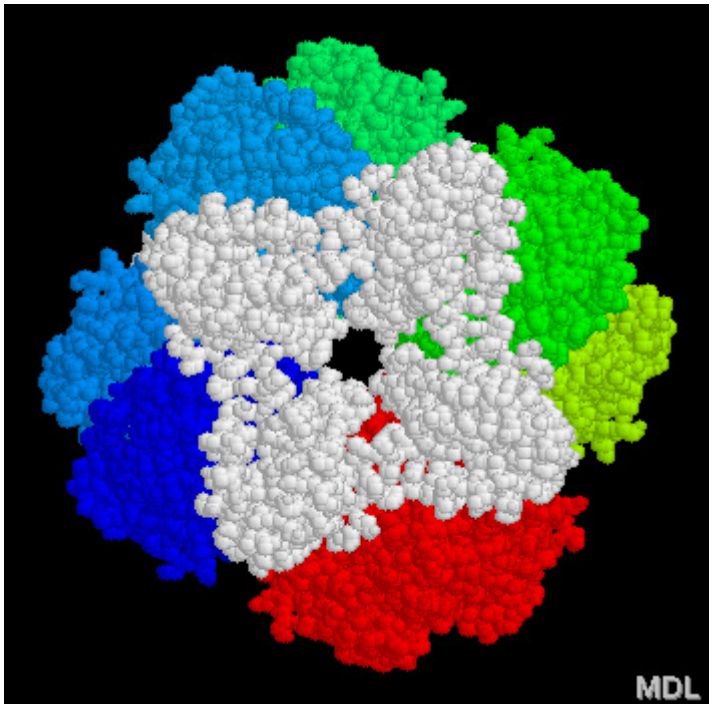
Some of these slides have been borrowed from Dr. Paul Lewis, Dr. Joe Felsenstein. Thanks!

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

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

RuBisCO enzyme



8 small subunits (white)
8 large subunits (colored)
Responsible for fixing CO_2

Green Plant rbc*L*

First 88 amino acids, translation is for *Zea mays*

```
M--S--P--Q--T--E--T--K--A--S--V--G--F--K--A--G--V--K--D--Y--K--L--T--Y--Y--T--P--E--Y--E--T--K--D--T--D--I--L--A--A--F--R--V--T--P--
Chara           (green alga; land plant lineage)   AAAGATTACAGATTAACTTACTATACTCCTGAGTATAAAACTAAAGATACTGACATTTTAGCTGCATTTCGTGTAACTCCA
Chlorella       (green alga)                       .....C...C.T...................T..CC..C.A.....C.....T...C.T..A..G..C...A.G.....T
Volvox          (green alga)                       ........TC.T.....A.....C..A.....C...GT.GTA.....C........C.....A.........A.G......
Conocephalum    (liverwort)                        ........TC.........T........G..T...G............G..T........A......A.AA.G.....T
Bazzania        (moss)                             ........T........C..T.....G.....A...G.G..C.....G..A..T.....G..A.........A.G.....C
Anthoceros      (hornwort)                         ........T........CC.T.....C.....T..CG.G..C..G........T.....G..A..G.C.T.AA.G.....T
Osmunda         (fern)                             ........TC....G...C..........C..T...G.G..C..G........T.....G..A.....C..AA.G.....C
Lycopodium      (club "moss")                      .GG..............C.T..C........T.....G..C.....A..C..T...C.G..A........AA.G.....T
Ginkgo          (gymnosperm; Ginkgo biloba)        ..............G.....T..........A...C....C...........T..C..G..A.....C..A........T
Picea           (gymnosperm; spruce)               ...................T..........A...C.G..C........G..T.....G..A.....C..A........T
Iris            (flowering plant)                  ..............G.....T..........T..CG....C...........T..C..G..A.....C..A........T
Asplenium       (fern; spleenwort)                 ........TC..C.G.....T..C..C..C..A..C..G..C........C..T..C..G..A..T..C..GA.G..C...
Nicotiana       (flowering plant; tobacco)         .....G....A...G.....T..............CC....C..G........T..A..G..A.....C..A........T
```

```
Q--L--G--V--P--P--E--E--A--G--A--A--V--A--A--E--S--S--T--G--T--W--T--T--V--W--T--D--G--L--T--S--L--D--R--Y--K--G--R--C--Y--H--I--E--
CAACCTGGCGTTCCACCTGAAGAAGCAGGGGCTGCAGTAGCTGCAGAATCTTCTACTGGTACATGGACTACTGTTTGGACTGACGGATTAACTAGTTTGGACCGATACAAAGGAAGATGCTACGATATTGAA
.....A..T........A........G..T..G........A........A..A........T.....G.....A........T..T...........A.....T........TC.T..T..T..C..C..G
.....A..T..............TGT..T.....T..T.....T.....A..A..A.....T.....A.....A........T..T.....A...C.T.....T........TC.T..T..T..C..C..G
..G.....G..A...G.A...........A..A.....T.....T.........................A...........T..TC.T....ACC.T..T..T..T.....TC.......T.G......C
.....G..A..A.................A..G...........T........A..C.....G.....C..G........C..T..GC.T..A...C.C..T..T........TC.......T..C..C...
T....A..G..G.................A..C...........T........A.........................C..T...C.T..C..CC.T.....T........TC..........C......
.....C..A..A..GG....G.....T..A..............G...........A.....G.....C.....A.....G..T...C.T..C...C.T..T..T..T..G..TC................
....T...A..A.....C..G.....G..A..C...........T........C.........................C..T...C.T..C...C.C..T..C........TC.G.....T..A......
........A..G........G.....G..A..............C........C..............C...........C..T...C.T..C...C.T..T..T.....G...........T..C..C..G
.....A..G..G..C..G.....G..A..A...........T........C..C...........C...........C..T...C.T......C.T..T..T.....G..GC.......T..C..C..G
.....C..A.....TG..........G.....C..G........C.....................A..A..G........T...C.T..C...C.T..T..T.........C........C.C..C..G
.....C..A..A...G..........C..A.................G..C.....A...........C.....G.....A.....G..G..C..CC.T.....T.....G..CC.............C..G
........A.....................C..G........C.....................A.....A.....C..T...C.T..C..CC.T..T..T........GC........CGC..C..G
```

All four bases are observed at some sites...

...while at other sites, only one base is observed

Question: Why is rate heterogeneity ubiquituous?

Answer: Differences in mutational rates and (mainly) **selective constraint**

- Many sites are under purifying (stabilizing) selection:
  - Any mutation results in a different amino acid, AND
  - A amino acid replacement at the site results in dramatically worse functioning of the protein.
  - These sites will show *low* rates of evolution on a tree.
- Other sites are less constrained.
  - A mutation results in the same amino acid, OR
  - Many amino acids will work equally well at that position in the protein.
  - These sites will show *high* rates of evolution on a tree.

Rate heterogeneity in protein-coding genes: terms

- **Synonymous** mutations result in the same amino acid.
- **Non-synonymous** mutations result in the different amino acid.
- **Conservative** changes are non-synonymous changes that result in a chemically similar amino acid.
- **Neutral** mutations result in a new genotype that has the same fitness as the genotypes currently fixed in the population.

Rate heterogeneity in protein-coding genes: generalities

- Synonymous changes are often neutral (or close to neutral),
- Third base positions and untranslated regions (introns and other non-coding regions) tend to have high rates because changes to these sites lead to synonymous changes.
- Transitions tend to lead to more synonymous or conservative changes.
- Amino acid residues that are embedded, involved in salt bonding, or part of the active site tend to be more constrained.
- Loops of amino acid residues on the outside of proteins often tolerate a wide range of substitutions (or even indels).

	2nd Base							
	U		C		A		G	
U	UUU	**F**	UCU	**S**	UAU	**Y**	UGU	**C**
	UUC		UCC		UAC		UGC	
	UUA	**L**	UCA		UAA	*	UGA	*
	UUG		UCG		UAG		UGG	**W**
C	CUU		CCU		CAU	**H**	CGU	
	CUC	**L**	CCC	**P**	CAC		CGC	**R**
	CUA		CCA		CAA	**Q**	CGA	
	CUG		CCG		CAG		CGG	
A	AUU		ACU		AAU	**N**	AGU	**S**
	AUC	**I**	ACC	**T**	AAC		AGC	
	AUA		ACA		AAA	**K**	AGA	**R**
	AUG	**M**	ACG		AAG		AGG	
G	GUU	**V**	GCU	**A**	GAU	**D**	GGU	**G**
	GUC		GCC		GAC		GGC	
	GUA	**L**	GCA		GAA	**E**	GGA	
	GUG		GCG		GAG		GGG	

Rate heterogeneity in RNA coding genes

- Stem regions
  - formed when RNA strand forms double-helix with itself
  - strongly conserved in general
  - evidence for compensatory substitutions
- Loop regions
  - some strongly conserved
  - some entire loops are found in only particular lineages

Accommodating rate heterogeneity in substitution models

- Site-specific rates approach
  - e.g. let 1st, 2nd and 3rd position sites each have their own relative substitution rate
- Proportion of invariable sites approach
  - assume that some proportion p_{invar} of sites have rate 0, while a proportion $1-p_{invar}$ have a rate > 0
- Discrete gamma distributed relative rates approach
  - assume that each site is evolving at one of n_{cat} relative rates, where the relative rates are determined using a gamma distribution having mean 1 and shape α
- Codon models (protein-coding genes only)
  - uses genetic code to determine appropriate relative rates
- Secondary structure models (RNA-coding genes only)
  - uses separate model for loops vs. stems, stem model takes account of compensatory substitutions

Site-specific rates

- You decide there are 3 classes of sites:
  - 1st positions evolve at relative rate r_1
  - 2nd positions evolve at relative rate r_2
  - 3rd positions evolve at relative rate r_3
- r_1, r_2 and r_3 are *relative* rates, not *actual* rates:
  - their average is 1.0: if each category has the same number of sites, $(r_1 + r_2 + r_3)/3 = 1.0$
  - the actual rates are $r_1 \alpha$ (for 1st positions), $r_2 \alpha$ (for 2nd positions) and $r_3 \alpha$ (for 3rd positions)
  - note that the average substitution rate over all sites is α $(r_1 \alpha + r_2 \alpha + r_3 \alpha)/3 = \alpha\ (1.0) = \alpha$
- Assuming k rate classes adds k-1 parameters to the model

Transition probabilities under the JC69 model

with no rate heterogeneity:

$$\begin{aligned} \Pr(i \to i|\nu) &= \frac{1}{4} + \frac{3}{4}\, e^{\frac{-4\nu}{3}} \\ \Pr(i \to j|\nu) &= \frac{1}{4} - \frac{1}{4}\, e^{\frac{-4\nu}{3}} \end{aligned}$$

Transition probabilities under the JC69 model

First base positions under a *site-specific rates* model:

$$\begin{aligned}
\Pr(i \to i|\nu) &= \frac{1}{4} + \frac{3}{4}\, e^{\frac{-4r_1\nu}{3}} \\
\Pr(i \to j|\nu) &= \frac{1}{4} - \frac{1}{4}\, e^{\frac{-4r_1\nu}{3}}
\end{aligned}$$

Site-specific rates in PAUP*

First, define a character partition that puts each site into one of several mutually exclusive categories (the category names are arbitrary):

```
charpartition codons = one:1-.\3, two:2-.\3, three:3-.\3;
```

Then tell PAUP* that you want site specific rates and provide the partition you defined previously:

```
lset rates=sitespec siterates=partition:codons;
```

Pinvar approach

- Unlike the site-specific rates approach, this approach does not require you to assign sites to rate categories
- Assumes there are only two classes of sites:
  - invariable sites (evolve at relative rate 0)
  - variable sites (evolves at relative rate *r*)
- Remarks:
  - mean of relative rates = $(p_{invar})(0) + (1-p_{invar})(r) = 1$
  - this means that $r = 1/(1-p_{invar})$
  - if all sites are variable, $p_{invar} = 0$ and $r = 1$

- **Constant site** – a site in which all of the taxa display the same character state.
- **Invariable site** – a site in which only one character state is allowed. A site that cannot change state.

All invariable sites are constant, but not all constant sites have to be invariable.

$$
\begin{aligned}
\Pr(i \to i|\ \text{invariable}) &= \frac{1}{4} + \frac{3}{4}\, e^{\frac{-40\nu}{3}} \\
&= \frac{1}{4} + \frac{3}{4}\, e^{0} \\
&= 1 \\
\Pr(i \to j|\ \text{invariable}) &= \frac{1}{4} - \frac{1}{4}\, e^{\frac{-40\nu}{3}} \\
&= 0
\end{aligned}
$$

A site's likelihood under the JC+ I model

x_i is the data pattern for site i. General form:

$$\Pr(x_i|\text{JC+I}) = p_{\text{inv}} \Pr(x_i|\text{ inv}) + (1 - p_{\text{inv}}) \Pr\left(x_i|\text{JC}, \frac{\boldsymbol{\nu}}{1 - p_{\text{inv}}}\right)$$

If x_i is a variable site:

$$\Pr(x_i|\text{JC+I}) = (1 - p_{\text{inv}}) \Pr\left(x_i|\text{JC}, \frac{\boldsymbol{\nu}}{1 - p_{\text{inv}}}\right)$$

If x_i is a constant site:

$$\Pr(x_i|\text{JC+I}) = p_{\text{inv}} \Pr(x_i|\text{ inv}) + (1 - p_{\text{inv}}) \Pr\left(x_i|\text{JC}, \frac{\boldsymbol{\nu}}{1 - p_{\text{inv}}}\right)$$

Why $\frac{\nu}{1-p_{\mathsf{inv}}}$?

We want the mean rate of change to be 1.0 over all sites (so we can interpret the branch lengths in terms of the expected # of changes per site).

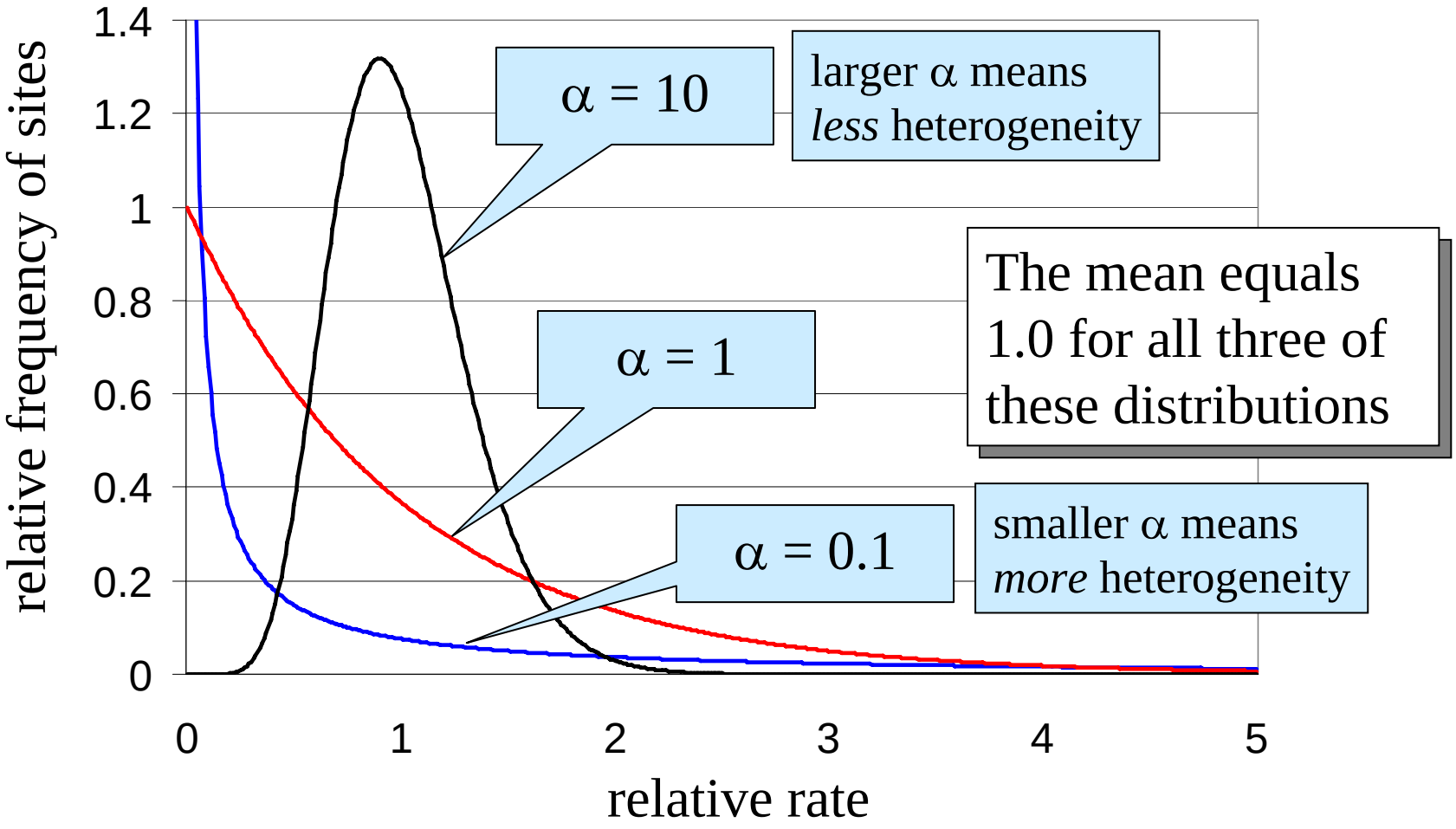
If r is the rate of change for the variable sites then:

$$\begin{aligned} 1 &= 0p_{\mathsf{inv}} + r\left(1 - p_{\mathsf{inv}}\right) \\ &= r\left(1 - p_{\mathsf{inv}}\right) \\ r &= \frac{1}{1 - p_{\mathsf{inv}}} \end{aligned}$$

Variable (but unknown) rates

- We expect more "shades of grey" rather than the on-or-off view of the pInvar model.
- *a priori* we do not know which sites are fast and which are slow
- We may be able to characterize the *distribution* of rates across sites – high variance or low variance.

Gamma distributions



Gamma distribution

$$
\begin{aligned}
f(r) &= \frac{r^{\alpha-1}\beta^{\alpha}e^{-\beta r}}{\Gamma(\alpha)} \\
\text{mean} &= \alpha/\beta \\
\text{mean (in phylogenetics)} &= 1 \\
\text{(in phylogenetics) } \beta &= \alpha \\
\text{variance} &= \alpha/\beta^2 \\
\text{variance (in phylogenetics)} &= 1/\alpha
\end{aligned}
$$

Using Gamma-distributed rates across sites

- We usually use a discretized version of the gamma with 4-8 categories (the computation time increases linearly with the number of categories).

$$\Pr(x_i|JC+G) = \sum_{j}^{\text{ncat}} \Pr(x_i|JC, r_j\boldsymbol{\nu}) \Pr(r_j)$$

where:

$$\sum_{j}^{\text{ncat}} r_j \Pr(r_j) = 1$$

Discrete gamma (continued)

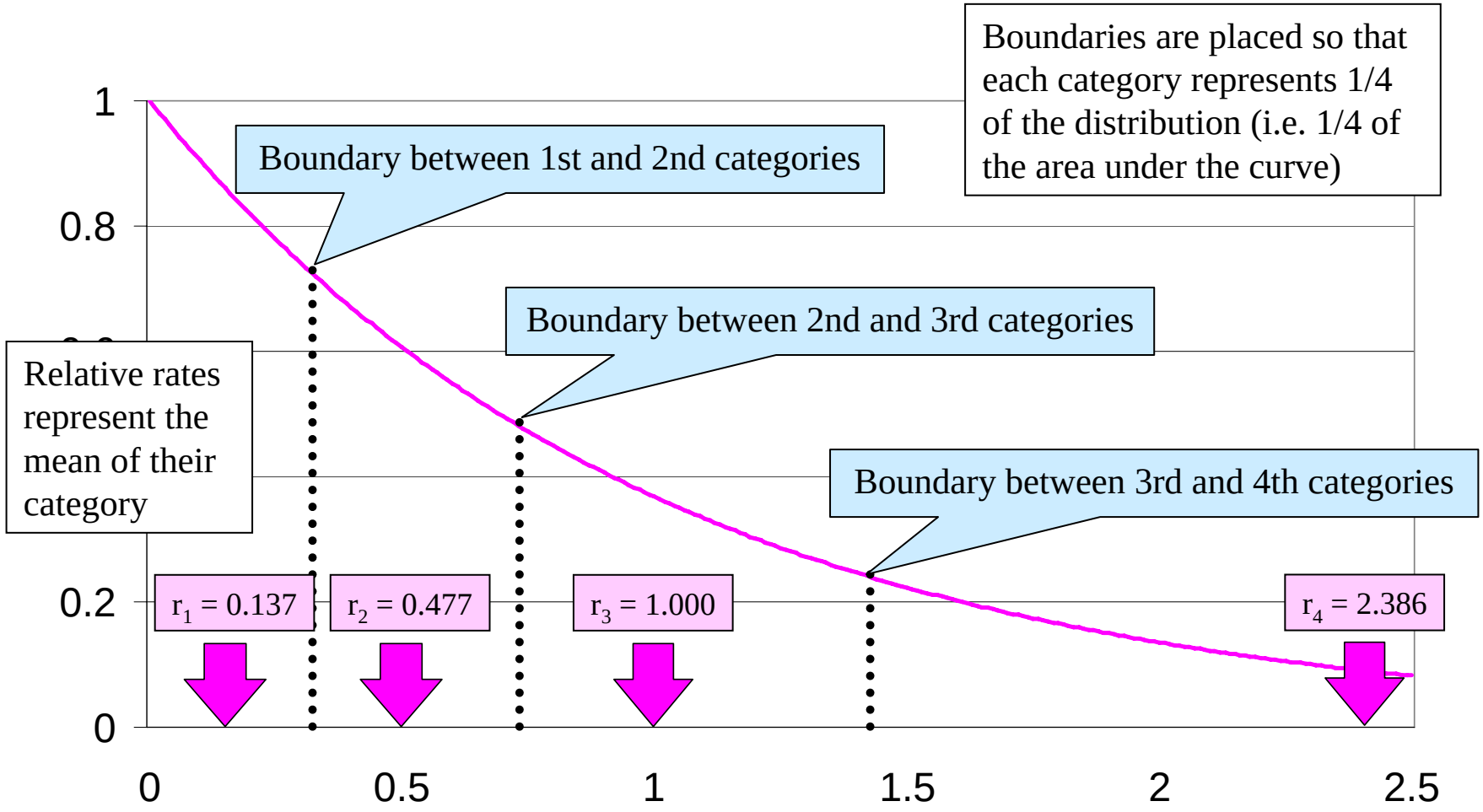
We "break up" the continuous gamma into intervals each of which has an equal probability, and use the mean rate within each interval as the representative rate for that rate category:

$$\Pr(r_j) = \frac{1}{\text{ncat}}$$

So:

$$\Pr(x_i|JC + G) = \frac{1}{\text{ncat}} \sum_{j}^{\text{ncat}} \Pr(x_i|JC, r_j\boldsymbol{\nu})$$

Relative rates in 4-category case



Discrete gamma rate heterogeneity in PAUP*

To use gamma distributed rates with 4 categories:

```
lset rates=gamma ncat=4;
```

To estimate the shape parameter:

```
lset shape=estimate;
```

To combine pinvar with gamma:

```
lset rates=gamma shape=0.2 pinvar=0.4;
```

Note: `estimate`, `previous`, or a specific value can be specified for both `shape` and `pinvar`

Rate homogeneity in PAUP*

Just tell PAUP* that you want all rates to be equal and that you want all sites to be allowed to vary:

```
lset rates=equal pinvar=0;
```

Note: these are the default settings, but it is useful to know how to go back to rate homogeneity after you have experimented with rate heterogeneity!

Likelihood ratio test

- Always compares an **unconstrained** to a **constrained** model
- Constrained model must be **nested within** the unconstrained model
- Parameter(s) take on their maximum likelihood estimates (**MLEs**) in the unconstrained model
- Parameters(s) set to **some other value** of interest in the constrained model
- *Unconstrained model must be able to attain a higher maximum likelihood than the constrained model*

Likelihood Ratio Test

$$LRT = 2 \log \frac{L(\hat{\theta})}{L(\theta)}$$

$\hat{\theta}$ is the MLE

θ is some other value

Coin-flipping example:

Data:	6 heads out of 10 flips
Constrained model:	fair coin ($\theta = 0.5$)
Unconstrained model:	biased coin ($\theta = \hat{\theta}$)

$$\Pr(y = 6|\theta = 0.6) = \left(\frac{10!}{6!\ 4!}\right)(0.6)^6(0.4)^4 = 0.25082$$

Example of likelihood calculation for case of $\theta = 0.6$

Likelihood Ratio Test

Coin-flipping example:

Data:	6 heads out of 10 flips
Constrained model:	fair coin ($\theta = 0.5$)
Unconstrained model:	biased coin ($\theta = \hat{\theta}$)

$$LRT = 2 \log \frac{L(0.6)}{L(0.5)} = 2 \log \frac{0.251}{0.205} = 0.4$$

$$\chi^2 = \frac{(6-5)^2}{5} + \frac{(4-5)^2}{5} = 0.4$$

Not significant:
P = 0.527
This means that the simpler, constrained model cannot be rejected

LRT approximates a chi-square random variable with d.f. equal to the difference in the number of free parameters between the two models

Examples of unconstrained vs. constrained model comparisons

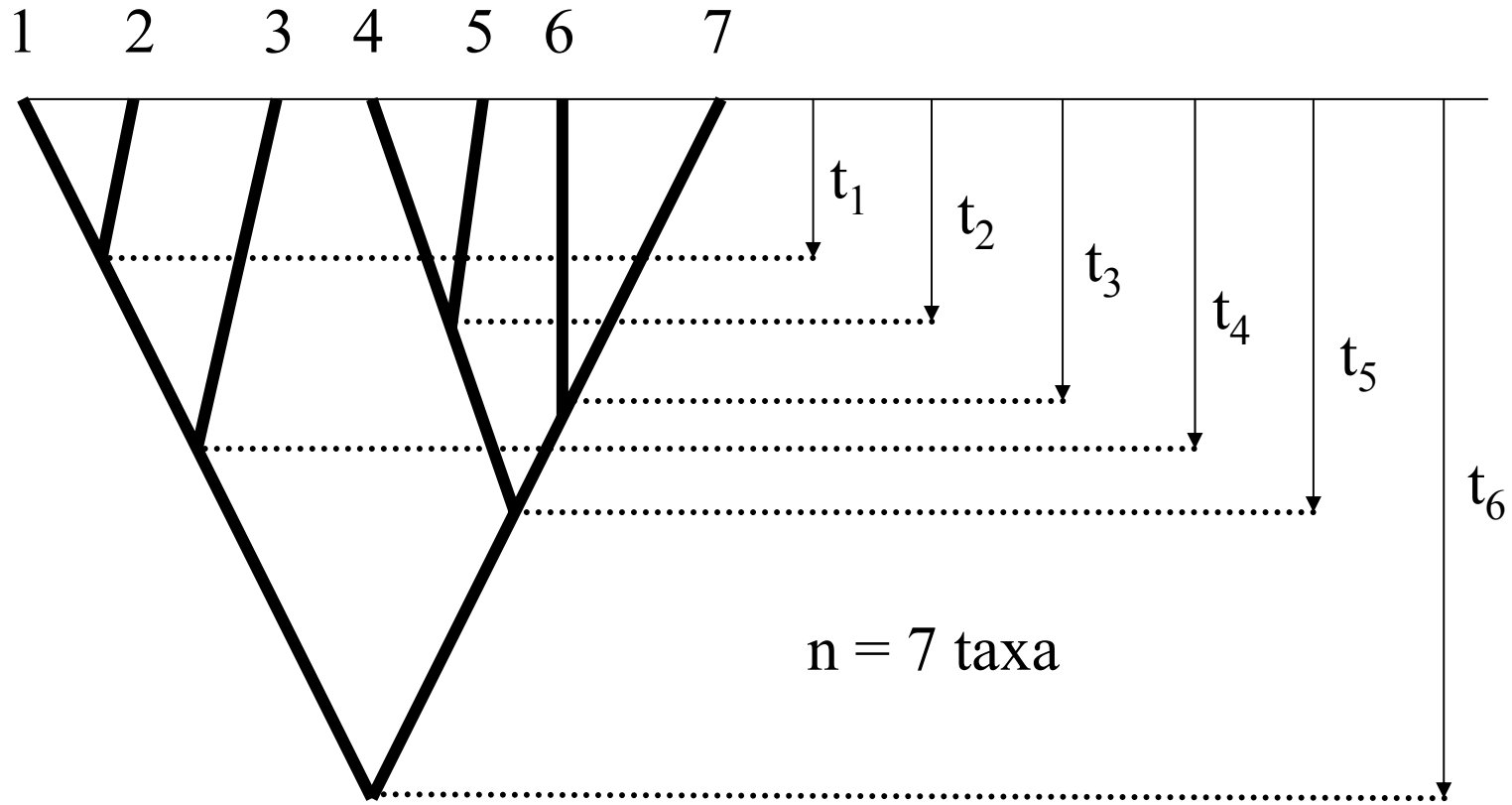
1. GTR+G (shape=MLE) vs. GTR (shape=∞)
2. K80 (κ=MLE) vs. JC (κ=1.0)
3. HKY+I+G (p_{inv}=MLE) vs. HKY+G (p_{inv}=0)
4. HKY+I+G (p_{inv}=MLE, shape=MLE) vs. HKY (p_{inv}=0, shape= ∞)

Note: cases in which the constrained model involves setting a parameter to the edge of its valid range (e.g. cases 1, 2 and 4 above) require special consideration (see Ota et al. 2000)

Ota, R., P. J. Waddell, M. Hasegawa, H. Shimodaira, and H. Kishino. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Molecular Biology and Evolution 17:798-803.

Testing the molecular clock

Unconstrained model: need to estimate 2n-3 = 11 branch lengths
Constrained model: need to estimate n-1 = 6 divergence times



Likelihood ratio test thus has (2n-3) - (n-1) = n-2 d.f.

Akaike Information Criterion

- AIC = -2 max(ln$L$) + 2K
- K is number of free model parameters
- Measures relative distance to true model
- Model with **smallest** AIC wins
- Advantage over LRT: **non-nested models**

Example: 6 heads/10 flips revisited
Unconstrained model: $\theta = 0.6$, AIC = -2(-1.383) + 2(1) = 4.766
Constrained model: $\theta = 0.5$, AIC = -2(-1.584) + 2(0) = 3.168 (best)

Bayesian Information Criterion

- BIC = $-2 \max(\ln L) + K \log(n)$
- K is number of free model parameters
- n is the sample size
- Model with **smallest** BIC wins
- Advantage over LRT: **non-nested models**
- Considered superior to both AIC and LRT

Example: 6 heads/10 flips one more time. Note: $\log(10) \approx 2.3$
Unconstrained model: $\theta = 0.6$, BIC = -2(-1.383) + (2.3)(1) = 5.066
Constrained model: $\theta = 0.5$, BIC = -2(-1.584) + 0 = 3.168 (best)

Likelihood ratio test favors more complex models

- Assume the simpler, constrained model is the true model
- If the LRT was statistically consistent, it would choose the true model with certainty as $n \rightarrow \infty$
- But the simpler model will be rejected 5% of the time, regardless of sample size
- Thus, LRT biased toward choosing the more complex, unconstrained model

References