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	-1

Kimura (1980) model or "the K80 model"

Transitions and transversions occur at different rates:

		To State			
		A	C	G	T
From State	A	$-2\beta - \alpha$	β	α	β
	C	β	$-2\beta - \alpha$	β	α
	G	α	β	$-2\beta - \alpha$	β
	T	β	α	β	$-2\beta - \alpha$

Kimura (1980) model or "the K80 model". Reparameterized.

Once again, we care only about the relative rates, so we can choose one rate to be frame of reference. This turns the 2 parameter model into a 1 parameter form:

		To State			
		A	C	G	T
From State	A	$-(2+\kappa)\beta-$	β	$\kappa\beta$	β
	C	β	$-(2+\kappa)$	β	$\kappa\beta$
	G	$\kappa\beta$	β	$-(2+\kappa)$	β
	T	β	$\kappa\beta$	β	$-(2+\kappa)$

Kimura (1980) model or "the K80 model". Reparameterized again.

		To State			
		A	C	G	T
From State	A	$-2-\kappa$	1	κ	1
	C	1	$-2-\kappa$	1	κ
	G	κ	1	$-2-\kappa$	1
	T	1	κ	1	$-2-\kappa$

Kappa is the transitiiton/transversion rate ratio:

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

(if $\kappa = 1$ then we are back to JC).

What is the instantaneous probability of an particular transversion?

$$\begin{aligned} \Pr(A \to C) &= \Pr(A) \Pr(\text{change to } C) \\ &= \frac{1}{4} (\beta dt) \end{aligned}$$

What is the instantaneous probability of an particular transition?

$$\begin{aligned} \Pr(A \to G) &= \Pr(A) \Pr(\text{change to } G) \\ &= \frac{1}{4} (\kappa \beta dt) \end{aligned}$$

There are four types of transitions:

$$A \to G, G \to A, C \to T, T \to C$$

and eight types of transversions:

$$A \to C, A \to T, G \to C, G \to T, C \to A, C \to G, T \to A, T \to G$$

$$\text{Ti/Tv ratio} = \frac{\Pr(\text{any transition})}{\Pr(\text{any transversion})} = \frac{4\left(\frac{1}{4}\left(\kappa\beta dt\right)\right)}{8\left(\frac{1}{4}\left(\beta dt\right)\right)} = \frac{\kappa}{2}$$

For K2P instantaneous transition/transversion ratio is one-half the instantaneous transition/transversion **rate ratio**

Felsenstein 1981 model or "F81 model"

		To State			
		A	C	G	T
From State	A	$-$	π_C	π_G	π_T
	C	π_A	$-$	π_G	π_T
	G	π_A	π_C	$-$	π_T
	T	π_A	π_C	π_G	$-$

HKY 1985 model

		To State			
		A	C	G	T
From State	A	$-$	π_C	$\kappa\pi_G$	π_T
	C	π_A	$-$	π_G	$\kappa\pi_T$
	G	$\kappa\pi_A$	π_C	$-$	π_T
	T	π_A	$\kappa\pi_C$	π_G	$-$

F84* vs. HKY85

F84 model:

μ rate of process generating *all types of substitutions*

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

Becomes F81 model if $k = 0$

HKY85 model:

β rate of process generating *only transversions*

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

Becomes F81 model if $\kappa = 1$

*First used in PHYLIP in 1984, first published byKishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of Molecular Evolution 29: 170-179.

General Time Reversible – GTR model

		To State			
		A	C	G	T
From State	A	$-$	$a\pi_C$	$b\pi_G$	$c\pi_T$
	C	$a\pi_A$	$-$	$d\pi_G$	$e\pi_T$
	G	$b\pi_A$	$d\pi_C$	$-$	$f\pi_T$
	T	$c\pi_A$	$e\pi_C$	$f\pi_G$	$-$

In PAUP, $f = 1$ indicating that $G \rightarrow T$ is the reference rate

References

Bulmer, M. (1991). Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution*, 8(6):868–883.

Desper, R. and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 9(5):687–705.

Desper, R. and Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution*.

Gascuel, O. and Steel, M. (2006). Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11):1997–2000.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120.

Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 2000(51):41–47.