Amino Acid Models and Deep-Time Phylogenetic Inference

Edward Susko

Department of Mathematics and Statistics, Dalhousie University

Dalhousie University (Halifax, Nova Scotia)

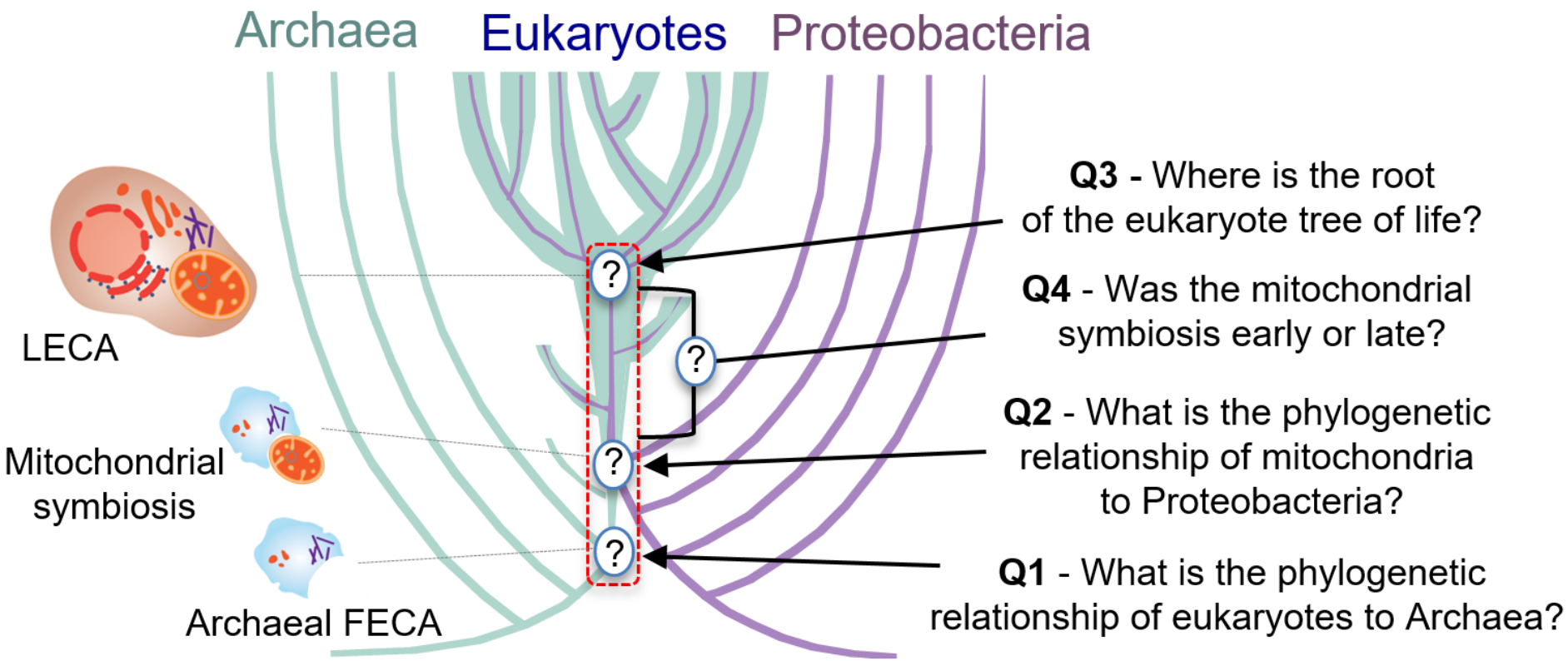
Joe Bielawski

- Codon Models
- Adaptive Evolution
- Non-adaptive Evolution
- Multi-level Selection

Andrew Roger

- Amino Acid Models
- Phylogenetic Estimation
- Deep Divergences

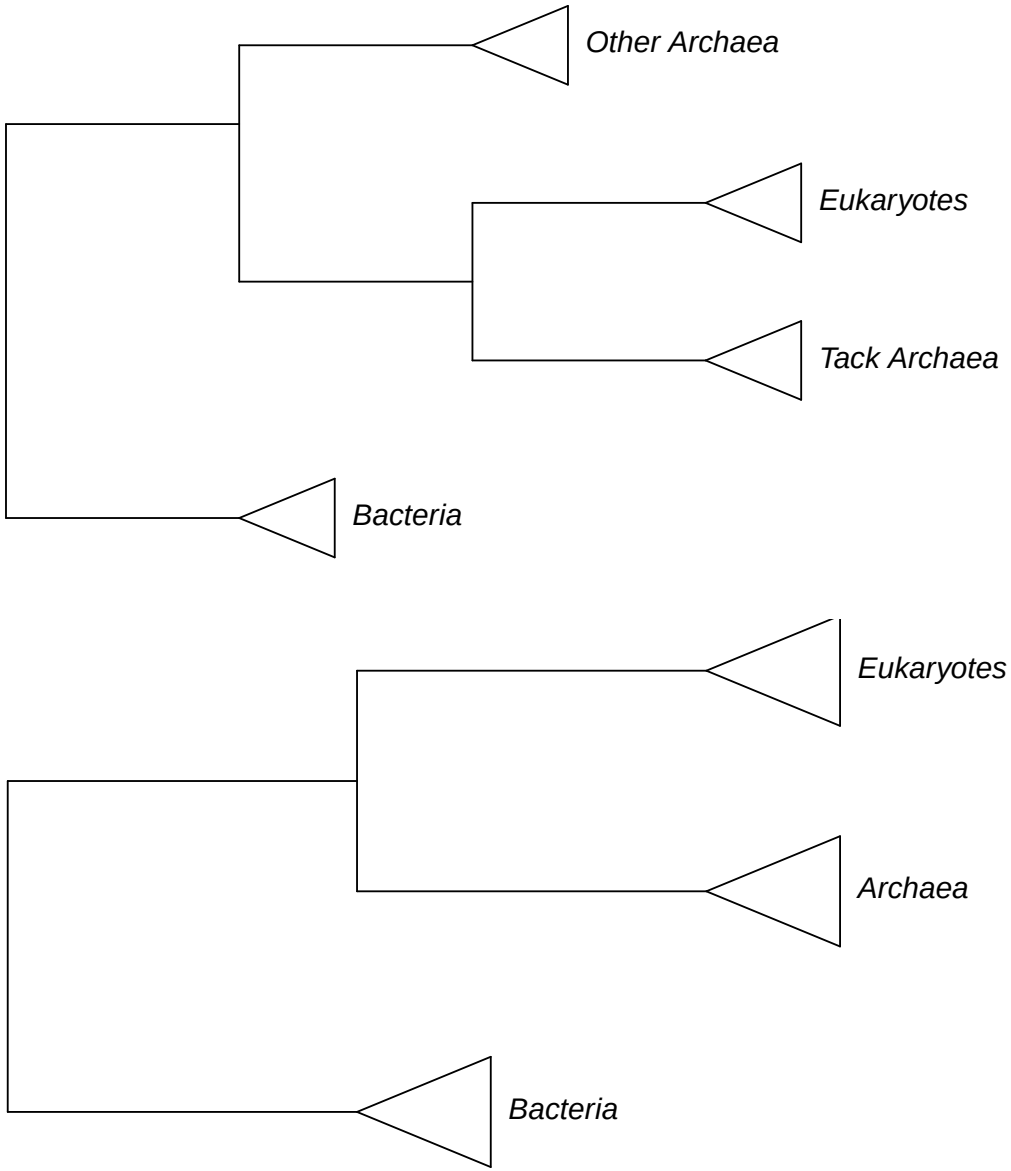
Eukaryogenesis Questions



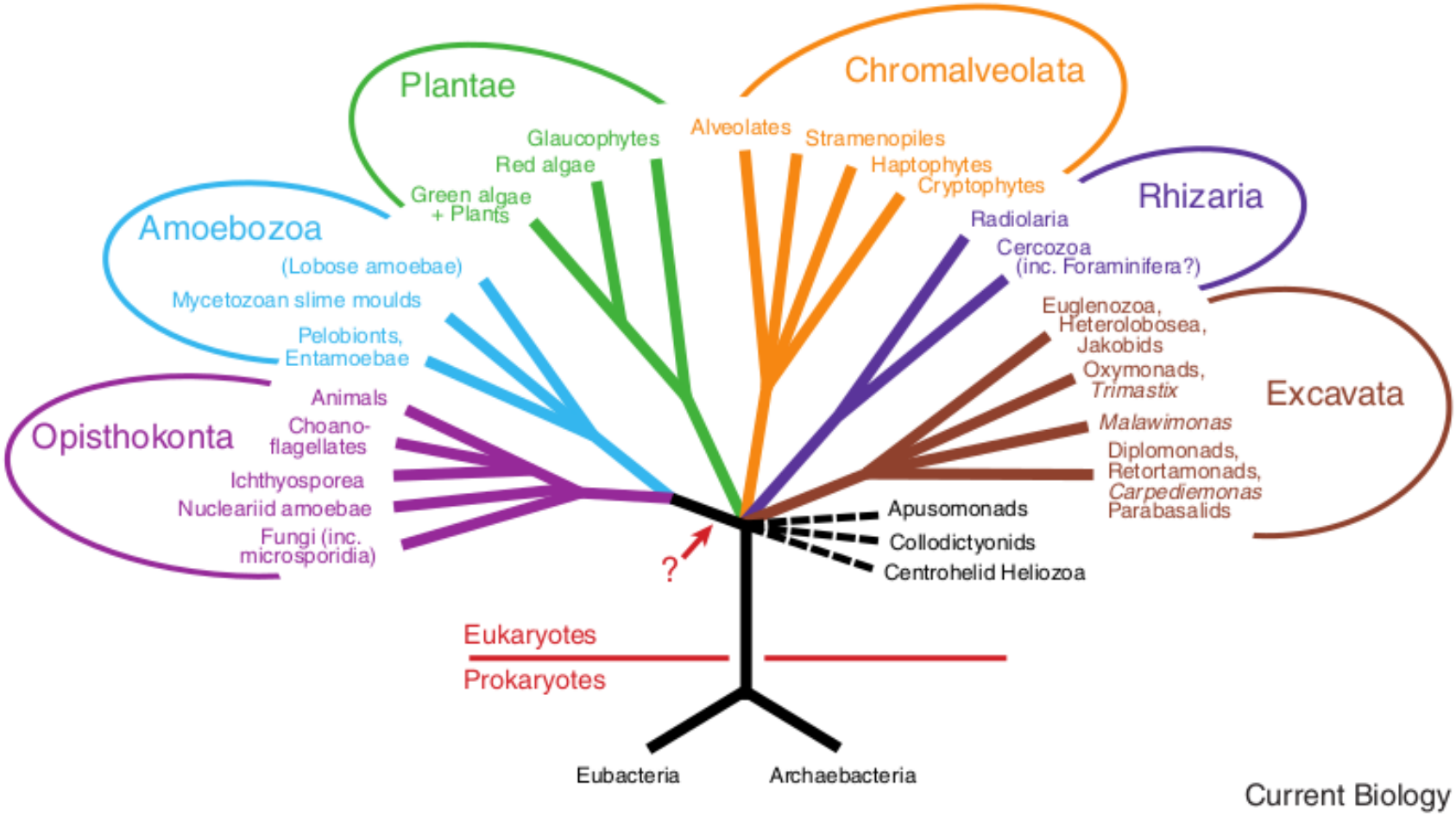
GORDON AND BETTY MOORE FOUNDATION

SIMONS FOUNDATION

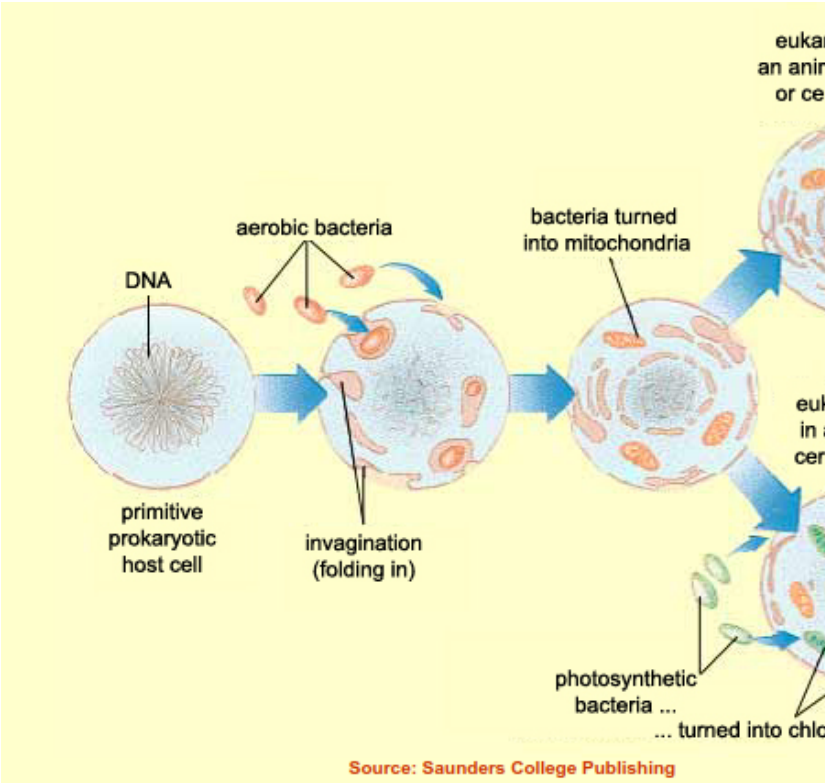
Phylogenetic Relationship of Eukaryotes to Archaea

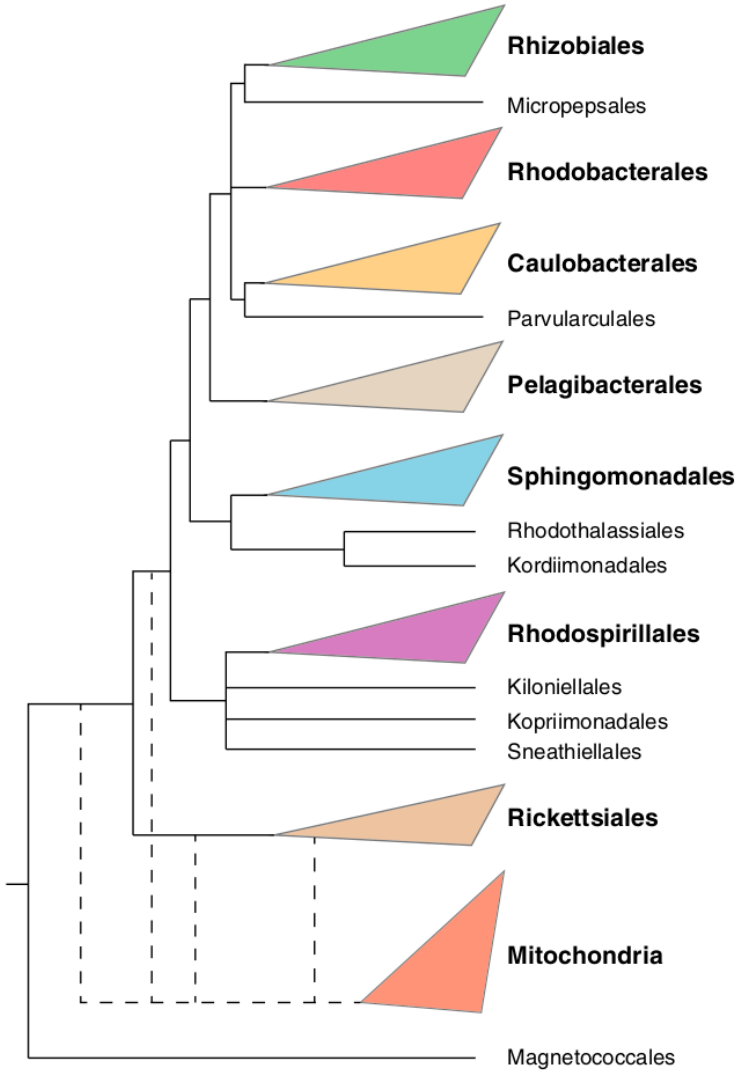


Root of the Tree of Eukaryotes



Mitochondrial Origins





Difficulties with Deep Divergences ($>$ 500Mya)

- Saturation of sequence changes over time
- Rapid radiations within clades

  $\Longrightarrow$ Enormous differences between taxa

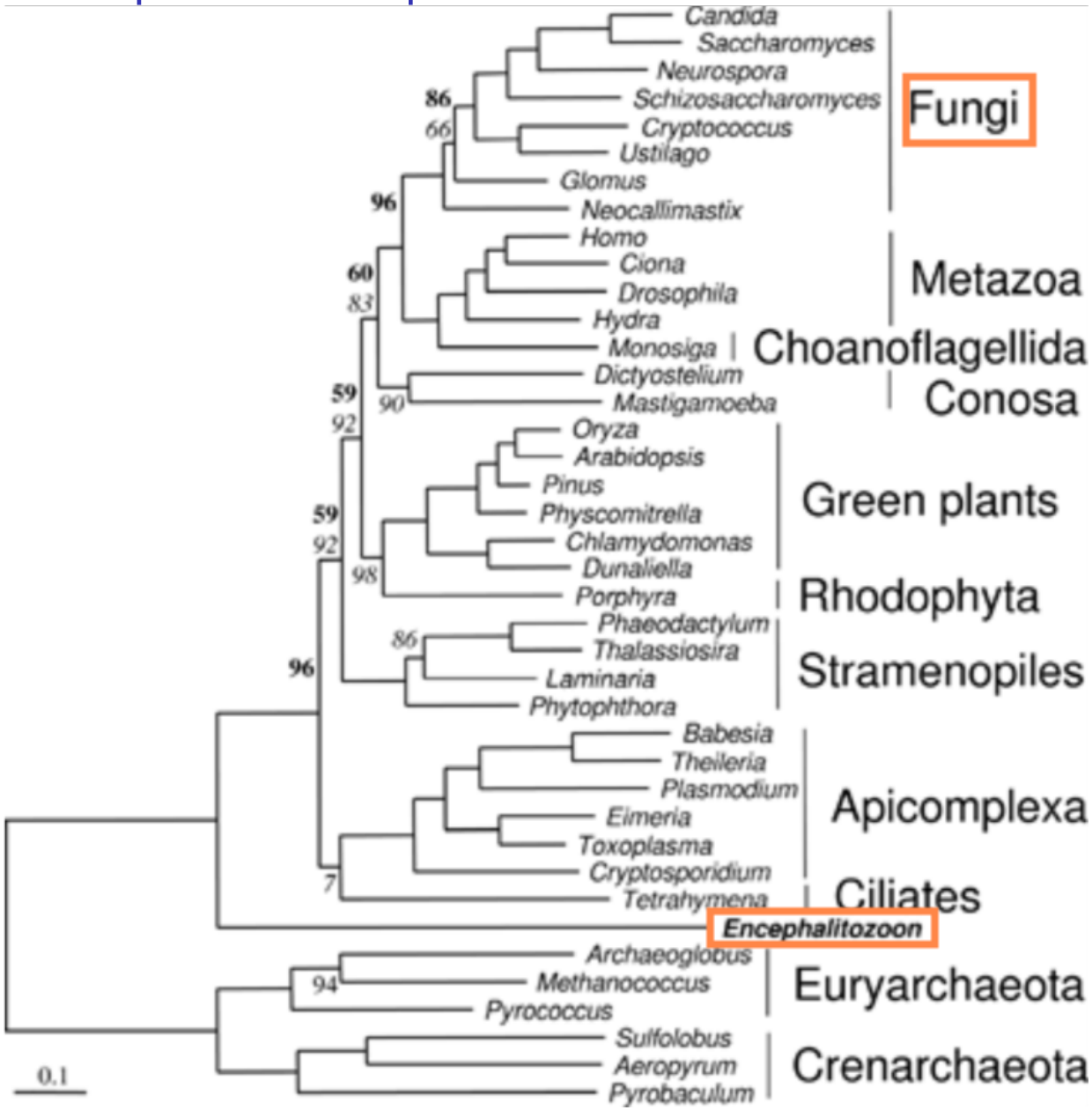
Amino acid data rather than DNA

GTG	CTG	CCT	GCC	GAC	AAG
...	...	G.C	...	...	...
...	...	..C	..T	...	...
...	..C	G.A	.AT	...	..A
...	..C	GA.	..T	...	...
becomes					
G	C	L	V	A	K
G	C	V	V	A	K
G	C	L	V	A	K
G	C	V	A	A	K
G	C	A	V	A	K

Difficulties with Deep Divergences ($>$ 500Mya)

- Saturation of sequence changes over time (loss of information)
- Process variation over time and genomic location
  - Phlyogenomic approaches: concatenation of multiple genes/proteins
- Gene tree vs species tree discrepancies
  - Incomplete lineage sorting
  - Lateral (horizontal) gene transfer

Microsprodia Example



Brinkman et al. (2005). Syst. Biol. 54:743–757.

Concatenated Protein Set

	Site				
	1	2	3	...	n
Homo	S	E	S	...	-
Enceph	Y	E	K	...	S
Schizo	I	E	N	...	S
Saccha	I	D	N	...	S
...					

- Each protein has $n \approx 300$. 133 proteins
- Concatenated sets large $n = 24291$
- **Observational unit ($\boldsymbol{x}_h$)**: vector of data at a site

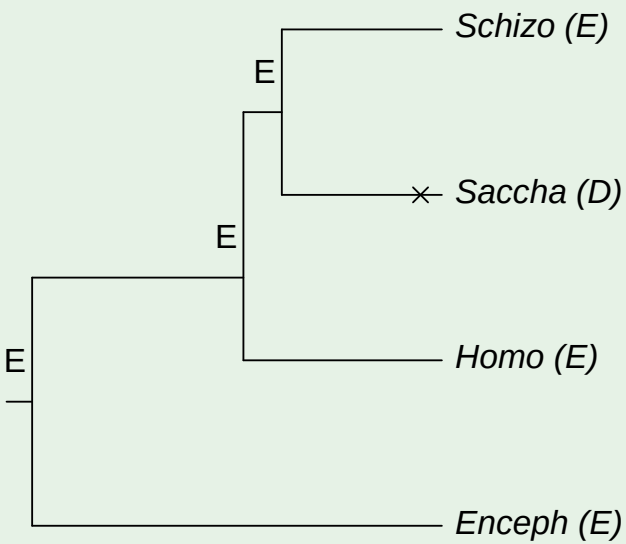
- Usually data at sites are treated as independent. Likelihood

$$\text{Likelihood} = L(\tau, \boldsymbol{t}, \theta) = \prod_h p(\boldsymbol{x}_h; \tau, \boldsymbol{t}, \theta)$$

τ - topology
$\boldsymbol{t}$ - edge lengths
θ - other parameters

Base Model for a Site

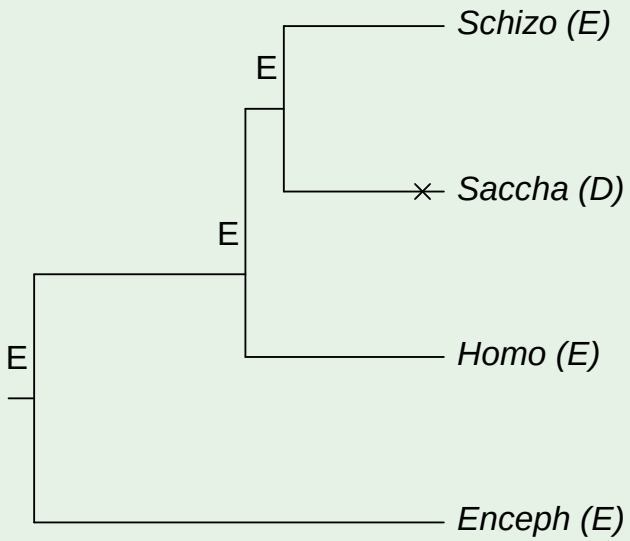
Evolution at a site



- Evolution is assumed independent across sites.
- Evolution along edge is conditionally independent, given ancestral node data.
- Evolution along edge is according to a stationary, time-reversible, continuous-time Markov Chain.

Likelihood Calculation

Evolution at a site



- x - observed data at tips
- a - unobserved ancestral data

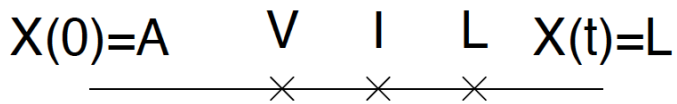
- Complete Data Site Likelihood

$$p(x, a; \tau, \boldsymbol{t}, \boldsymbol{\theta}) = \pi_E P_{EE}(t_M) P_{EE}(t_{HF}) P_{EE}(t_F) P_{EE}(t_{Sc}) P_{ED}(t_{Sa})$$

- Site Likelihood $P(x; \tau, \boldsymbol{t}, \boldsymbol{\theta}) = \sum_a P(x, a; \tau, \boldsymbol{t}, \boldsymbol{\theta})$
  - Pruning algorithm (Felsentstein 1981) for efficient computation

Base Model for an Edge

- Substitution matrix: $P(t)$ (20×20) $P(t)_{ij} = P(X(t) = j | X(0) = i)$

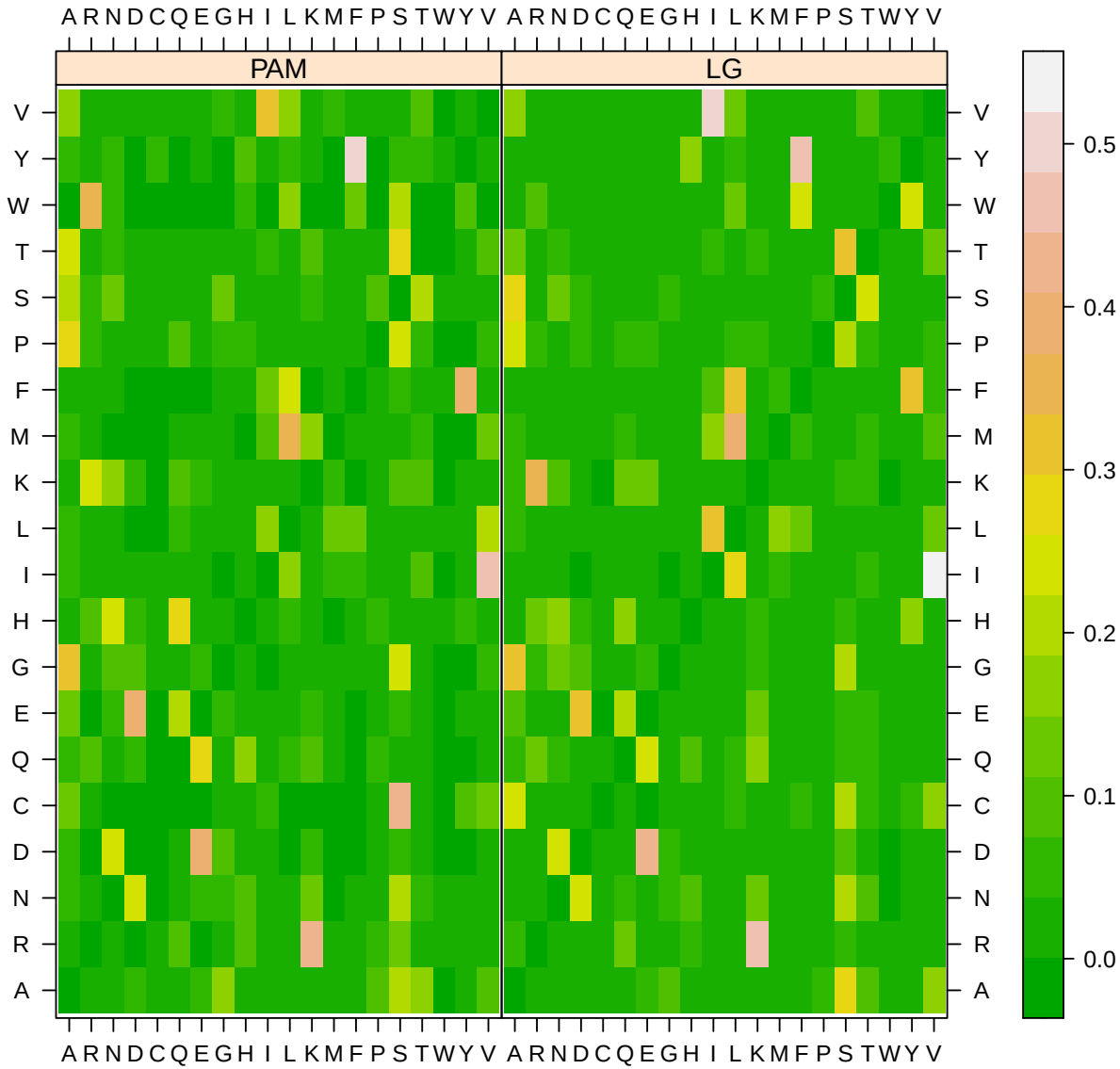


$$P(t) = \exp[Qt] = \sum_{k=0}^{\infty} (Qt)^k / k!.$$

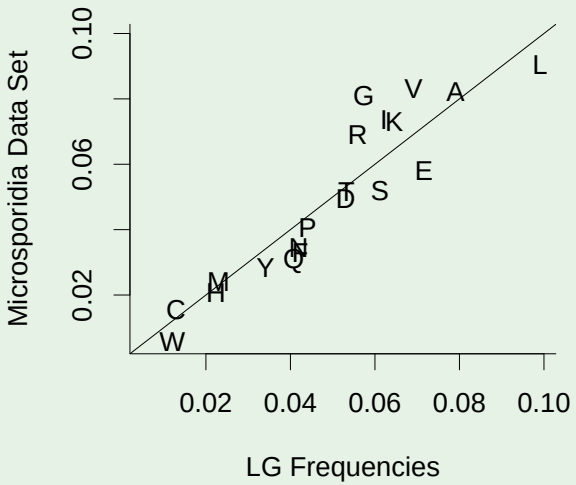
$$P[\text{Next amino acid} = j | \text{Current} = i] \propto Q_{ij}$$

- Q: Empirically derived from large data base and then fixed. PAM (1979), JTT (1999), WAG (2001), LG (2008)
  - PAM - Parsimony
  - LG - ML estimation with rate variation

Rate Matrix Comparison - P[Next amino acid = j|Current =i]



Frequency Comparison



- LG comes with frequencies
- Data set frequencies often differ
- Data set frequencies: simple proportion over all sites and taxa

- Stationary, time-reversible model $\iff Q_{ij} = S_{ij}\pi_j$ where $S_{ij} = S_{ji}$
  - $S_{ij} = Q_{ij}/\pi_j$ - exchangeabilities
- Model with data set frequencies: $\hat{Q}_{ij} = S_{ij}\hat{\pi}_j$

Rate Variation

- Rates of evolution vary substantially across sites

Data for Two Sites (Low and High Rate)

```
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
KTK-MANNAAM-D-AAASASSANKKAKNKAS-AA---AA-
```

- Simple Approach: Include them as parameters. Likelihood

$$L(\tau, \boldsymbol{t}, \boldsymbol{\theta}) = \prod_h p(\boldsymbol{x}_h; \tau, \boldsymbol{t}, \boldsymbol{\theta})$$

becomes

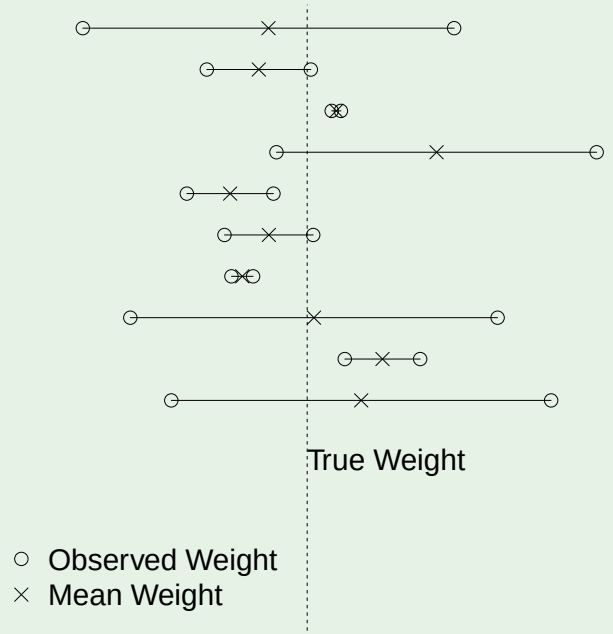
$$L(\tau, \boldsymbol{t}, \boldsymbol{\theta}, \boldsymbol{r}) = \prod_h p(\boldsymbol{x}_h; \tau, r_h \boldsymbol{t}, \boldsymbol{\theta})$$

- Additional parameters: $r_1, \ldots, r_n$
  - Number of parameters $\to \infty$ as $n \to \infty$

Neyman-Scott Problem

- Observed Weights $X_{ij} \sim N(\mu_i, \sigma^2)$
- μ_i - True Weight
- σ^2 - Variance of Scale
- Additional parameters: $\mu_1, \ldots, \mu_n$
- Number of parameters $\to \infty$ as $n \to \infty$

Example Data



- Want variability of X_{ij} about μ. Instead estimate variability of X_{ij} about $\bar{X}_i$.
- As $n \to \infty$,

$$\hat{\sigma}^2 \to \sigma^2/2$$

- Observed Weights $X_{ij} \sim N(\mu_i, \sigma^2)$
- Mixture Model: Assume μ_i i.i.d. from G (eg. $G \sim N(\mu_0, \tau_0^2)$)

$$p(x_{ij}, \mu_i; \sigma^2, G) = p(x_{ij}|\mu_i; \sigma^2) g(\mu_i; \mu_0, \tau_0)$$

- ML estimation: $(\hat{G}, \hat{\sigma}^2)$ maximize

$$L(G, \sigma^2) = \prod_{ij} \int p(x_{ij}|\mu_i; \sigma^2) g(\mu_i; \mu_0, \tau_0) \, d\mu_i.$$

- Even if μ_i are fixed constants satisfying that $|\mu_i| \leq M$,

$$\hat{\sigma}^2 \to \sigma^2$$

General Finite Mixtures

- General Mixture: $X_i|\theta_i \sim p(x_i|\theta_i, \zeta)$ and $\theta_1, \ldots, \theta_n$ iid from G

$$L(G, \zeta) = \prod_i \int p(x_i|\theta, \zeta) g(\theta)\, d\theta$$

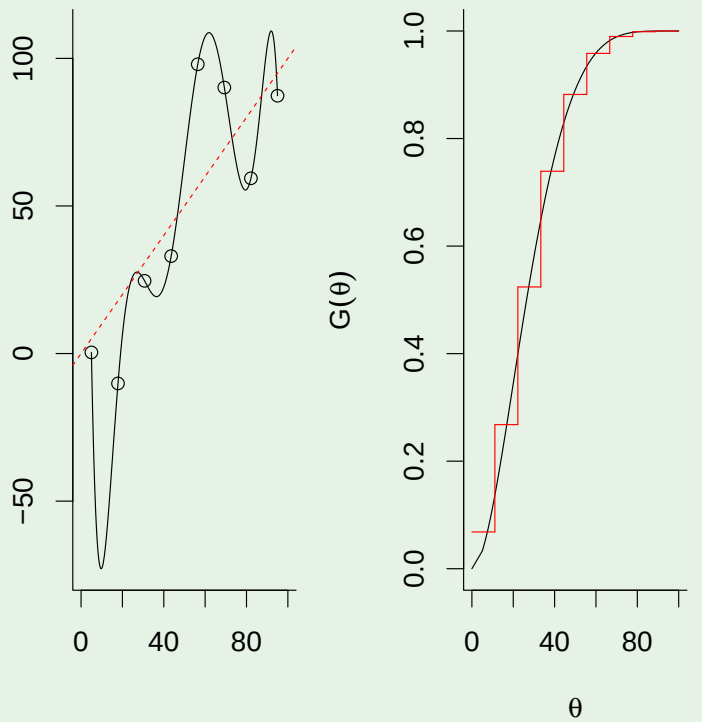
- Finite Mixture: G is a finite distribution: w_c is probability of θ_c, $c = 1, \ldots, K$.

$$L(\boldsymbol{w}, \boldsymbol{\theta}, \zeta) = \prod_i \sum_c w_c p(x_i|\theta_c, \zeta)$$

- Lindsay (1983): Maximizer of $L(G, \zeta)$ will always be the same as $L(\boldsymbol{w}, \boldsymbol{\theta}, \zeta)$ for some choice of K.
- Kiefer and Wolfowitz (1956): 'Usually' $(\hat{G}, \hat{\zeta}) \to (G, \zeta)$ as $n \to \infty$.
  - Note: $K \to \infty$ if G is continuous

Mixtures are not Overparameterized

Function Spaces



- As a space of functions,
  - space of polynomials are large
  - space of distribution functions are small

Rates Across Sites

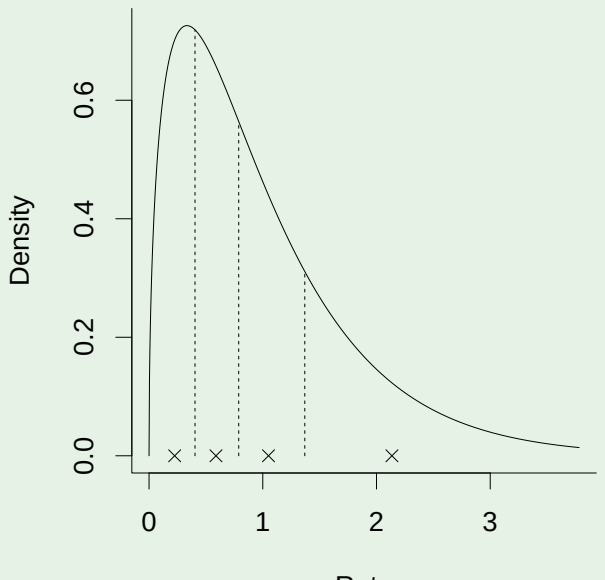
- Rates of evolution usually vary substantially across sites
- Mixture Approach: $r_1, \ldots, r_n$ i.i.d. w_c probability that rate is r_c

$$L(\tau, \boldsymbol{t}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{r}) = \prod_h \sum_c w_c p(\boldsymbol{x}_h; \tau, r_c \boldsymbol{t}, \boldsymbol{\theta})$$

- Finite mixtures are necessary. Integration breaks the pruning algorithm

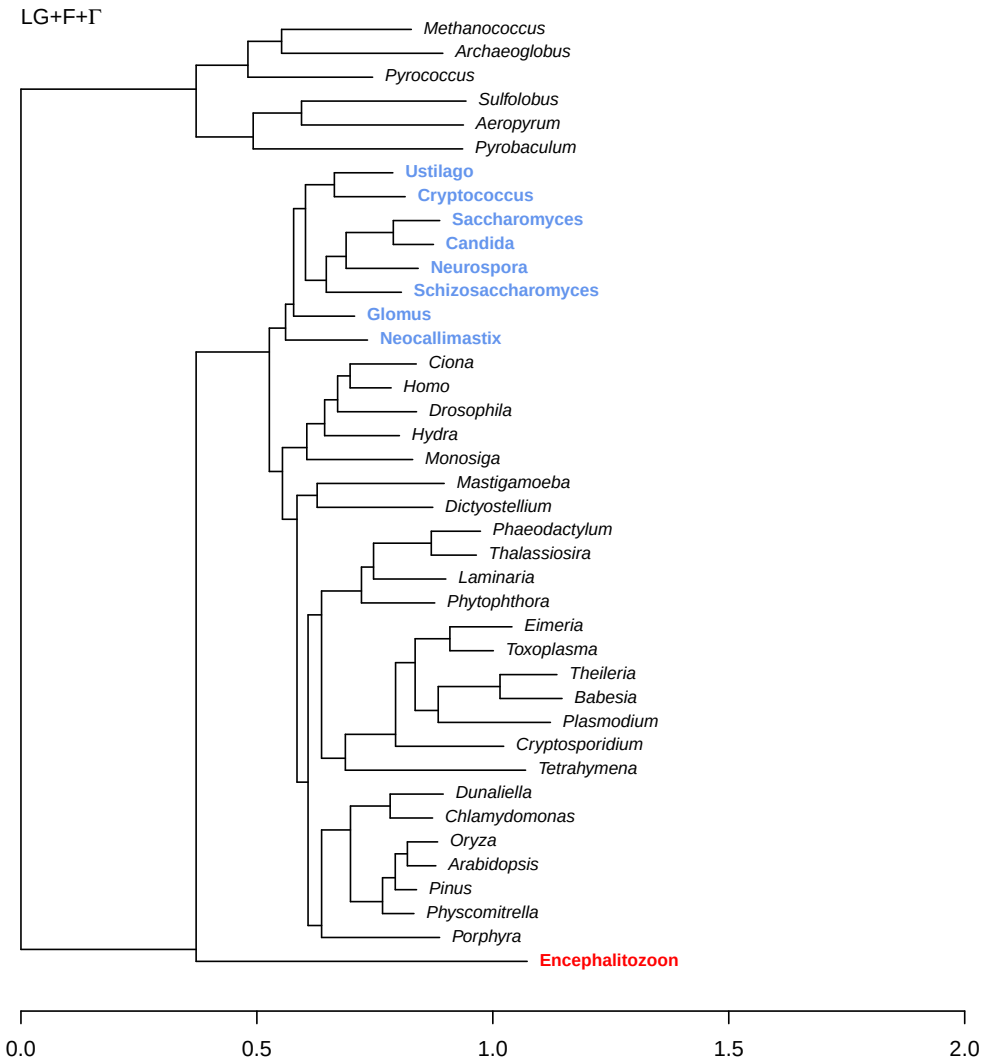
- Gamma model (Yang 1994)
- $w_c = 1/K$ and $r_c(\alpha)$ conditional mean of Gamma(α,α)
- $L(\tau, \boldsymbol{t}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{r})$ becomes $L(\tau, \boldsymbol{t}, \boldsymbol{\theta}, \alpha)$

Gamma Distribution $\alpha = 1.5$

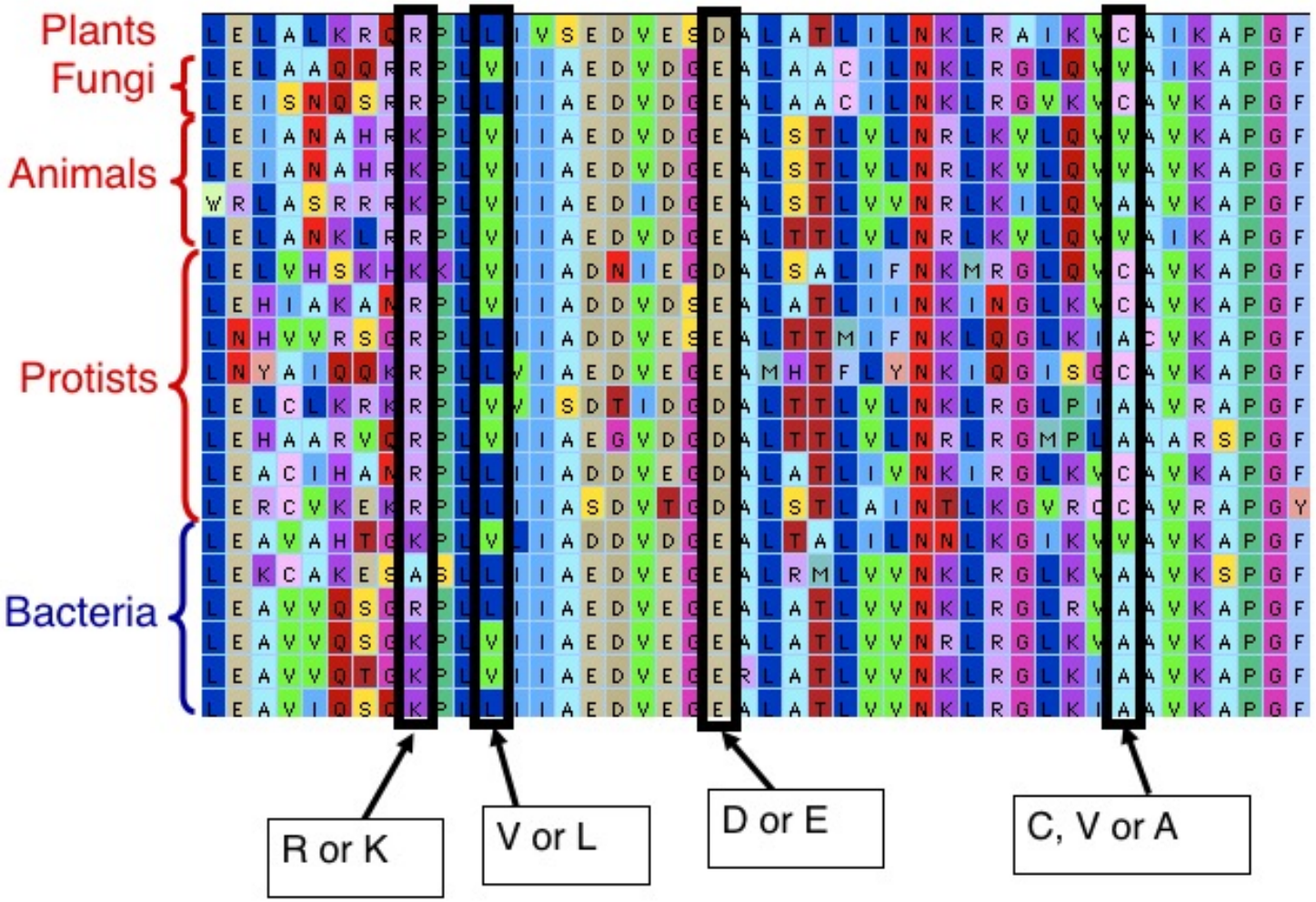


Microsporidia Example (Base Model)

```
$ iqtree -s microsporidia.phy -m LG+F+G
```



Amino Acid Frequency Variation



Frequency Mixture Models

- Similar problem and solution as for rates across sites: $\pi^{(1)}, \ldots, \pi^{(n)}$ i.i.d. w_c probability that frequency vector is $\pi^{(c)}$.

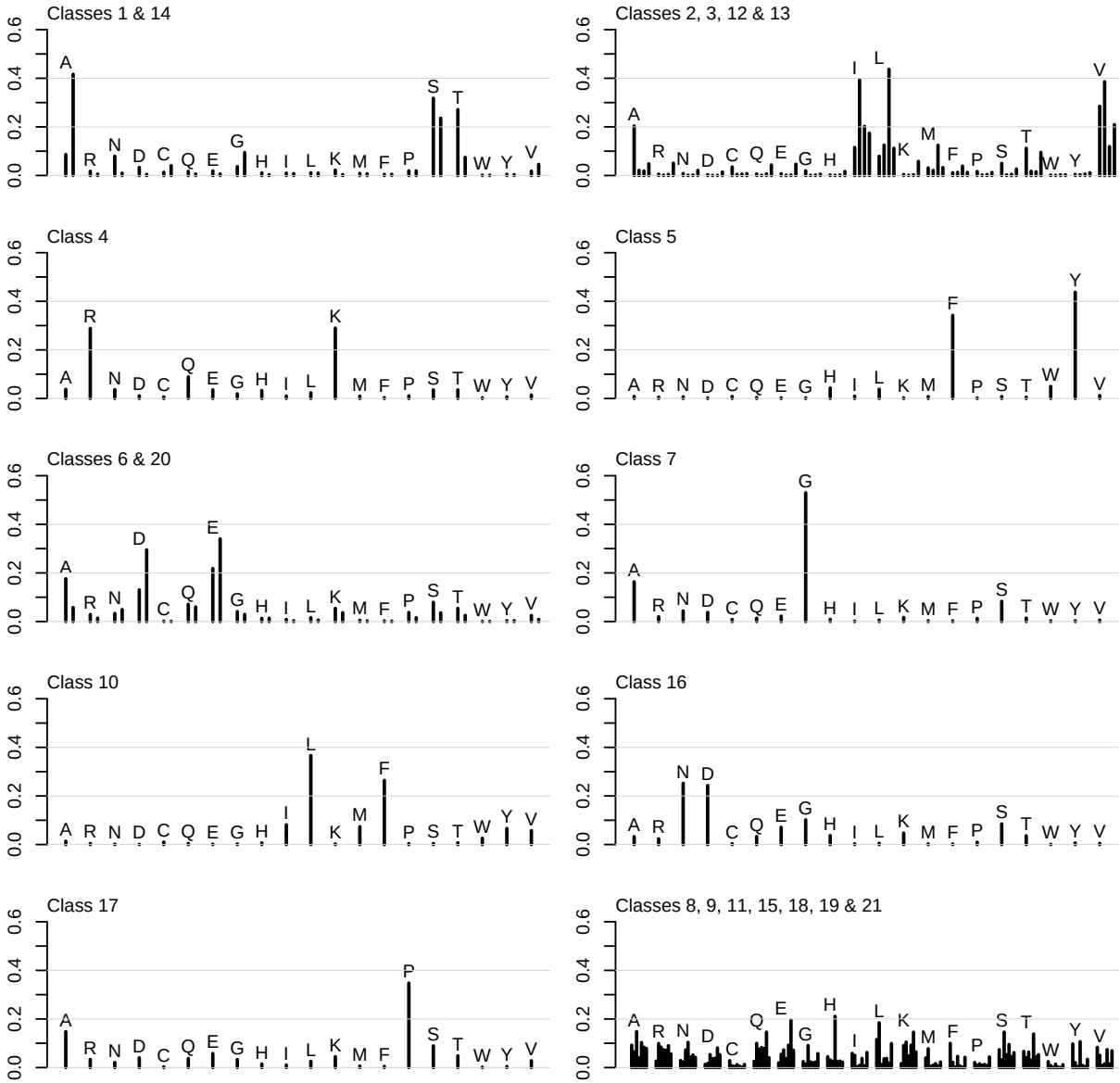
$$L(\tau, \boldsymbol{t}, \alpha, \boldsymbol{w}, \boldsymbol{\pi}) = \prod_h \sum_c w_c p(\boldsymbol{x}_h; \tau, \boldsymbol{t}, \alpha, \pi^{(c)})$$

- Each frequency vector, $\pi^{(c)}$, is 20-dimensional. ML estimation difficult
- Similar to Exchangeablility. Use fixed $\pi^{(c)}$ estimated from a large data base and fix throughout.

$$L(\tau, \boldsymbol{t}, \alpha, \boldsymbol{w}) = \prod_h \sum_c w_c p(\boldsymbol{x}_h; \tau, r_c \boldsymbol{t}, \alpha, \pi^{(c)})$$

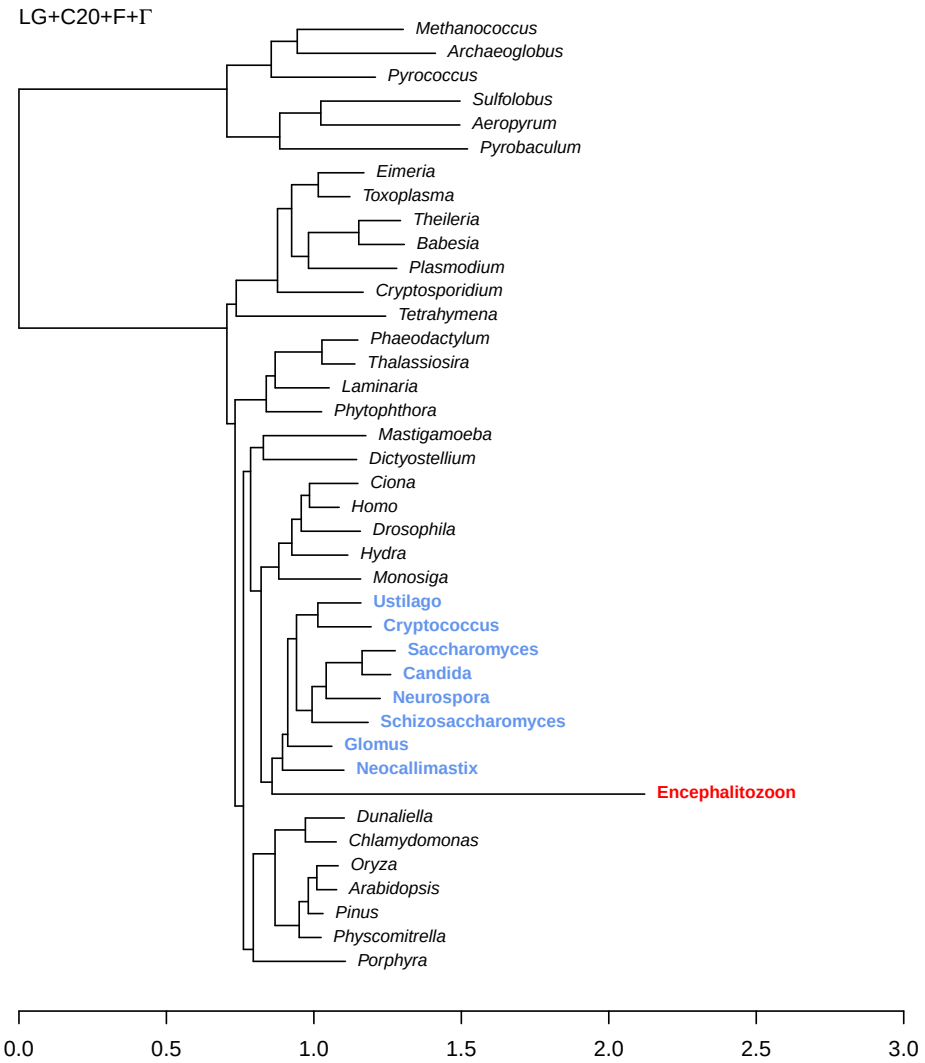
- C-series models (Le et al. 2012). C10, C20, ... C60

C20 Frequencies & LG Frequencies (Class 21)



Microsporidia Example (C20 Mixture)

```
$ iqtree -s microsporidia.phy -m LG+C20+F+G
```



Frequency Mixture Models - Frequency Class Estimation

$$L(\tau, \boldsymbol{t}, \alpha, \boldsymbol{w}, \boldsymbol{\pi}) = \prod_h \sum_c w_c p(\boldsymbol{x}_h; \tau, r_c \boldsymbol{t}, \alpha, \pi^{(c)})$$

- Estimation using data at hand?
- C classes $\Longrightarrow$ $C * 19 + C - 1$ additional parameters
- Numerical derivative approximations required $\Longrightarrow$ Repeated pruning algorithm applications
- ML estimation infeasible in practice

Composite Likelihood

- Setting: $P(\text{Full Data}; \theta)$ difficult to calculate or maximize
- Events E_k can be found where $P(E_k; \theta)$ is easily calculated

$$L_C(\theta) = \prod_k P(E_k; \theta)^{w_k}$$

- Each $P(E_k; \theta)$ is a partial likelihood.
- Often produces consistent estimation

- Composite likelihoods implicit in phylogenetics.
- Model for full data: Markov chain of sequences not sites
- E_k event x_k is observed at site k

Frequency Setting:

- $p(x|\pi^{(c)})$ is difficult calculate or maximize
- Let E_k be event Taxa k has amino acid x_k:
  $P(E_k|\pi^{(c)}) = \pi^{(c)}_{x_k}$ & composite likelihood contribution becomes

$$\prod_k \pi^{(c)}_{x_k}$$

Composite Likelihood for Frequency Variation

- Replace $p(x|\pi^{(c)})$ with product of conditional marginals, $\prod_s \pi^{(c)}_{x_s}$
- Site likelihood using marginals for single taxa:

$$\sum_c w_c \prod_s \pi^{(c)}_{x_s}$$

- Maximize

$$\sum_h \log[\sum_c w_c \prod_s \pi^{(c)}_{x_{hs}}]$$

Software available at
https://www.mathstat.dal.ca/ tsusko/software.html
under Susko, Lincker & Roger (2018)

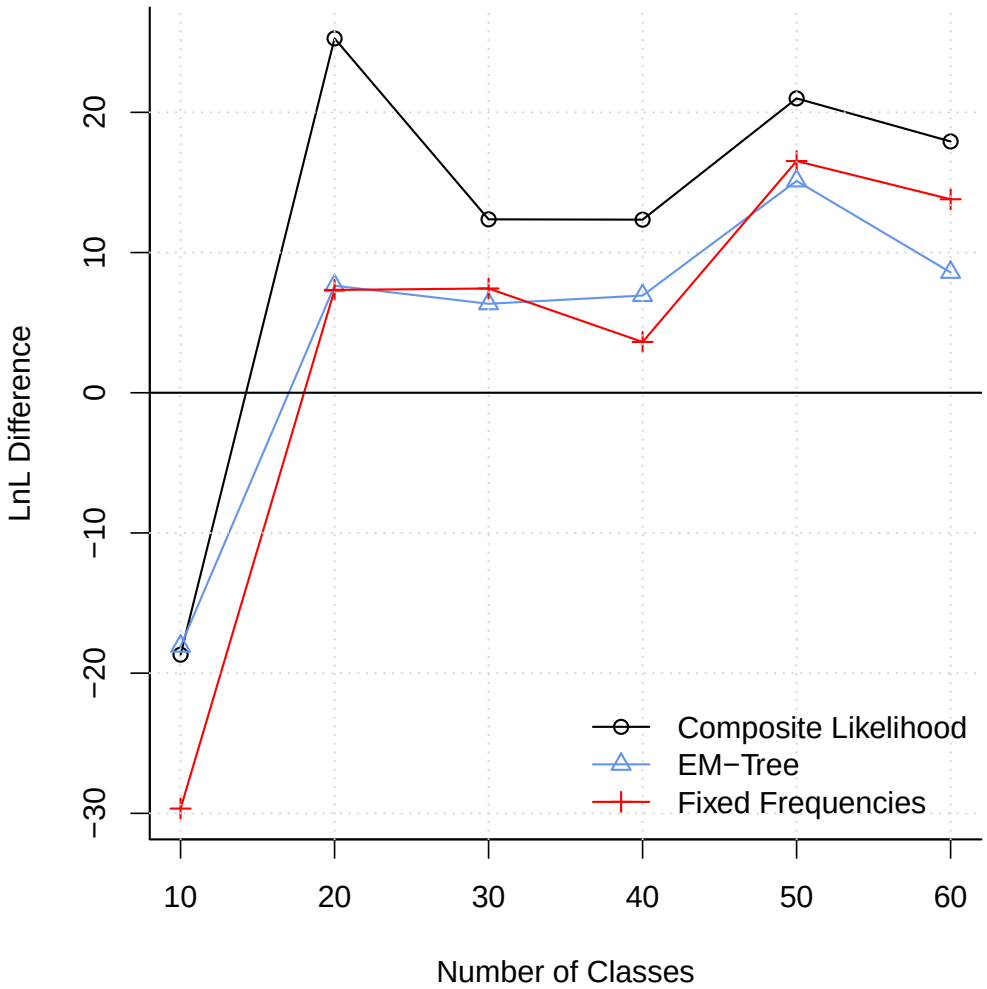
```
$ mammal -s microsporidia.phy
         -t microsporidia.phy.treefile -c 20
```

Creates a nexus file, `esmodel.nex` that can be used with iqtree

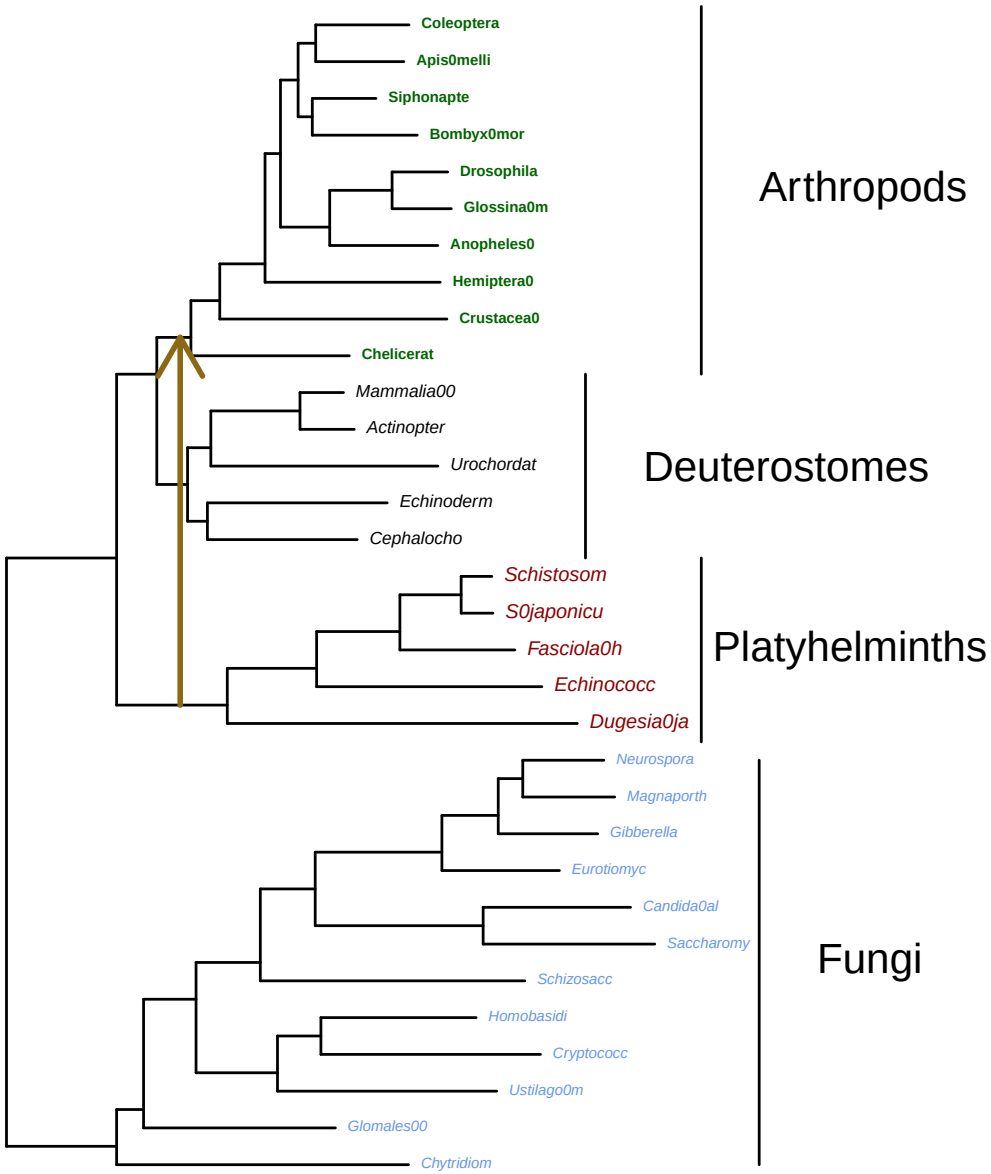
```
$ iqtree -s microsporidia.phy
         -mdef esmodel.nex -m LG+ESmodel+G
```

Microsporidia - Likelihood Improvement

△LnL (Correct Tree - Default Tree)

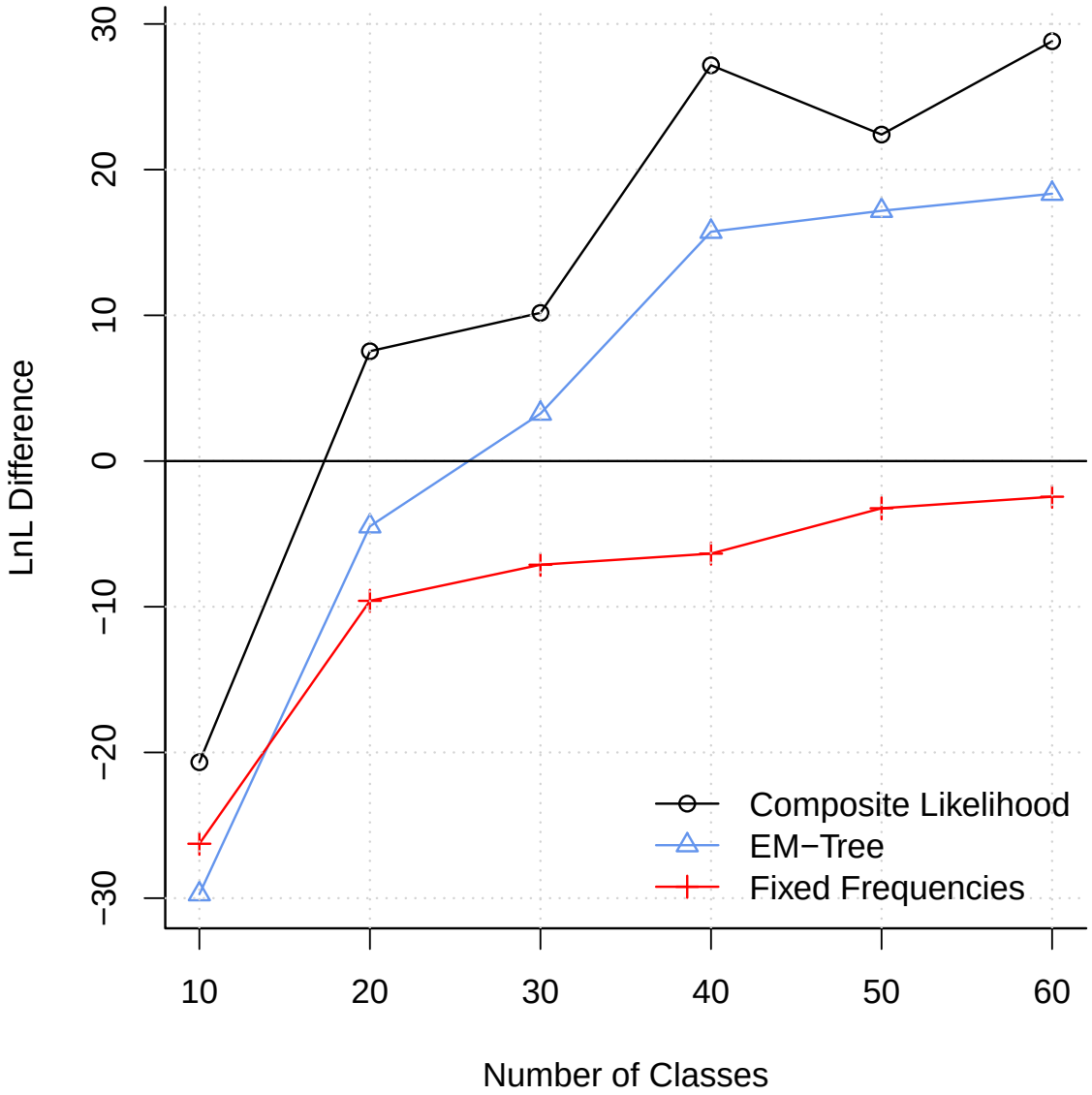


Platyhelminths Example

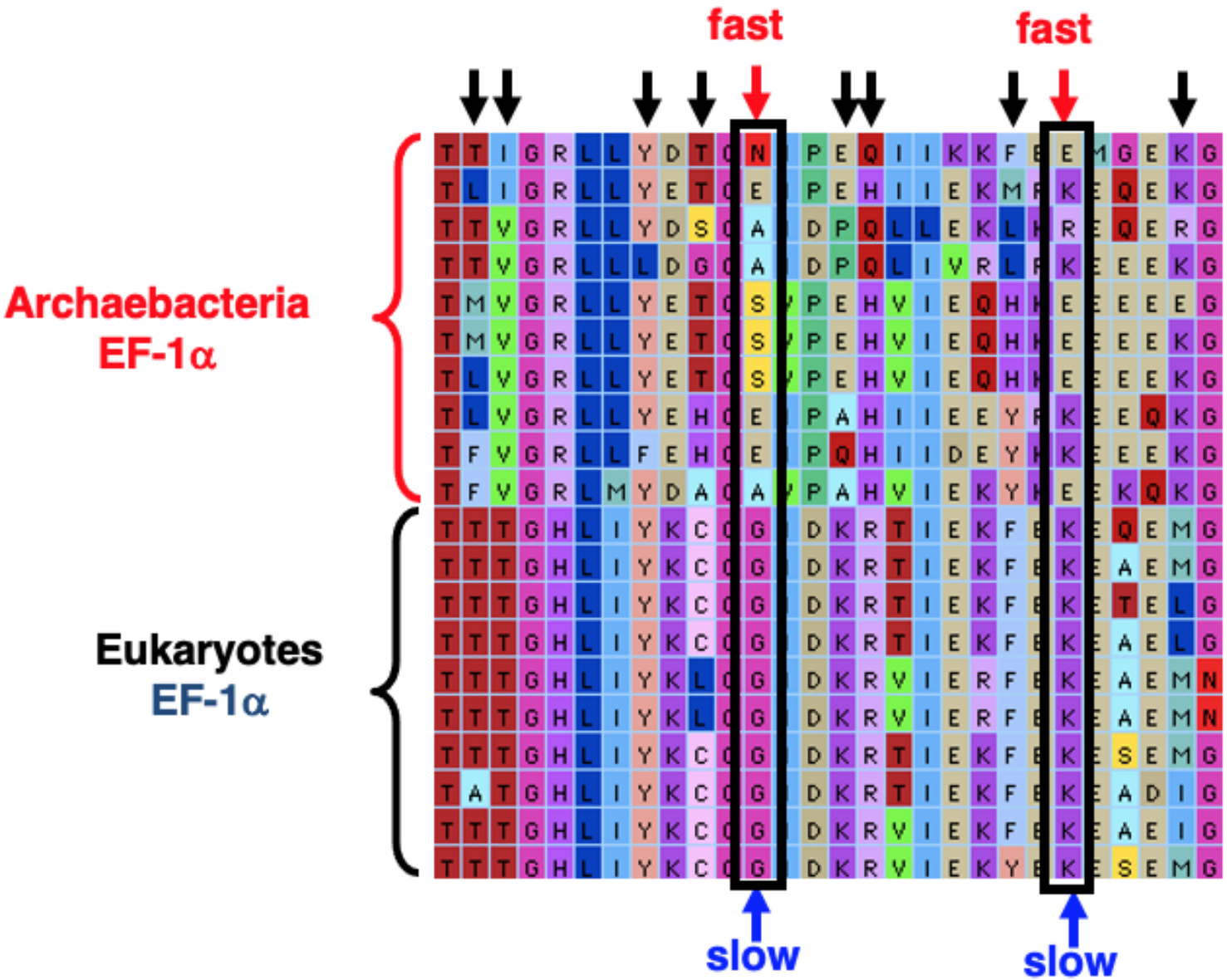


Likelihood Improvement

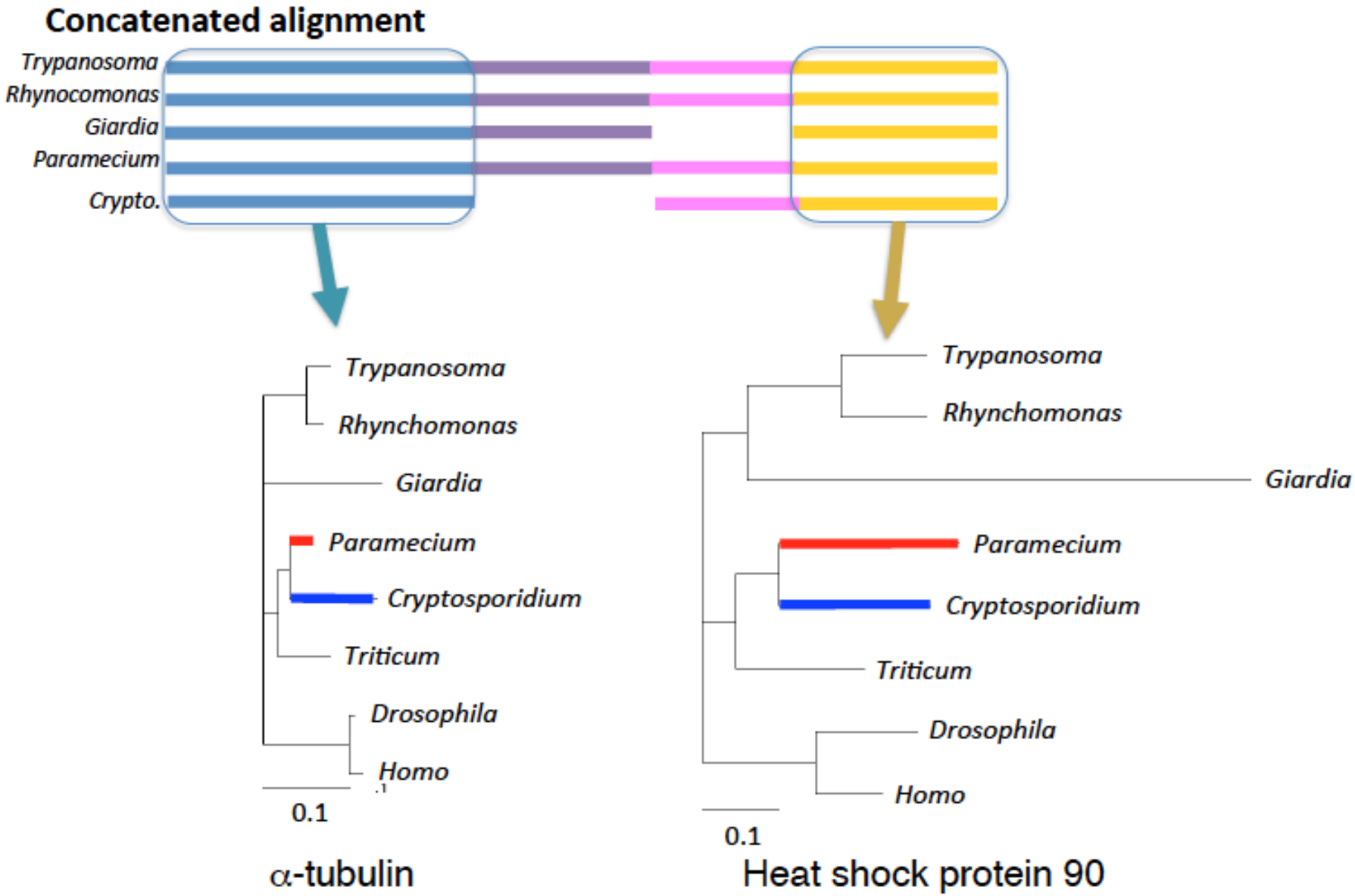
△LnL (Correct Tree - Default Tree)



Heterotachy (Rate Variation over Sites and Trees)



Heterotachy (Gene-wise)



- Heterotachy over Genes
  - Unlinked Branch Length Model (UBL): Each gene has its own set of edge-lengths
  - Linked Branch Length Model (LBL): Single different rate multipliers for each gene
- Heterotachy over Sites: Free Rates Model: Each site has its own set of edge-lengths.
  - Mixture model: $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$ i.i.d. w_c probability of edge-length class t_c (IQ-TREE: `LG+F+H4` in place of `LG+F+G4`)
- UBL & LBL models
  - Estimate $\boldsymbol{t}_g$ or r_g separately for each gene g
  - No mixture ($n \approx 300$)

Rates Across Sites (RAS) and Linked Branch Length (LBL)

Rates Vary over Genes

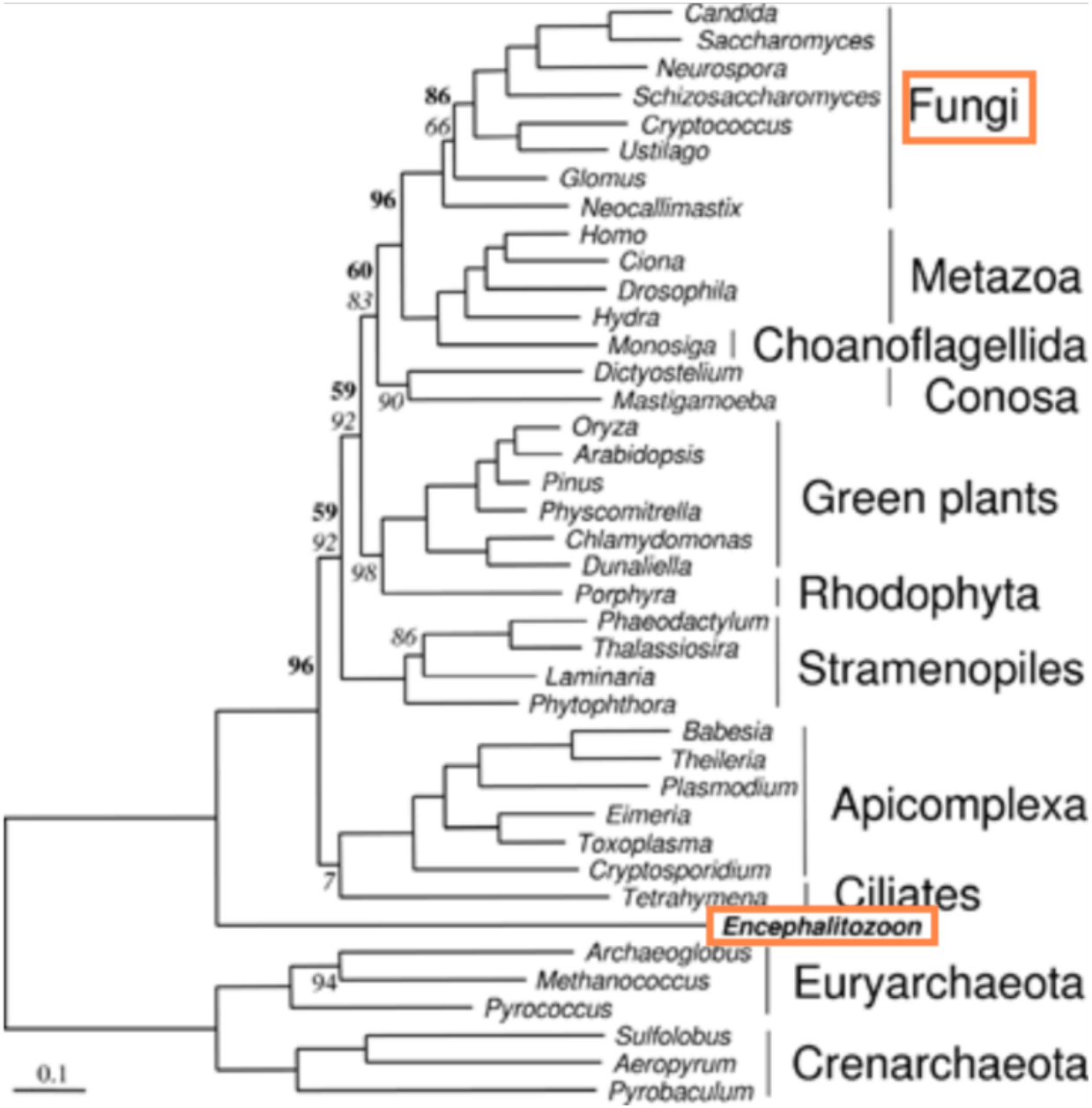
Gene 1	Gene 2
TNKQE	TGHLI
LEKAE	TGHLI
TARTE	TGHLI
TAKAE	TGHLI
MSEAE...	TGHLI...
MSKAE	TGHLI
LSKSE	TGHLI
LEKAD	TGHLI
FEKAE	TGHLI

Permutation (Vary within)

Gene 1	Gene 2
TTNGK	HQLEI
LTEGK	HALEI
TTAGR	HTLEI
TTAGK	HALEI
MTSGE...	HALEI...
MTSGK	HALEI
LTSGK	HSLEI
LTEGK	HALDI
FTEGK	HALEI

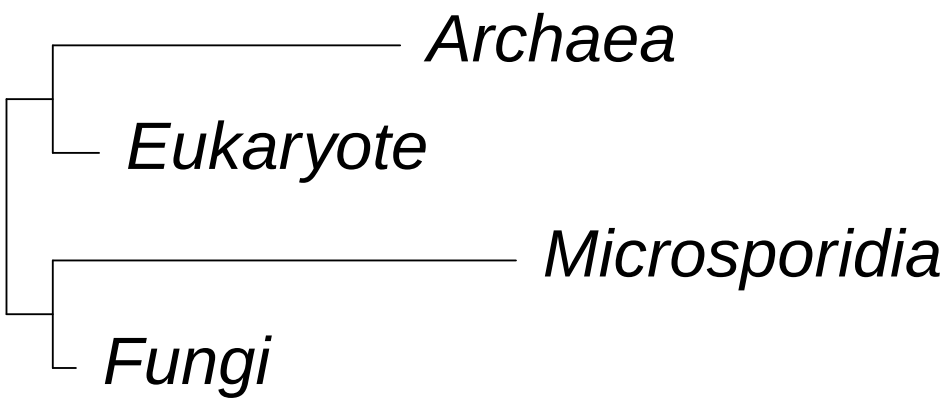
- RAS model: r_i, site i
- Order doesn't matter for RAS. Same LnL for both.
- LBL model: r_i, site i. r_i constant for gene g
- LBL model: Richer r_i variation

Microsprodia Example

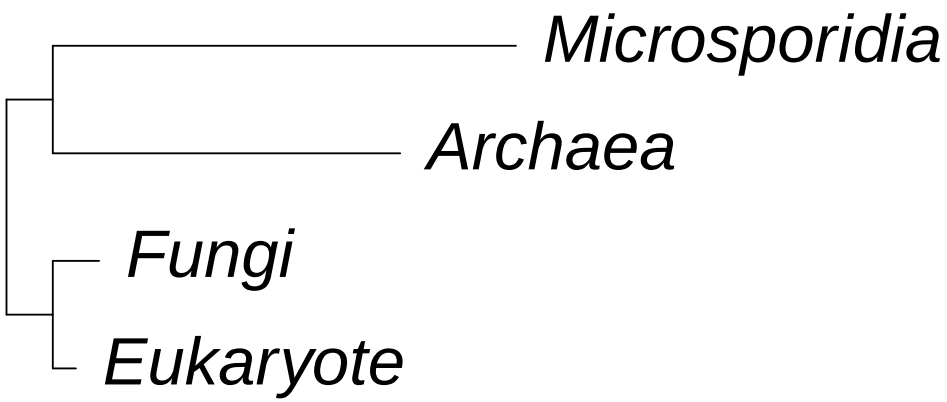


Microsprodia Example (Long Branch Attraction)

Correct Tree



LBA Tree

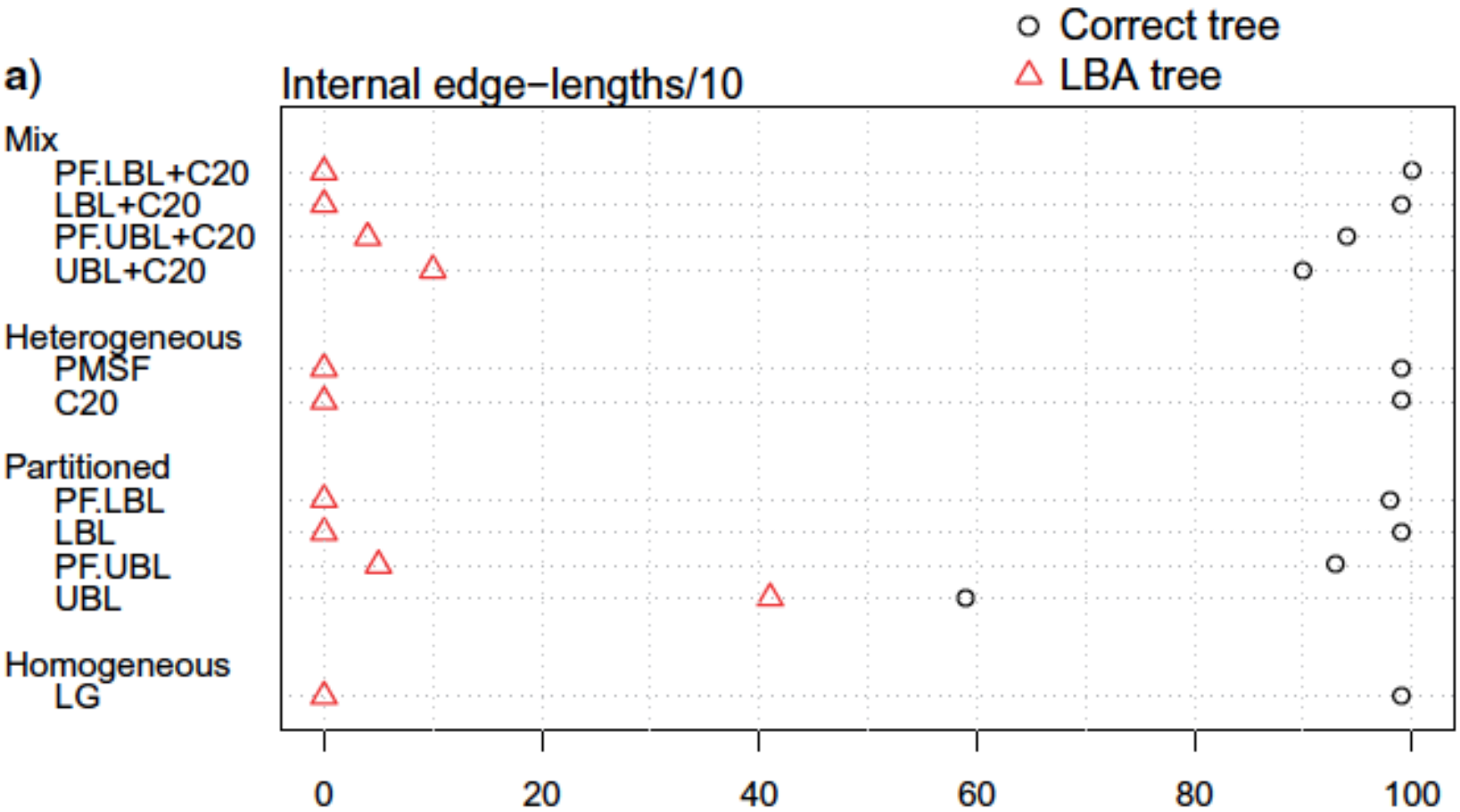


Simulation Comparison

Wang et al. Syst Biol (2019). 68:1003

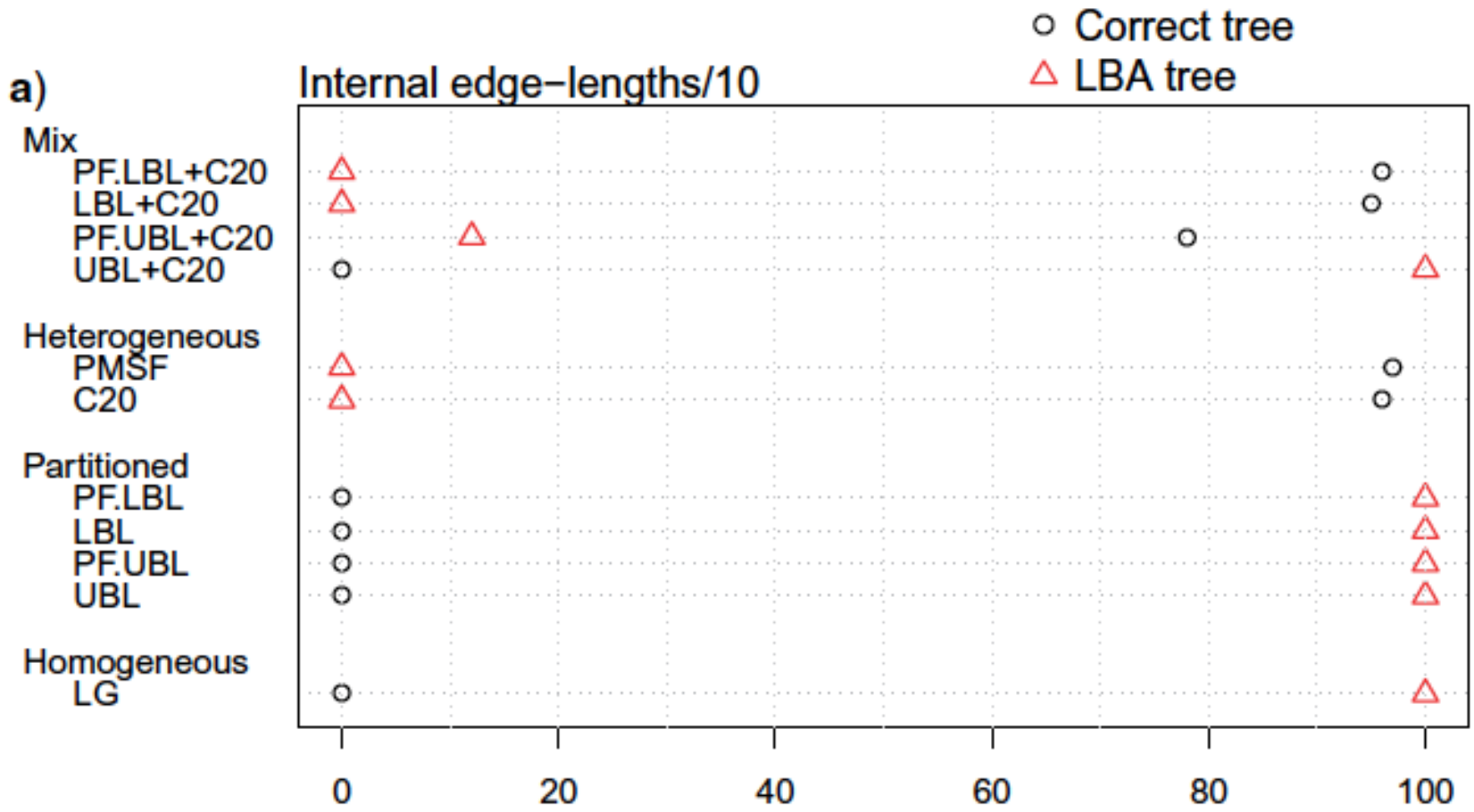
- Empirical-based simulation study
- Extracted estimated 4-taxon gene trees from Microsporidia. $\Longrightarrow$ UBL model
- Compared fitted partitioned models and frequency mixtures
- Included results for PartitionFinder (Lanfear et al. (2012). 29:1695)
- Included results for PMSF (single set of frequencies/site)
- All estimation methods included `+F+G`
  - All methods available in IQTREE

LG+F+Gamma+UBL Simulating Model



- Almost all methods do well at estimating correct tree
- UBL is the correct model??
- PF.UBL does well

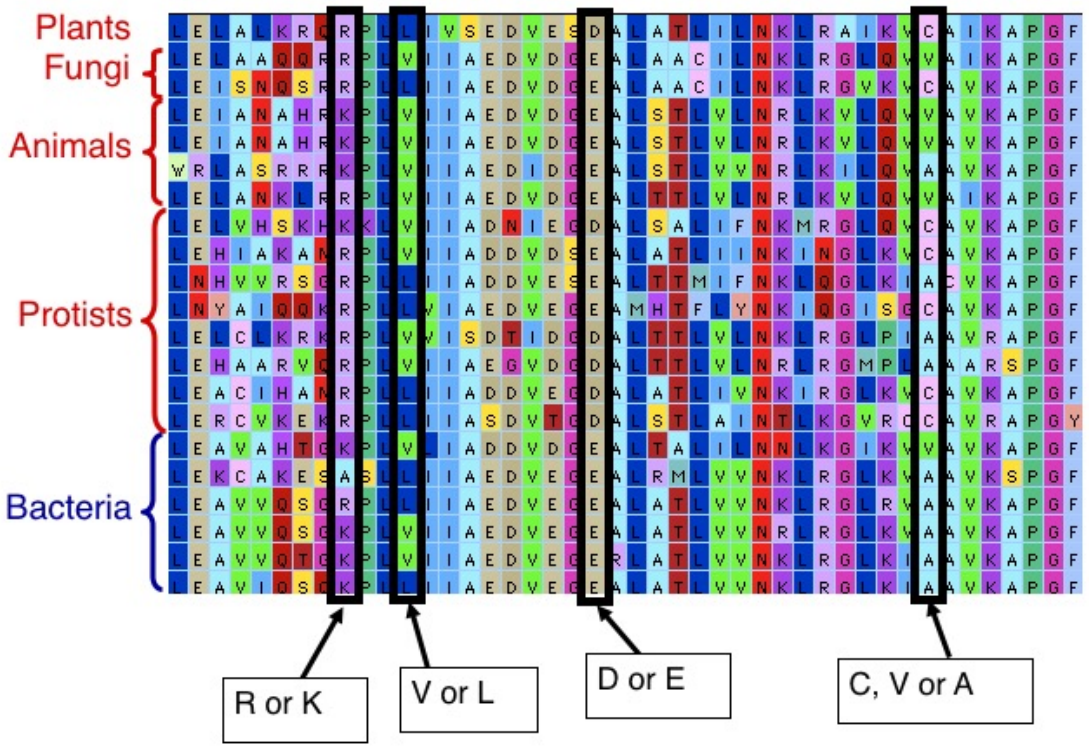
LG+C20+F+Gamma+UBL Simulating Model



- Frequency mixture do well
- Partition models do very poorly
- UBL+C20 is the correct model??
- PF.UBL+C20 does well

Why is LBA so prevalent?

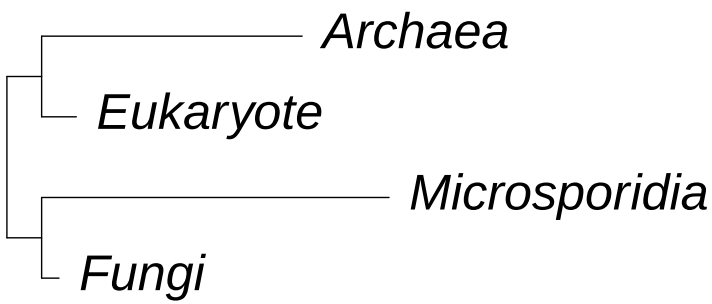
Evolution of chaperonin 60 over ~1.5 billion years



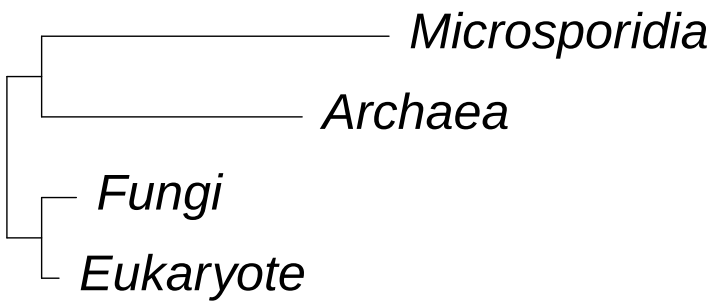
- Rate not small at ‘V or L’/‘D or E’ site
- Single matrix models expect more amino acids
  $\Longrightarrow$ under-estimate number of substitutions
- Greater underestimation for larger edge-lengths than shorter

Long Branch Attraction (Single Matrix Model)

Correct Tree

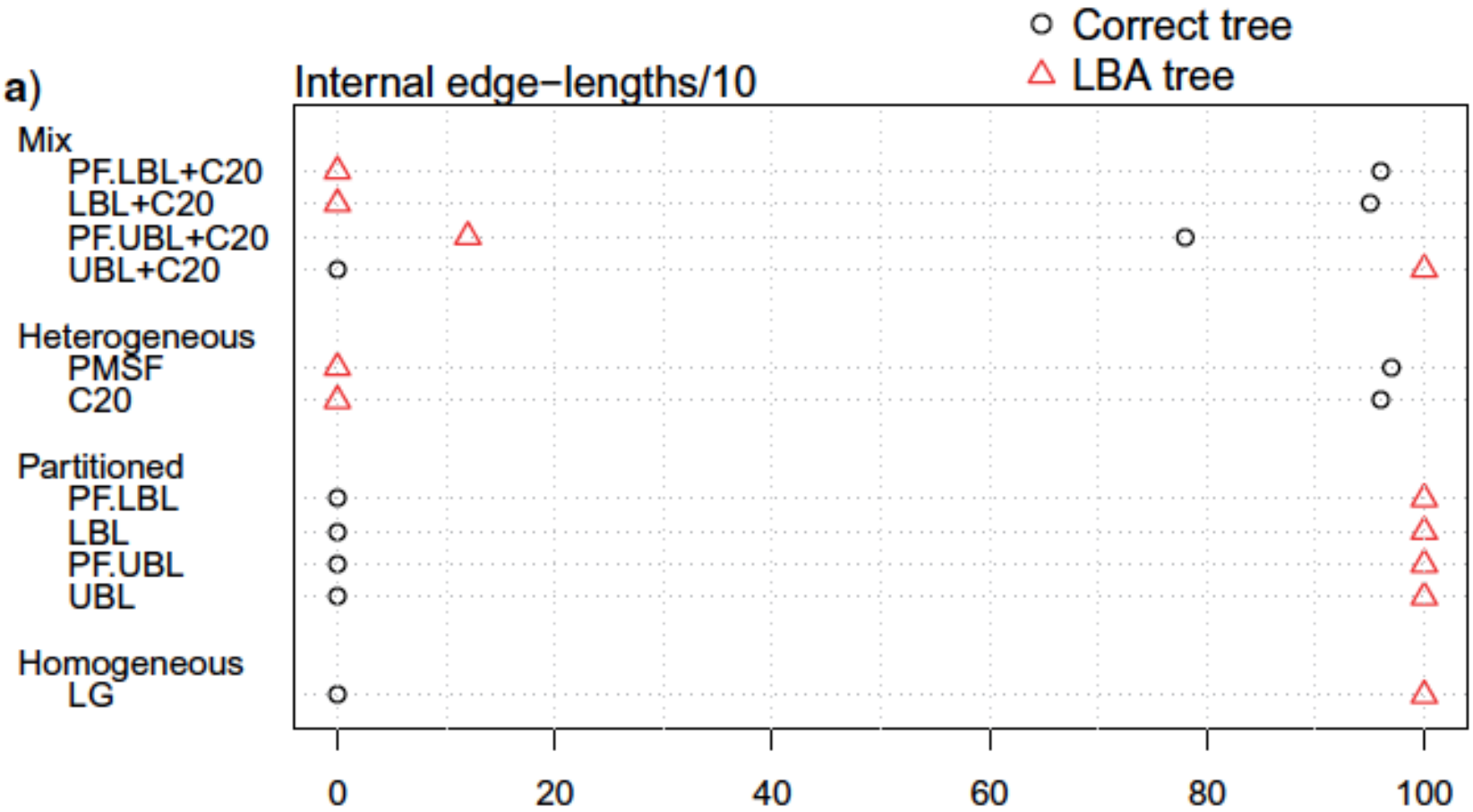


LBA Tree



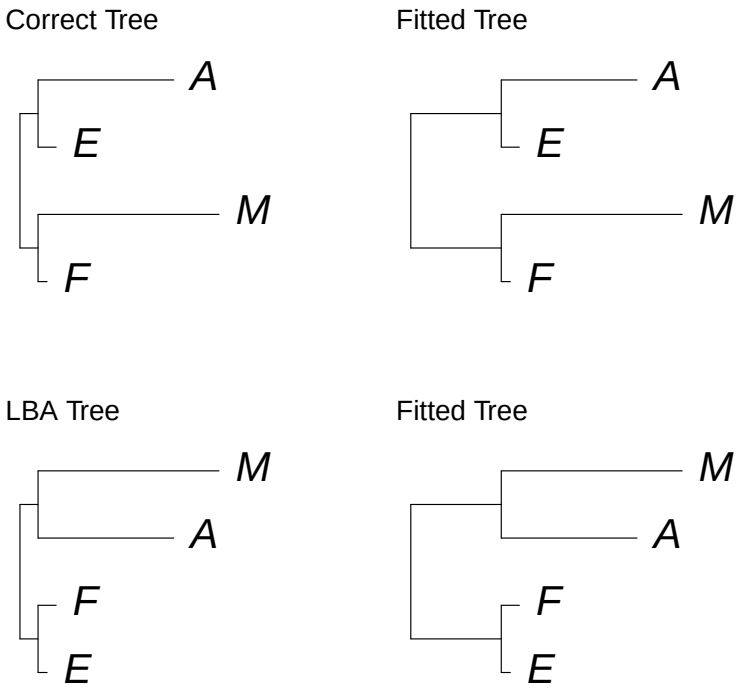
- Distance between *M* and *A* inferred shorter.
- Distance between *F* and *E* roughly same
- LBA tree accomodates

Long Branch Attraction - No Misspecification



- C20+UBL is the correct model

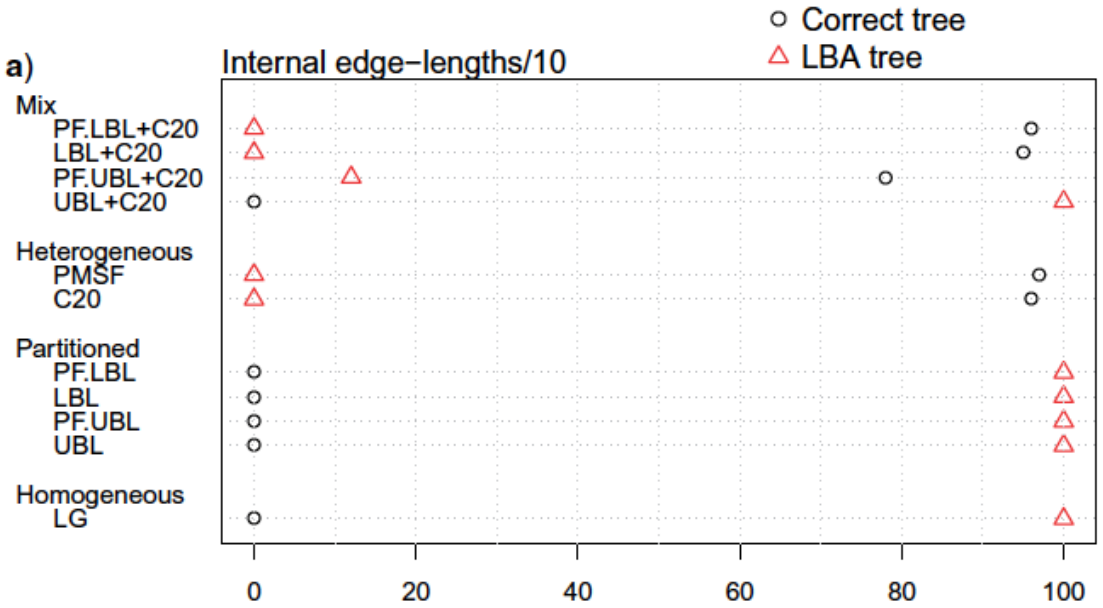
Long Branch Attraction - No Misspecification



- Similar to reason for ASTRAL inconsistency
- Need small distance between F and E
- Small middle edge-lengths for Correct Tree
- Large middle edge-lengths possible for LBA Tree

- Greater model flexibility for LBA Tree $\Longrightarrow$
  - Bias $E[l_g(\text{LBA}) - l_g(\text{Correct})] > 0$
  - Bias should go away with large n (Correct Model)

Long Branch Attraction - No Misspecification



- UBL maximizes separately over genes. $LnL = \sum_g l_g$, l_g max LnL.
- Slight bias $E[l_g(\text{LBA}) - l_g(\text{Correct})] = 0.2$
- Single gene $\text{Var}[l_g] = 1$ more important than bias
- $G = 133$ genes: Largest Average LnL wins

$$E[\text{ave}_g l_g(\text{LBA}) - \text{ave} l_g(\text{Corr})] = E[l_g(\text{LBA}) - l_g(\text{Corr})] = 0.2$$

so same slight bias. But $\text{Var}[\sum_g l_g/G] = \text{Var}[l_g]/G = 0.008$.

- Large G, bias more important than variance

Key Points

- Amino acid data minimizes problems with saturation
- Frequency and rate variation are important to adjust for
- Long branch attraction is a common problem.
- Partition Models
  - Linked branch models reasonable but Gamma adjusts to some degree
  - Unlinked branch lengths with PartitionFinder