Maximum Likelihood in Phylogenetics

Rachel Bevan

PhD Student McGill University

DIMACS 2006

Outline

- Maximum Likelihood
  - General concepts
  - Hypothesis Testing
  - Computing the probability of a site along a tree
- DNA and Protein Models
- Site rate heterogeneity
- Recent advances
  - Accounting for gene rate heterogeneity

Likelihood

- Probability of data given a model of interest
- For probability density function *f*, data points $x_1, \ldots, x_n$ and parameter vector θ

$$
\begin{aligned}
L(\theta|x_1, \ldots, x_n) &= L(\theta|x) \\
&= f(x|\theta) \\
&= \prod_{i=1}^{n} f(x_i|\theta)
\end{aligned}
$$

Example

- Series of coin flips (1 is heads, 0 is tails)

$$x_1, ..., x_7 = \{1, 0, 0, 0, 0, 1, 0\}$$

- Bernoulli random variable pmf IID

$$P(X = x|p) = p^x(1-p)^{(1-x)}$$

$$x = 0, 1; 0 \leq p \leq 1$$

Example cont…

- Normally for a fair coin $p=0.5$, thus the probability of observing 5 tails and 2 heads is

$$\prod_{i=1}^{7} P(X_i = x_i | p = 0.5)$$

$$= 0.5^{(5)} 0.5^{(2)}$$

Example cont…

- What if we don't know whether or not the coin is fair?
- We want to find the best estimate of p given our observed data
  - Recall: each observed coin toss is a bernoulli trail observe heads or tails $x_1, ..., x_7 = \{1, 0, 0, 0, 0, 1, 0\}$

$$L(p|x_1, ..., x_7) = \prod_{i=0}^{7} (p)^{x_i} (1-p)^{(1-x_i)}$$
$$= p^2(1-p)^5$$

  - Thus find the maximum value of p given the data and our model of interest

Hypothesis testing

- What if we don't know that the data are IID bernoulli?
- Compare the probability under a NULL hypothesis (the model that we think describes the data) to the alternative hypothesis
- LRT: if the models are *nested*

$$\lambda(x) = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)}$$

For $\hat{\theta}_0$ the MLE of data x under the NULL hypothesis and $\hat{\theta}$ the MLE of data x under the alternative hypothesis

Significance level of LRT

- Rejection region:

$$\{x : \lambda(x) \leq c\}; 0 \leq c \leq 1$$

- If the value of the LRT is within this region then we reject the NULL hypothesis in favour of the alternative
- Test level: choose c so that the NULL hypothesis is rejected incorrectly in favour of the alternative at a level α

$$\sup_{\theta \in \Theta_0} P_\theta(\lambda(X) \leq c) \leq \alpha$$

Asymptotic Distribution of LRT

- Under certain regularity conditions

$$-2log\lambda(X)$$

converges to a chi-squared distribution, with degrees of freedom the difference in number of parameters specified by NULL and alternative hypotheses

Non-nested hypotheses

- Akaike Information Criterion
  - For model *i* with likelihood L_i and parameters p_i

$$AIC_i = -2log(L_i) + 2p_i$$

- Bayesian Information Criterion
  - For model *i* with likelihood L_i, parameters p_i and data size n

$$BIC_i = -2log(L_i) + p_i ln(n)$$

Model Fit

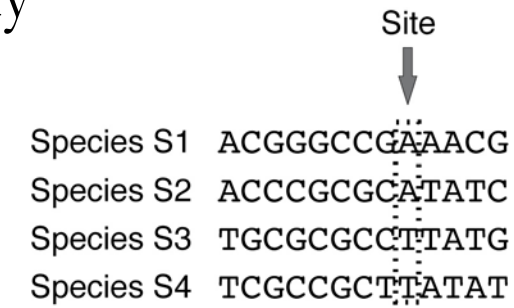
- Under AIC/BIC models with more parameters are 'punished'
  - This is due to the fact that adding more parameters to a model in general leads to a better likelihood
  - finding a better fit to the current data of interest, but not any data set
  - Over-fitting the model to the data

Likelihood on Trees

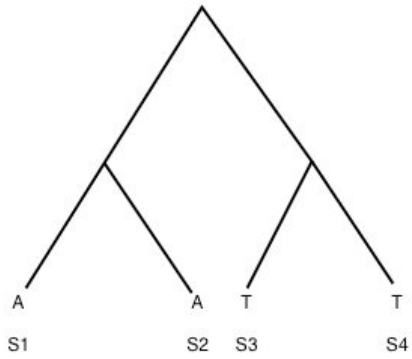
- Why do we care
  - Allows us to assess the probability of observing data under a particular model of interest
  - It is now possible to compare different models to determine which provides the best 'fit' to the data

Likelihood on Trees: The Model

- Each site in an alignment is assumed to evolve independently



- Want to calculate the probability of a site according to a particular tree



The model cont…

- Calculate the likelihood of all sites n, for a given tree topology T and parameters θ

$$L(\theta, T|s_1, ..., s_n) = L(\theta, T|S)$$

$$= \prod_{i=1}^{n} f(s_i|\theta, T)$$

- Need to calculate for each site:

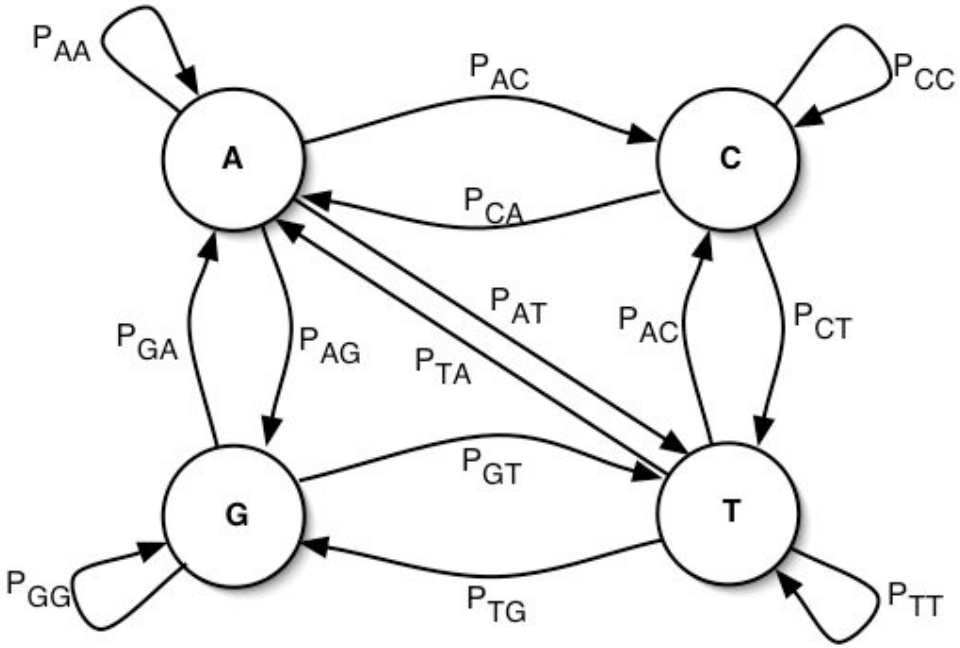
$$f(s_i|\theta, T)$$

Probability of change along a branch

- To compute $f(s_i|\theta, T)$ the probability of change along a branch must be defined
- What does this mean?
  - Want to calculate for all pairs of species, the probability of changing from state k to state l in site s_i along branch length t
- How is this done?
  - Continuous time discrete Markov chain

Discrete Markov Chain

- An example of a discrete Markov Chain for DNA



- We want to calculate the probability of change between each of the characters in a given site

Markov Condition

- Current state at time t (c_t)depends only upon previous state at time t-1 (c_{t-1})

$$P(c_t|c_{t-1}, c_{t-2}, ..., c_1) = P(c_t|c_{t-1}) = P_{c_t, c_{t-1}}$$

- For an m state Markov chain with states labelled from 1, ..., m, define transition matrix X

$$\mathbf{X} = \begin{pmatrix} P_{1,1} & \cdots & P_{1,m} \\ \vdots & \ddots & \vdots \\ P_{m,1} & \cdots & P_{m,m} \end{pmatrix}$$

- An n-step transition matrix is defined as

$$\mathbf{X}^n = \mathbf{X}\mathbf{X}\cdots\mathbf{X}$$

Properties of a Markov Chain

- It is possible to reach every state from every other state
  - If the chain is run for an infinite amount of time every state *i* will be reached with non-zero probability π_i known as the **stationary probability**
  - It is aperiodic
- Condition of detailed balance:

$$\pi_i X_{i,j} = \pi_j X_{j,i}$$

  - For $X_{i,j}$ the transition probability from state *i* to state *j*
  - Detailed balance implies the Markov Chain is **Time Reversible**

Probability of a given state…

- The ergodic property implies there is a stationary probability of being in state *i* π_i
- Based on this it is possible to calculate the probability of starting in state *i* and ending in state *j* after *k* time points

$$\begin{aligned} P(c_k = j, k \text{ ticks}, c_0 = i) &= P(c_k = j, k \text{ ticks} | c_0 = i) P(c_0 = i) \\ &= X_{i,j}^k \pi_i \\ &= R_{i,j}^k \end{aligned}$$

What if the time points aren't constant?

- In evolution, the changes from one state to another occur at different (non-constant) time intervals
- Let K be the random variable that describes the number of state transitions

$$K \sim Po(\mu t)$$

$$\text{and } P(K = k|\mu, t) = e^{-\mu t}\frac{(\mu t)^k}{k!}$$

Number of transitions…

- Furthermore, the number of transitions is **unknown**
- Let *P(t)* be the probability of change from one state to the next in time interval *t*. To calculate we sum over all possible number of transitions

$$
\begin{aligned}
P(t) &= \sum_{k=0}^{\infty} \mathbf{R}^k e^{-\mu t} \frac{(\mu t)^k}{k!} \\
&= e^{\mu t} \sum_{k=0}^{\infty} \frac{(\mathbf{R}\mu t)^k}{k!} \\
&= e^{\mu t} e^{\mathbf{R}\mu t} \\
&= e^{(\mathbf{R}-\mathbf{I})\mu t}
\end{aligned}
$$

Instantaneous Rate Matrix

- Thus we can define the instantaneous rate matrix from one state to another as

$$\mathbf{Q} = (\mathbf{R} - \mathbf{I})\mu$$

- Where the probability of change from one state to another along branch length *t*

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

DNA Models

- For DNA, the most general instantaneous matrix (GTR) has six rate parameters:

$$\mathbf{Q} = \begin{pmatrix} * & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ a\mu\pi_A & * & d\mu\pi_G & e\mu\pi_T \\ b\mu\pi_A & d\mu\pi_C & * & f\mu\pi_T \\ c\mu\pi_A & e\mu\pi_C & f\mu\pi_G & * \end{pmatrix}$$

- Diagonal elements are set so that rows sum to 0
- This is time reversible because

$$\pi_i \mathbf{Q}_{i,j} = \pi_j \mathbf{Q}_{j,i}$$

DNA models cont…

- For example

$$\pi_A \mathbf{Q}_{A,C} = \pi_C \mathbf{Q}_{C,A}$$

$$\pi_A a \mu \pi_C = \pi_c a \mu \pi_A$$

- Time reversibility is important because we don't know the position of the root of the tree

DNA Models cont…

- A more restricted model with fewer rate parameters the HKY model

$$\mathbf{Q} = \begin{pmatrix} * & \mu\pi_C & \kappa\mu\pi_G & \mu\pi_T \\ \mu\pi_A & * & \mu\pi_G & \kappa\mu\pi_T \\ \kappa\mu\pi_A & \mu\pi_C & * & \mu\pi_T \\ \mu\pi_A & \kappa\mu\pi_C & \mu\pi_G & * \end{pmatrix}$$

  - Allows for a rate parameter that describes the difference in the number of transitions versus transversions

Amino acid Models

- General amino acid model (amino acids labelled 1 through 20)

$$\mathbf{Q} = \begin{pmatrix} * & s_{1,2} & \cdots & s_{1,20} \\ \vdots & \ddots & & \vdots \\ s_{20,1} & \cdots & s_{20,19} & * \end{pmatrix} diag\,(\pi_1, \cdots, \pi_{20})$$

- Rate parameters are calculated based upon large protein alignment databases
- E.g. WAG
  - Estimate NJ tree T_i for each alignment A_i
  - Find maximum likelihood model as

$$L(M|T, A) = \prod_{\text{protein families } i} L(M|T_i, A_i)$$

Codon Models

- Not used for tree inference due to computational complexity of estimating transition probabilities
- However, popular for inferring sites under positive selection

$$q_{ij} = \begin{cases} 0 & \text{if i and j differ at two or three nucleotide positions} \\ \pi_j & \text{if i and j differ by one synonymous transversion} \\ \kappa\pi_j & \text{if i and j differ by one synonymous transition} \\ \omega\pi_j & \text{if i and j differ by one non-synonymous transversion} \\ \omega\kappa\pi_j & \text{if i and j differ by one non-synonymous transition} \end{cases}$$

Rate Heterogeneity

- Recall likelihood of data given model

$$L(\theta, T|s_1, ..., s_n) = L(\theta, T|S)$$

$$= \prod_{i=1}^{n} f(s_i|\theta, T)$$

- Where θ consists of branch lengths, and rate parameters in the Q matrix
- Some sites evolve more quickly, others more slowly
  - Want to modify the length of time for the site to evolve in the model accordingly by a rate R

$$P(tR)_{i,j} = e^{QRt}$$

Gamma rates across sites model

- Allowing for a rate of evolution for each site introduces n parameters for a data set with n sites

$$L(\theta, T, R_1, ..., R_n|S) = \prod_{i=1}^{n} f(s_i|\theta, T, R_i)$$

  - Model overfit
- Allow the rate of a site to vary according to the Gamma distribution

$$L(\theta, T, \alpha|S) = \int_0^\infty f(S|\theta, T, R)g(R|\alpha)dR$$

$$\approx \sum_{i=1}^{C} f(S|\theta, \lambda, T, \alpha, R_i)\frac{1}{C}$$

Bootstrap

- Data assumed to be IID according to the true distribution
- When sample size large enough empirical distribution $\hat{f}$ approximates true distribution f
- Thus can sample with replacement from empirical distribution in order to obtain new estimates for parameters
  - Typically for trees, internal node support is calculated based upon the number of times a branching occurs in the ML tree of the bootstrap samples

Problem

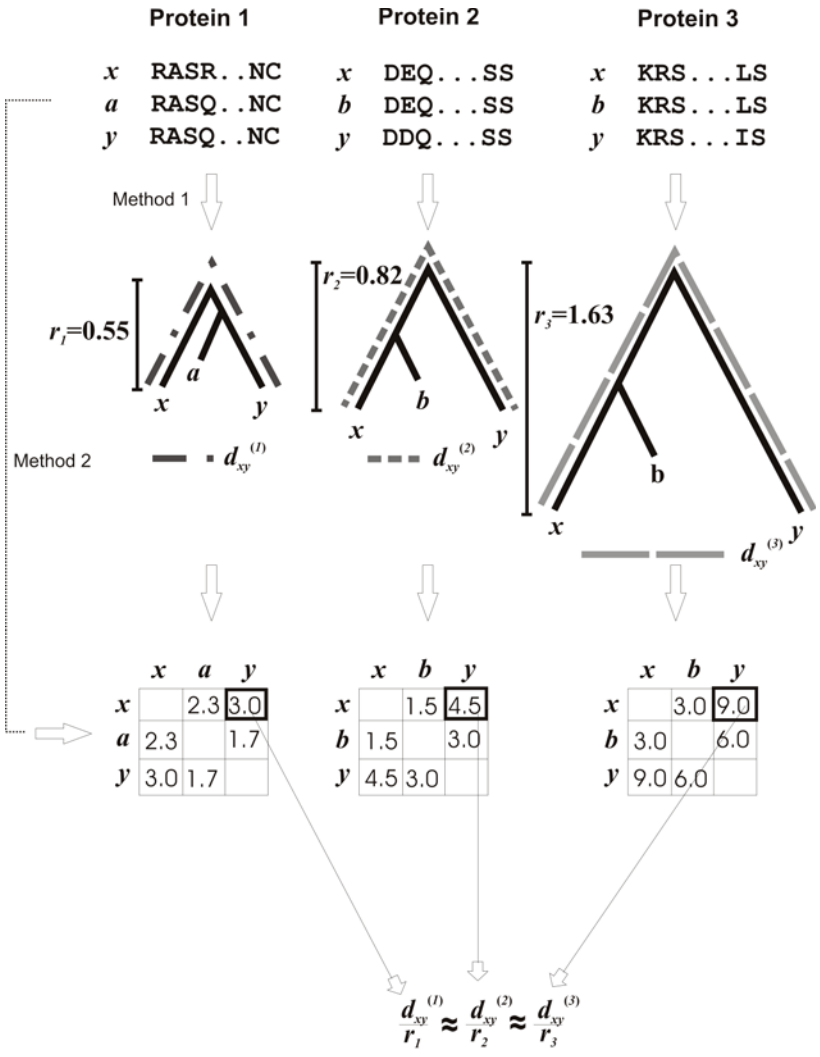
- Genes undergo different selective pressures
- Yang (1996, MBE), Pupko et al. (2002, MBE)
  - likelihood for a given tree is significantly better when incorporating gene rates into the model
- But…
  - Maximum likelihood methods for estimating such selective pressures are slow
  - Current methods assume a separate gene rate for each gene
    - Infinitely many parameters problem when there are many genes?

Approach

- Calculate gene rates *a priori*
- DistR method (Bevan et al.,Systematic Biology December 2005)
  - Calculates gene rates quickly
  - Gene rates closely approximate maximum likelihood rates
- Incorporate these rates into the ML tree search
  - Which models work best?

Estimating gene rates

- Estimate rates using DistR method
  - Fast, accurate method that doesn't need initial tree topology
  - uses pairwise distances between species for each gene



Method 1: Estimate tree from sequence data, then measure distances as sum of pairwise distance between taxa

Method 2: Estimate distances directly from sequence data

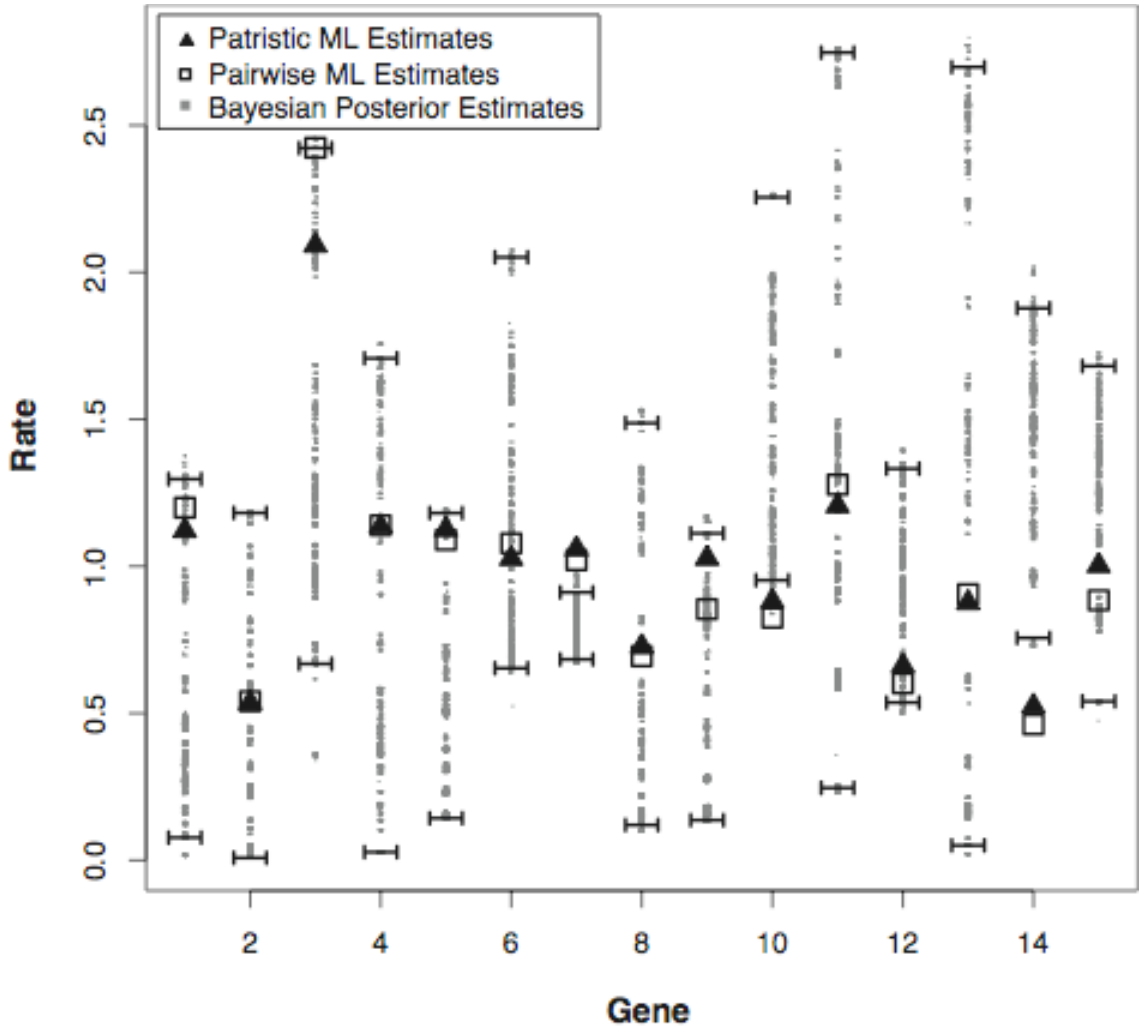
DistR gene rates

- Assume consensus distances $p_{x,y}$ and unknown rates $r_1, \ldots, r_n$
- use a weighted least squares framework to solve for $p_{x,y}, r_1, \ldots, r_n$

$$\sum_{k=1}^{n} \sum_{x,y \in G_k} \left(p_{x,y} - \frac{d_{x,y}^{(k)}}{r_k} \right)^2 Var(d_{x,y}^{(k)})^{-1}$$

- Note: This solution has an identifiability problem, thus a constraint must be used to find a unique solution

How good are DistR Esimates?

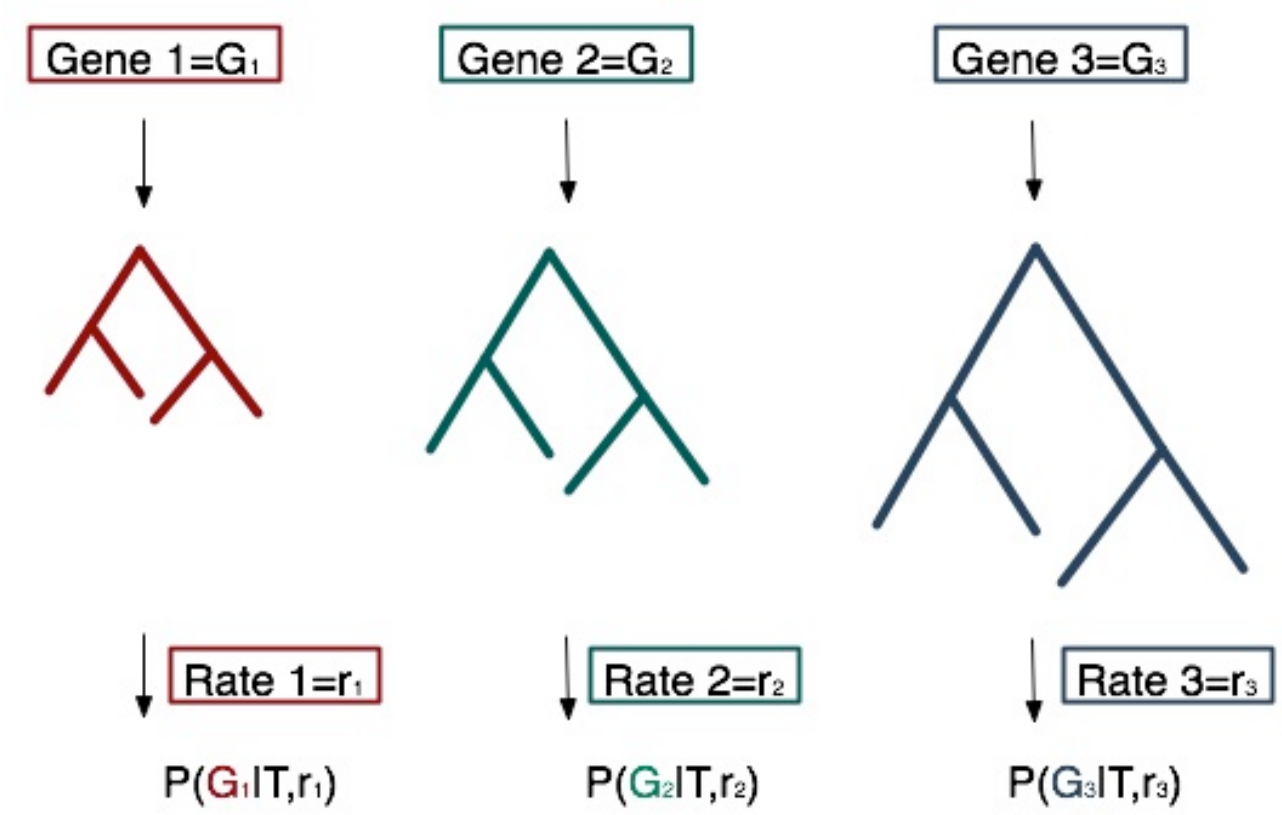


Based on 15 gene, 29 species fungal data set

How do we use DistR estimates?

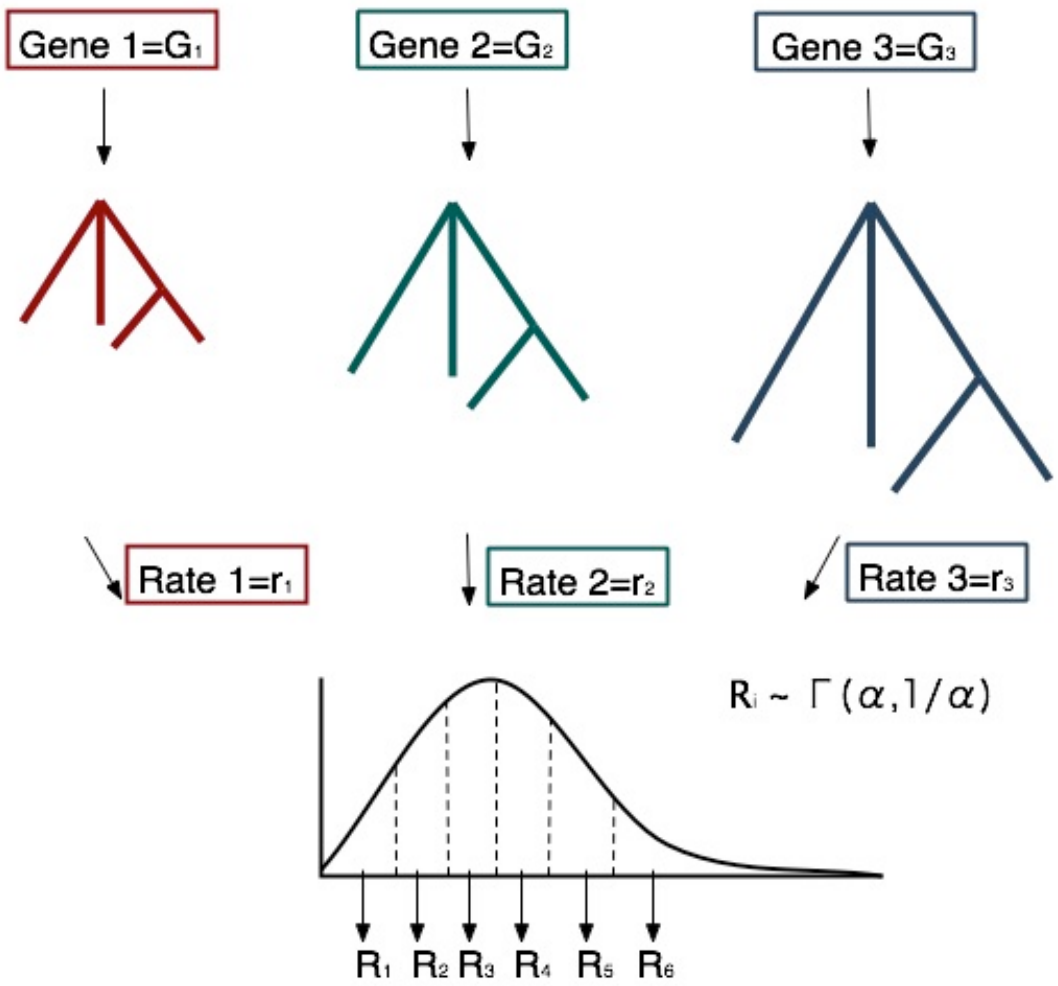
- n-parameter model: Each gene has it's own rate
  - Optimize over these rates as searching topology space
- α-parameter model: Each gene rate is drawn from distribution of rates
  - Optimize over this distribution as searching topology space

n-parameter model: Each gene has rate



$$P(\text{Data}|T,r_1,r_2,r_3) = P(G_1|T,r_1)\ P(G_2|T,r_2)\ P(G_3|T,r_3)$$

α-parameter model: Integrate



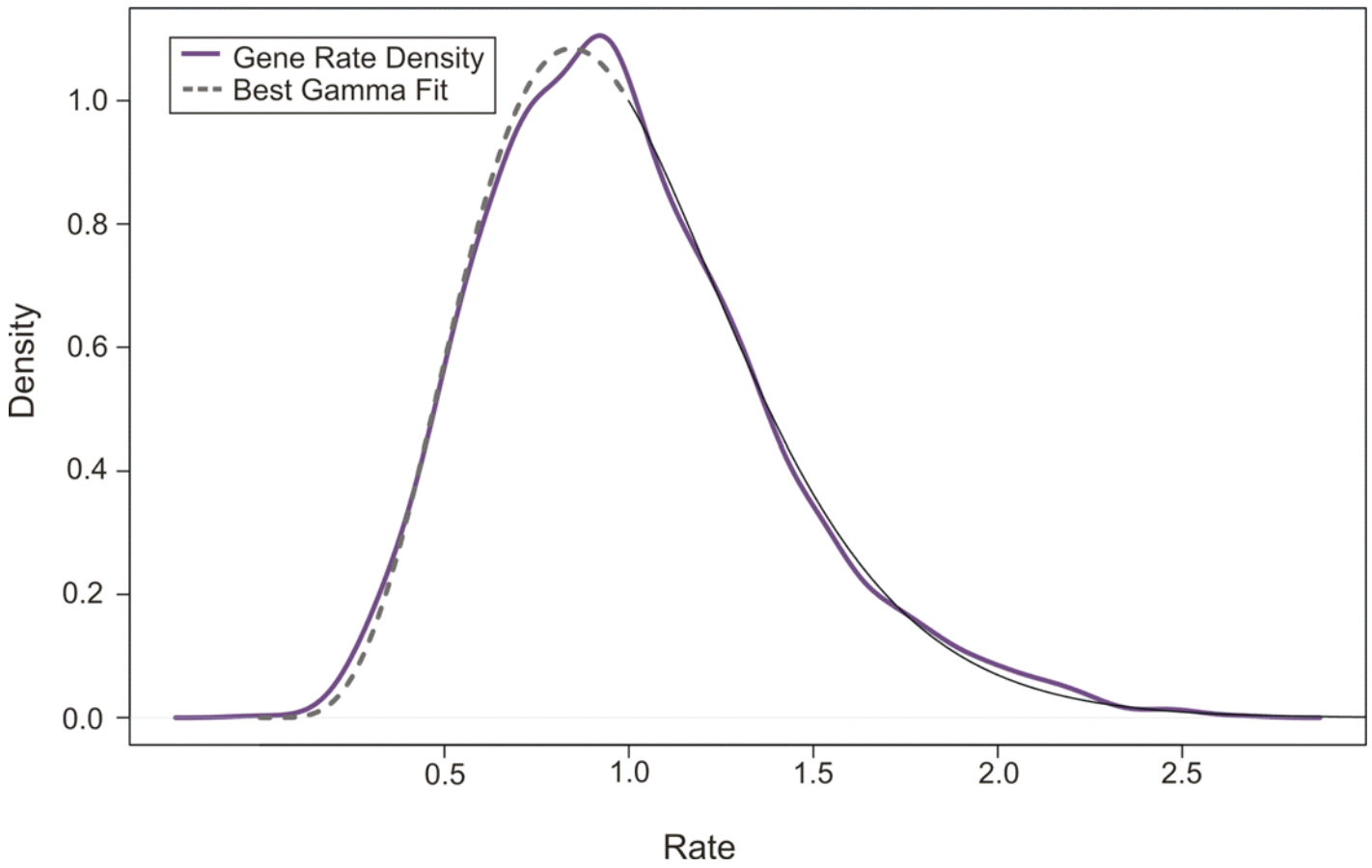
$$P(G_1|T,\alpha) =$$

$$P(R_1)\,P(G_1|T,R_1) + P(R_1)\,P(G_1|T,R_2) + \ldots + P(R_1)\,P(G_1|T,R_6)$$

$$P(\text{Data}|T,\alpha) = P(G_1|T,\alpha)\,P(G_2|T,\alpha)\,P(G_3|T,\alpha)$$

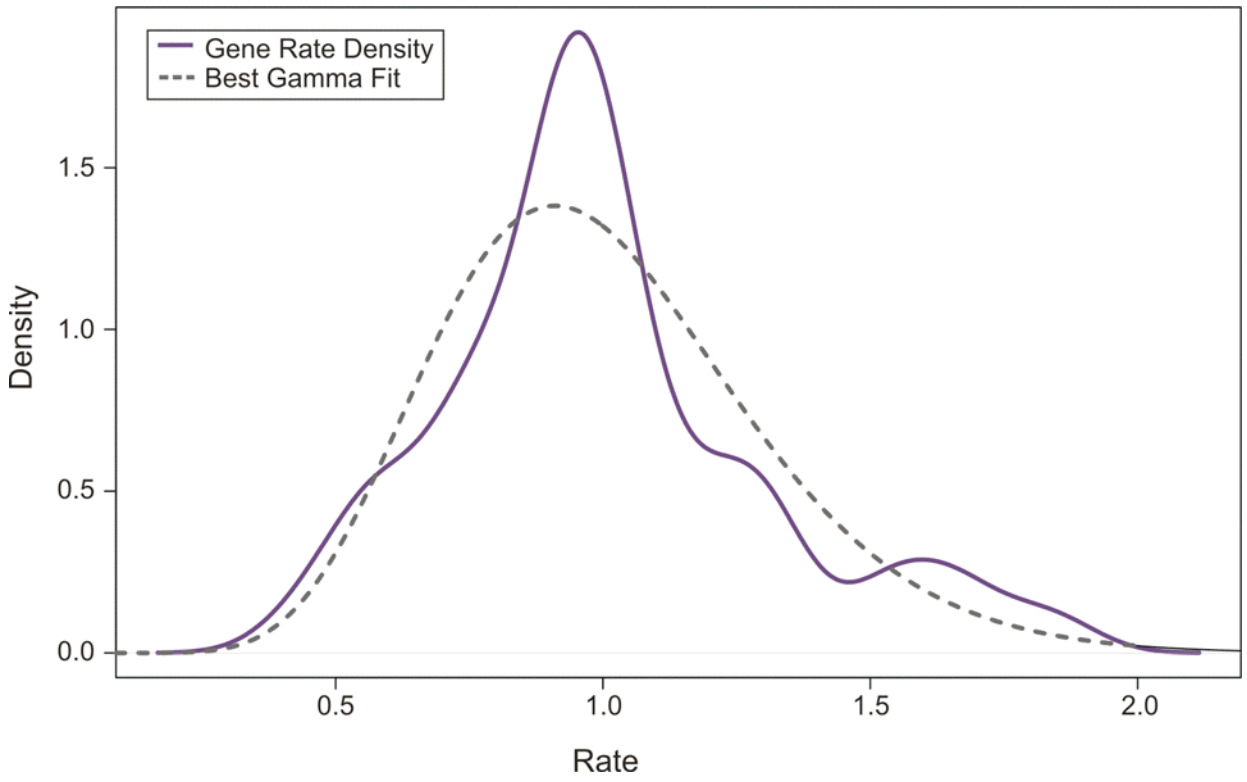
Gene rates distribution: Fit of gamma

- gamma distribution has excellent fit to data despite missing distances, few species per gene etc.

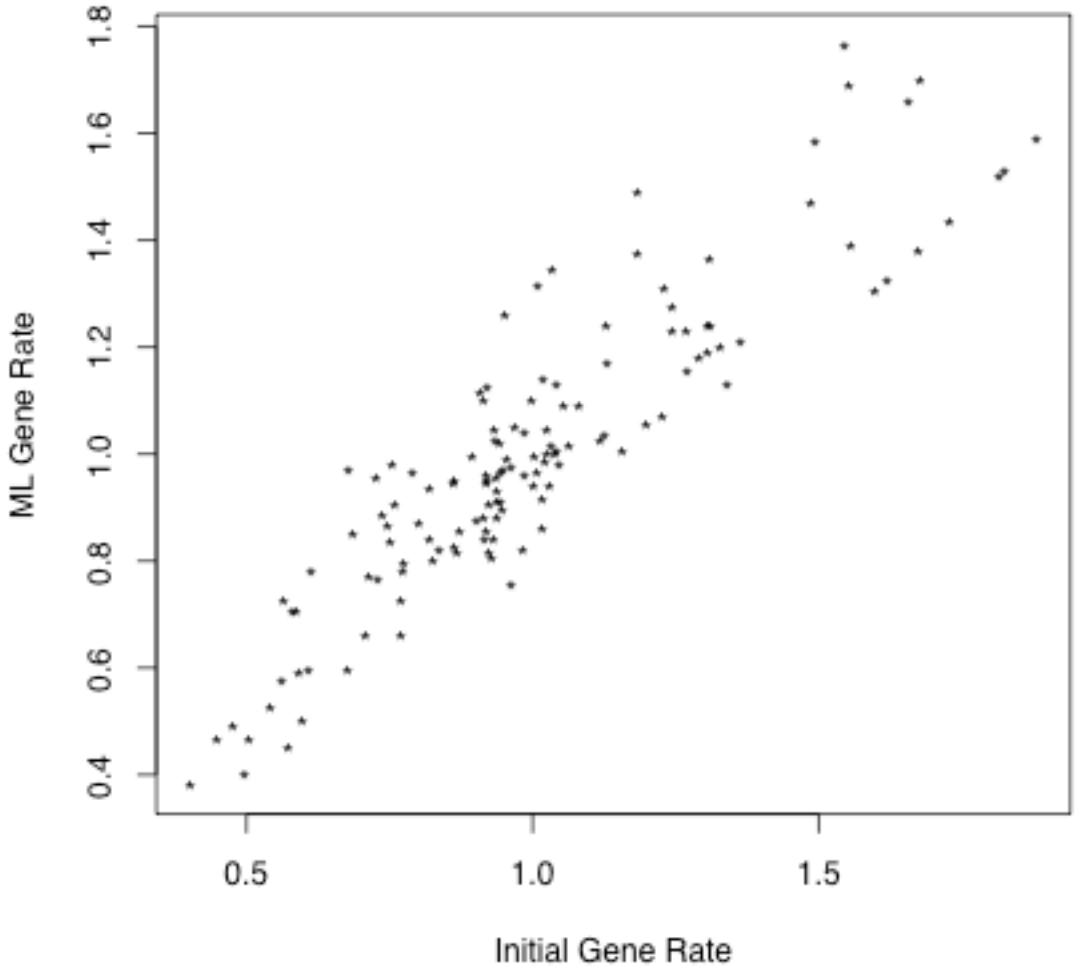


However…

- For individual data sets the fit is not as good
  - 133 genes, 44 species Brinkman et al. Sys. Bio. 2005



How accurate are DistR estimates?



n-parameter model versus α–parameter model

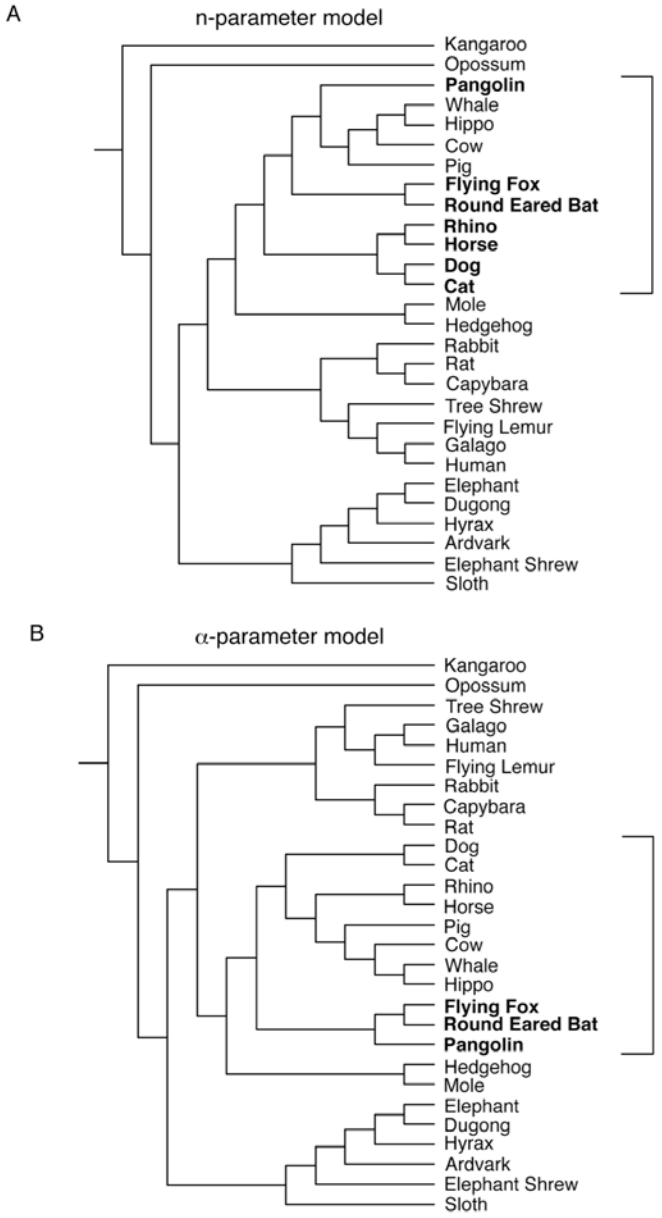
- Does the added computational time of α-parameter model help to find a better fit to the data?
  - ΔAIC for concatenated versus specified model, with a positive value indicating preference for gene rates model

Data Set	Num. Genes	Num. Species	n–parameter one–gamma	n–parameter gene–gamma	α–parameter one–gamma	α–parameter gene–gamma
Fungal mtDNA	15	29	1026.30	1152.42	893.04	1011.61
Eukaryotic	133	44	1529.58	2481.76	1298.86	2203.15
Madsen	4	28	153.05	426.75	149.95	423.96
Madsen – PT	4	28	162.24	435.78	152.79	427.46
Animal mtDNA	12	56	248.47	379.50	221.50	332.97
Murphy	6	46	189.08	296.25	186.91	284.56

Change in AIC favouring gene rates model over concatenated model

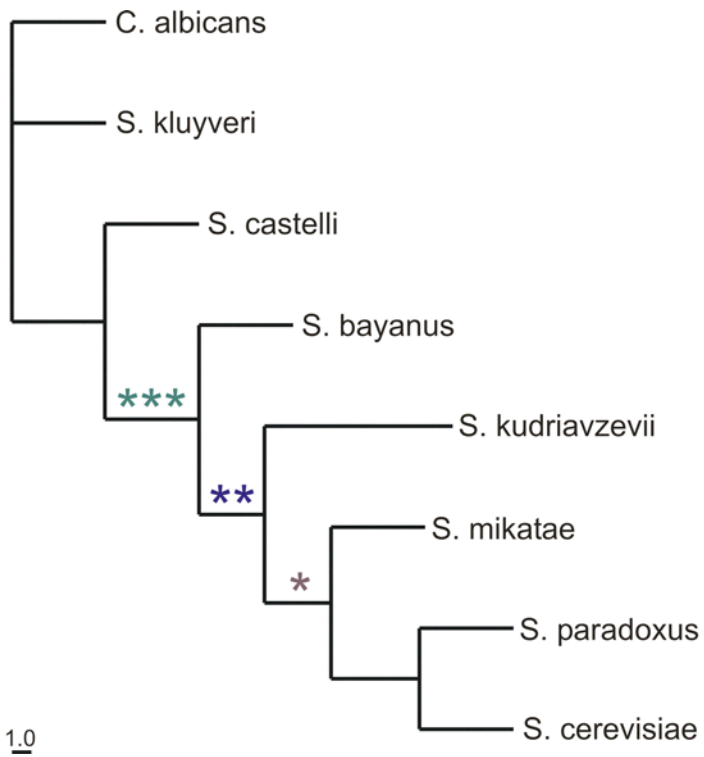
Madsen data set

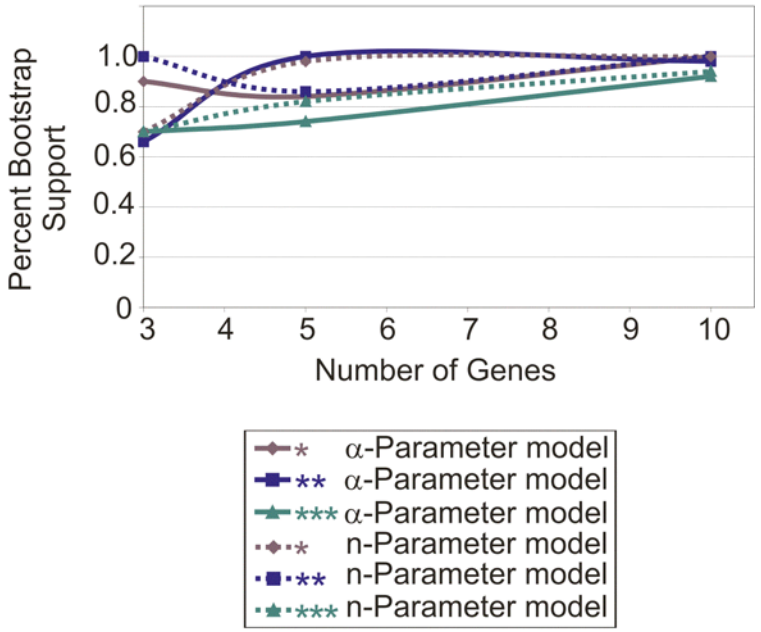
- How is the maximum likelihood topology affected n-parameter model versus α-parameter model?



Rokas data set

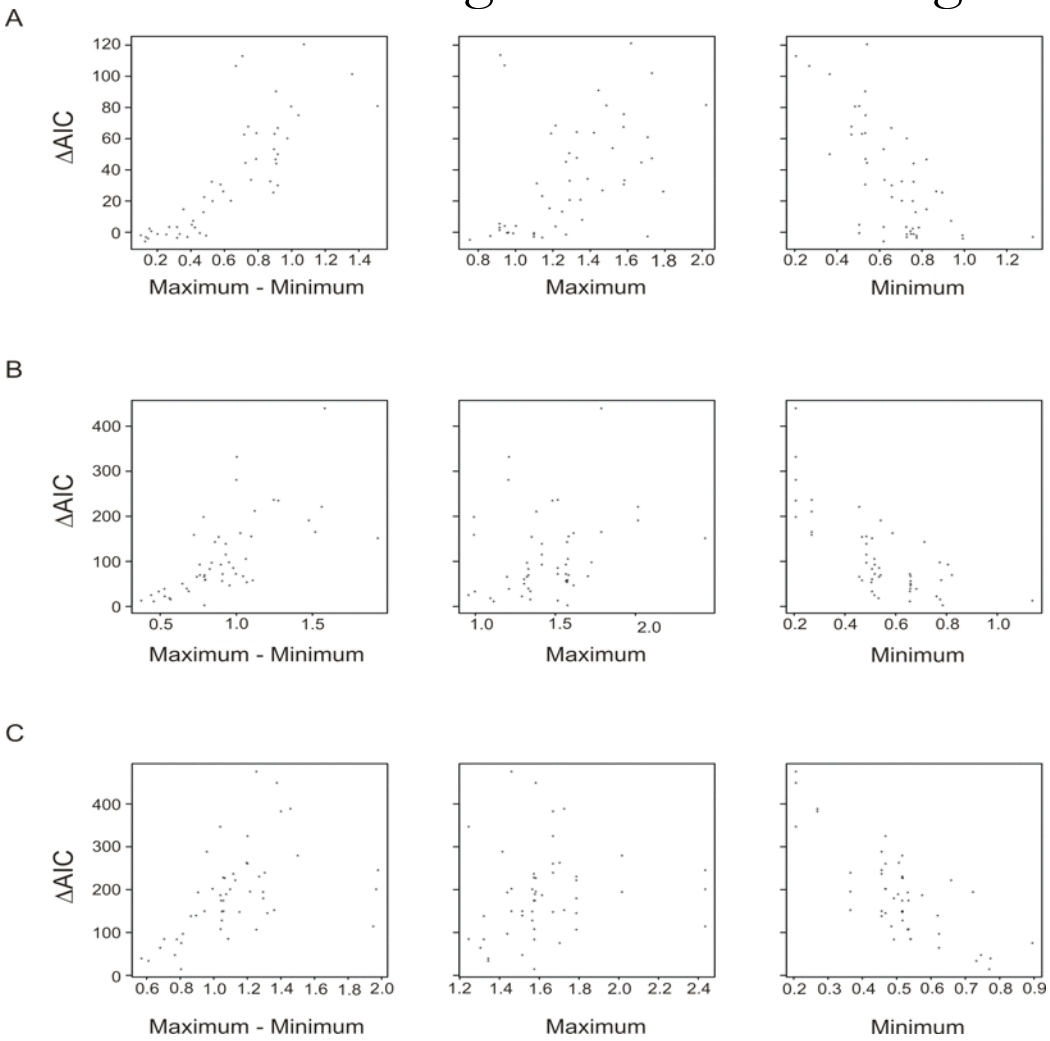
- At what point do the two methods find congruent topologies at least 95% of the time?
- Sample 3 genes, 5 genes and 10 genes from 106
  - Repeat 50 times



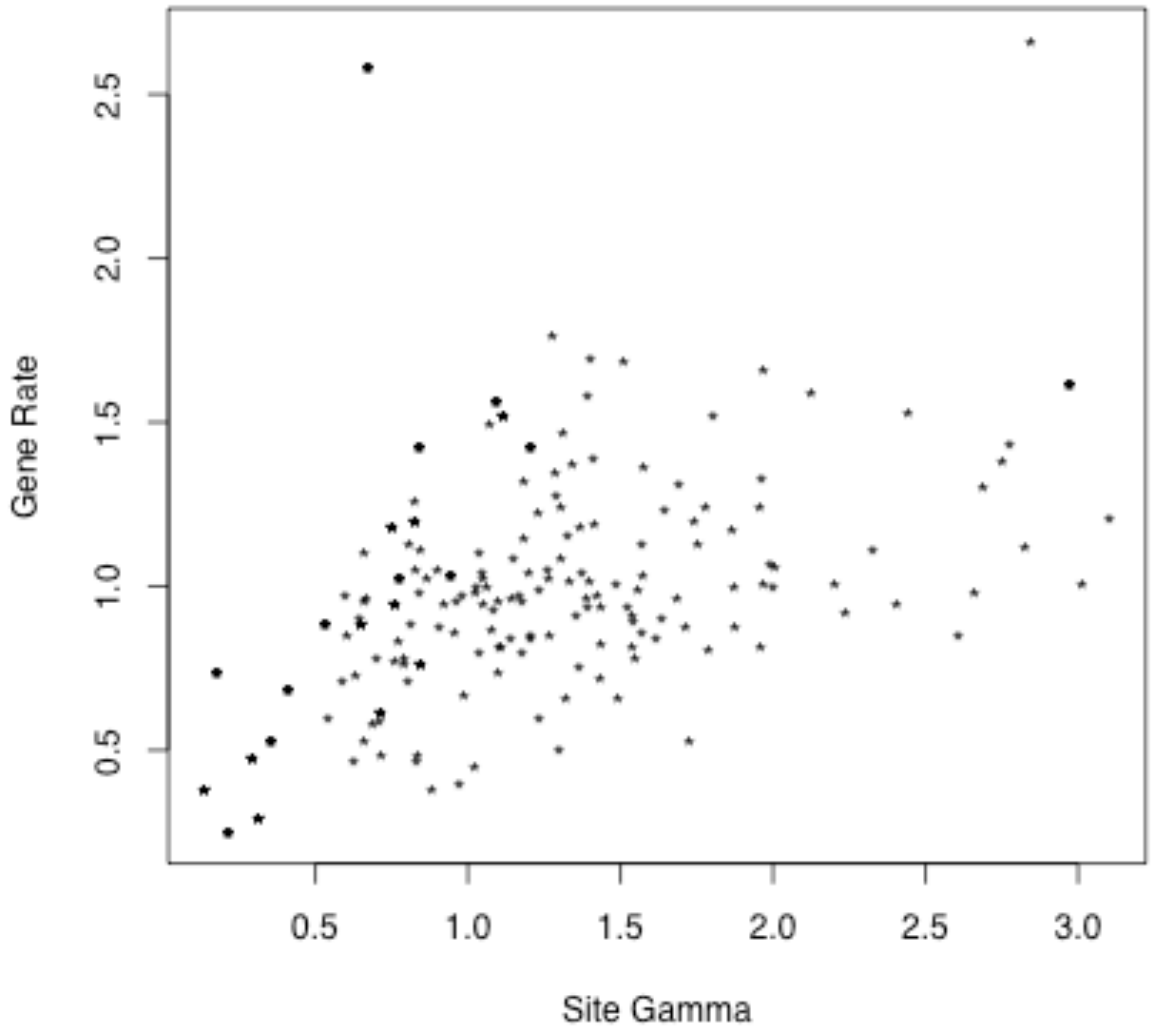


What explains the ΔAIC?

- Rokas data set, resampled genes
- Correlation of ΔAIC highest with slowest gene rate



Site Rate Heterogeneity vs Gene Rate



Conclusions

- ML inference in phylogenetics presented
  - Calculating probability of a site along a tree
  - Accounting for site rate heterogeneity
  - Bootstrap
- Extension of the basic model to include gene rate heterogeneity
  - DistR gene rate estimates are excellent starting estimates in phylogenetic analyses
  - More improvement with the gene rates models is found when slower genes are present in the data set
  - The computational effort required by the α-parameter model does not (on average) lead to a better fit of the model to the data
- How best to account for both site rate and gene rate heterogeneity is not clear
- Thanks for listening!
- Thanks to DIMACS.