Modelling prokaryote gene content

Matthew Spencer

Department of Mathematics and Statistics & Department of Molecular Biology and Biochemistry, Dalhousie University

Acknowledgements

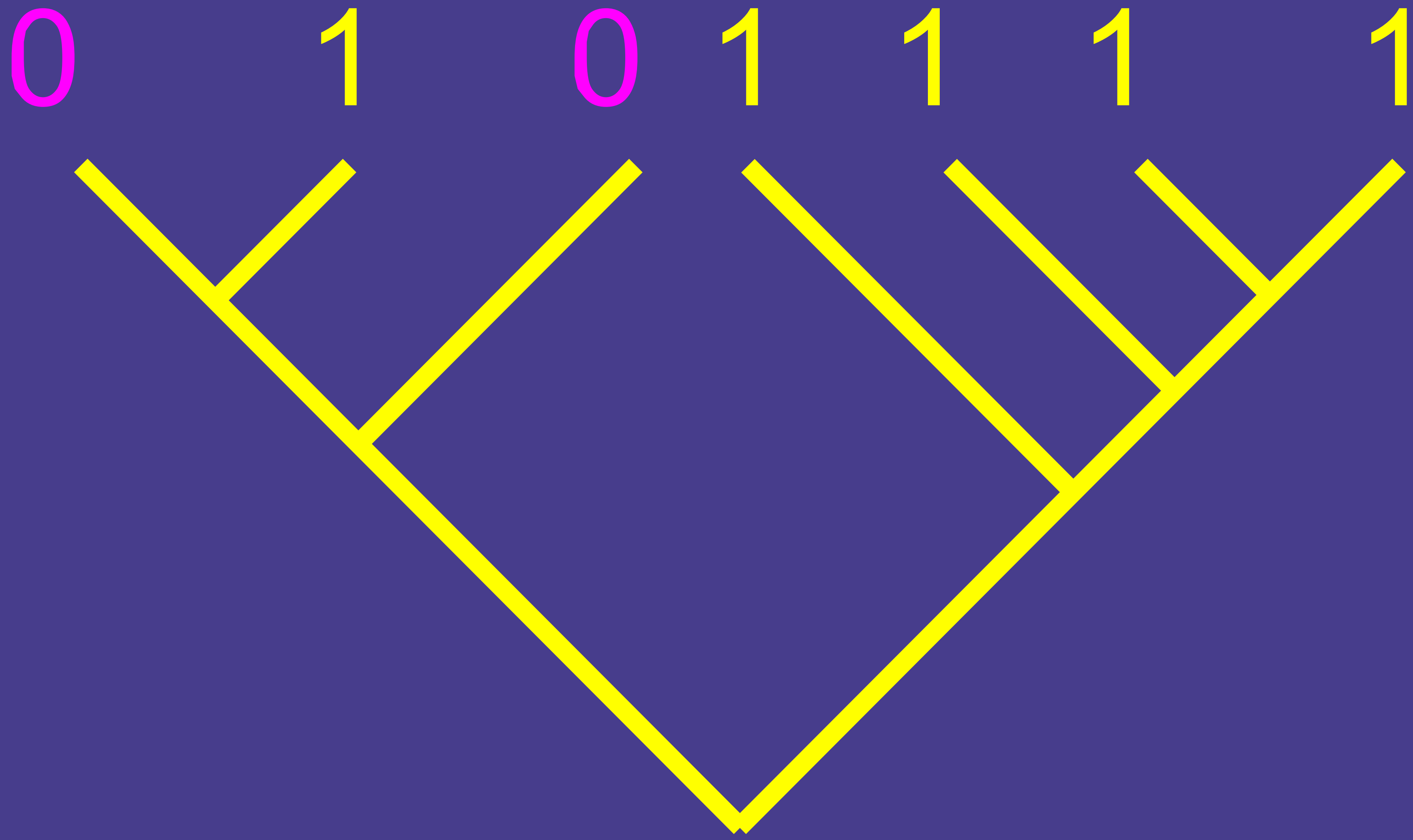
- Andrew Roger
- Ed Susko
- Dalhousie Statistical Evolutionary Bioinformatics group
- Genome Atlantic

Outline

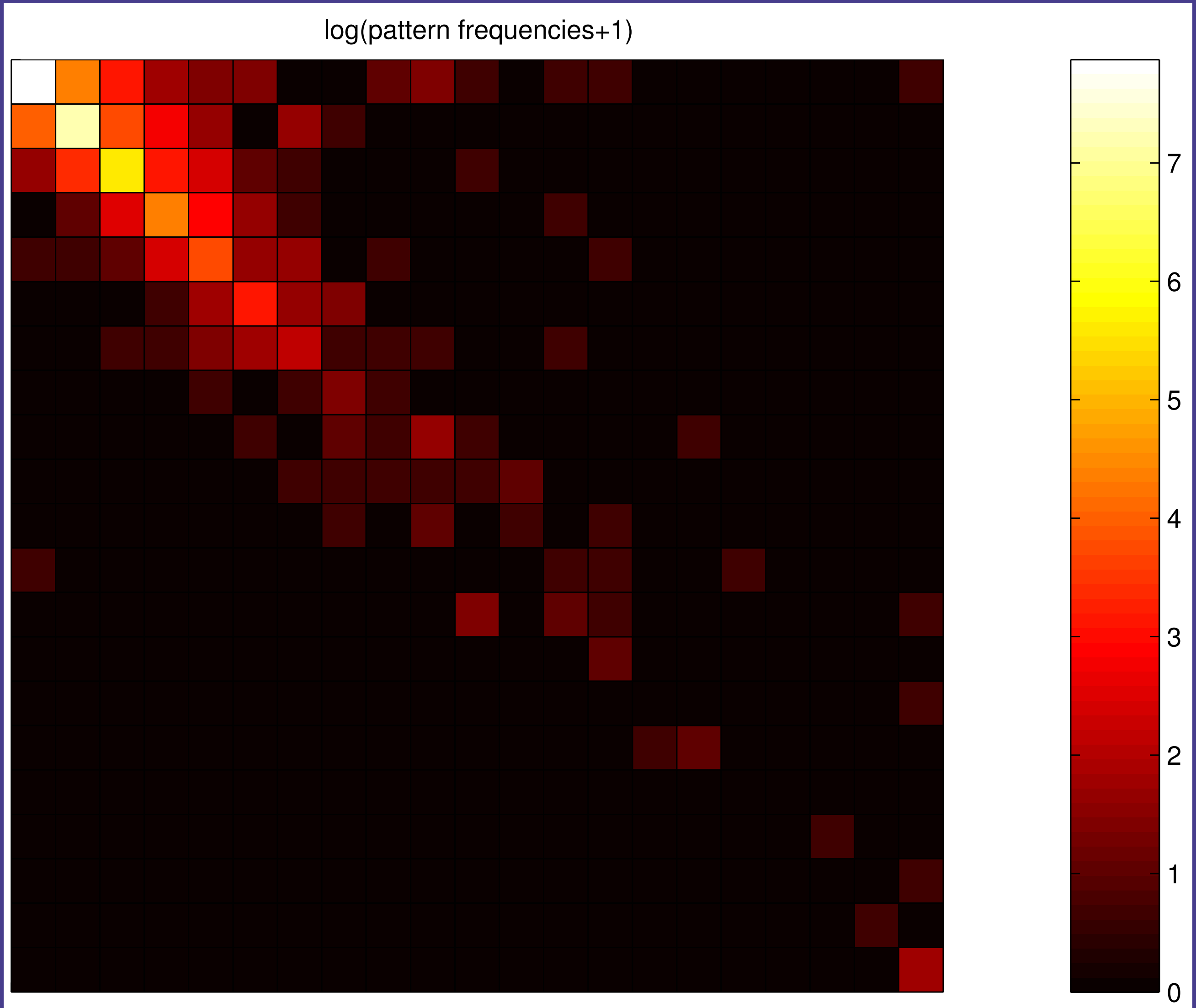
- Gene distributions: lateral transfer or multiple loss?
- Birth-death models vs. models with multi-gene events
- Lateral transfer rates
- ML distance phylogenies
- Residence times of genes

`http://www.mathstat.dal.ca/~matts/`

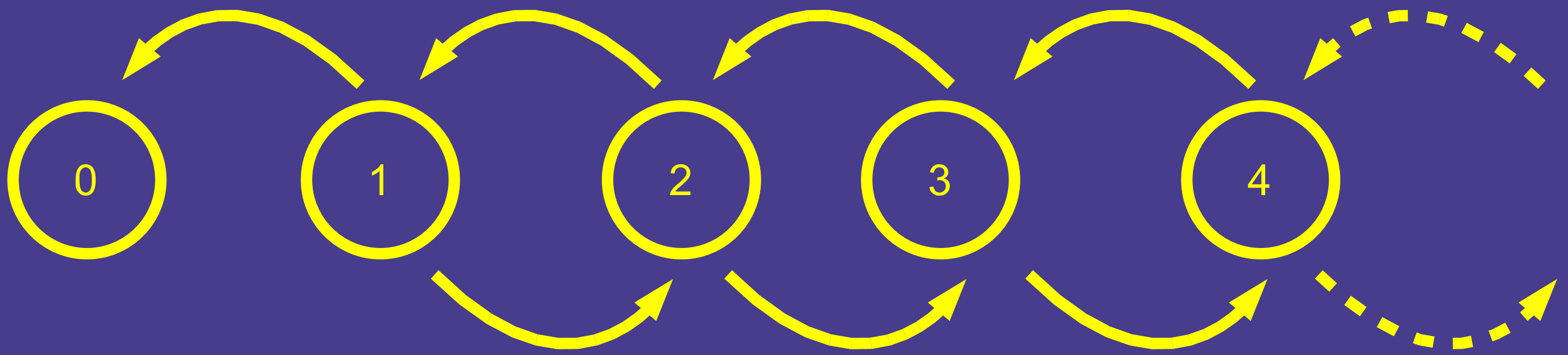
Lateral transfer or gene loss?



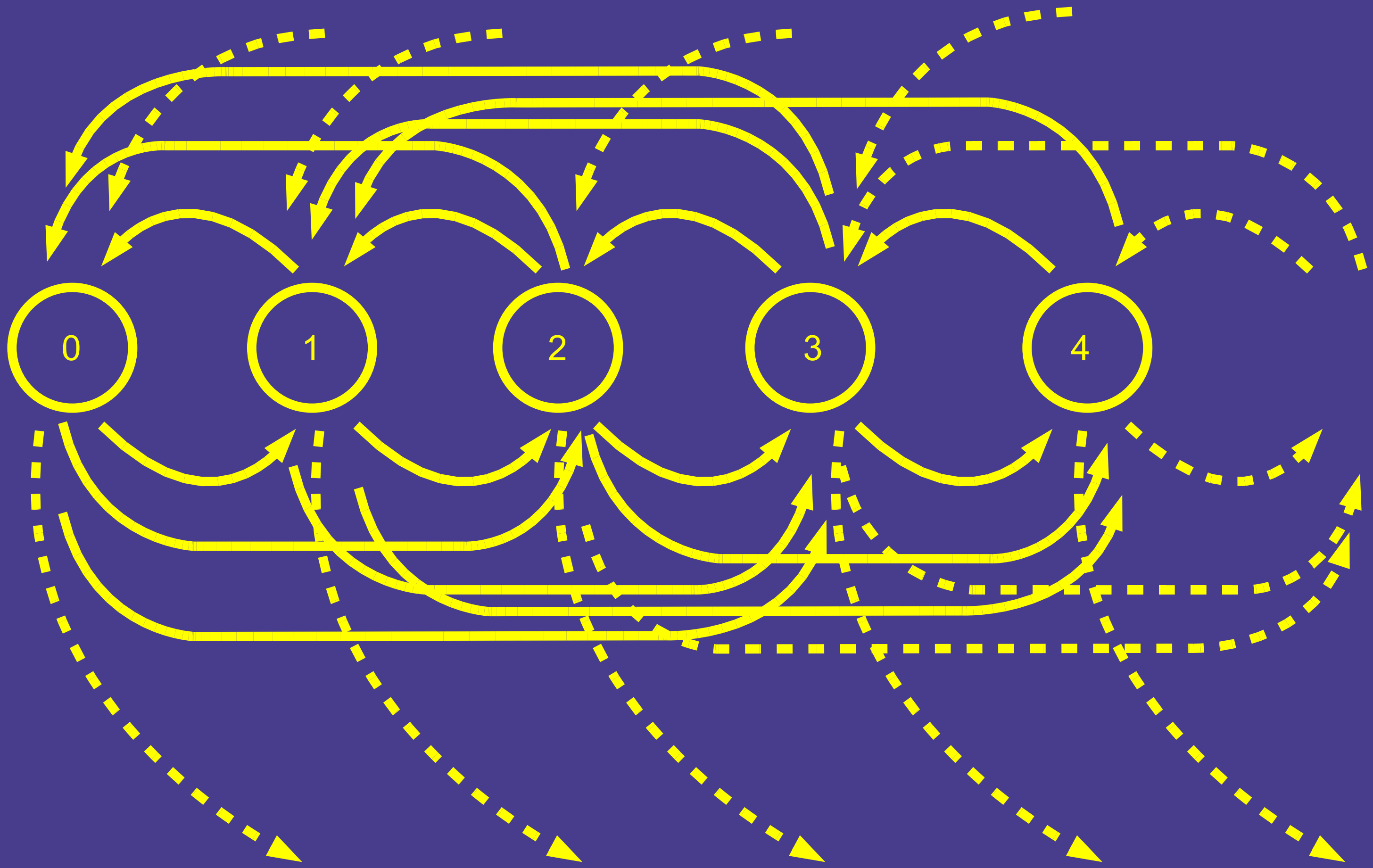
Gene family size data



Birth-death model



Models with multi-gene events



Assumptions

- Family independence
- Finite maximum number of genes in family
- Frequent rearrangements
- Lateral transfers come from outside the set of sampled organisms

Rate categories

- Deletions of single genes
- Gains of single genes
- Deletions of > 1 gene
- Gains of > 1 gene where the gain could be duplication
- Gains of more genes than could be duplicated
- Loss of entire gene family
- Transition from 0 to 1 members of family

Results

Log likelihoods:

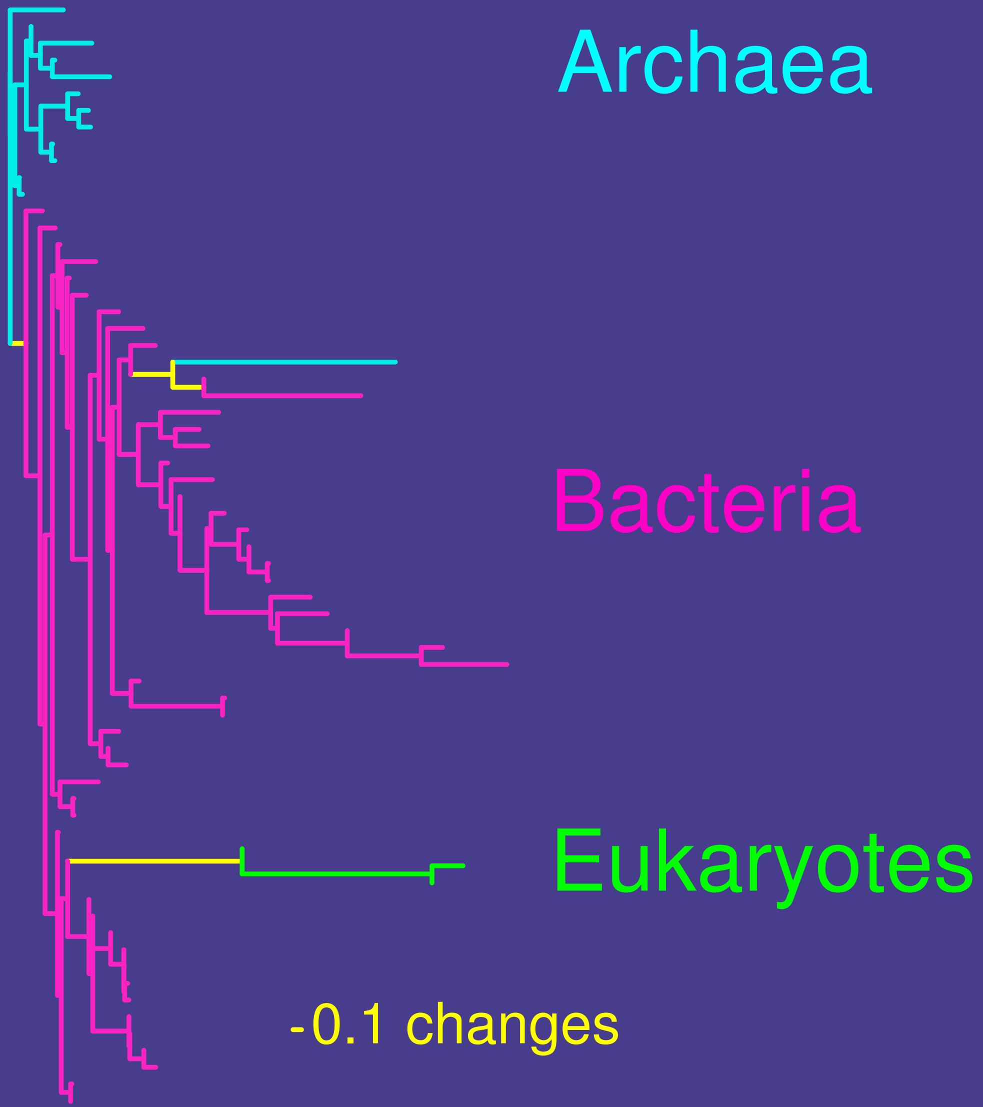
Species	blocks	birth-death
E. coli	-7.55×10^3	-7.89×10^3
A. fulgidus & B. subtilis	-9.13×10^3	-9.17×10^3

- Strongly prefer blocks model for both pairs
- Evidence for deletions and duplications of multiple genes

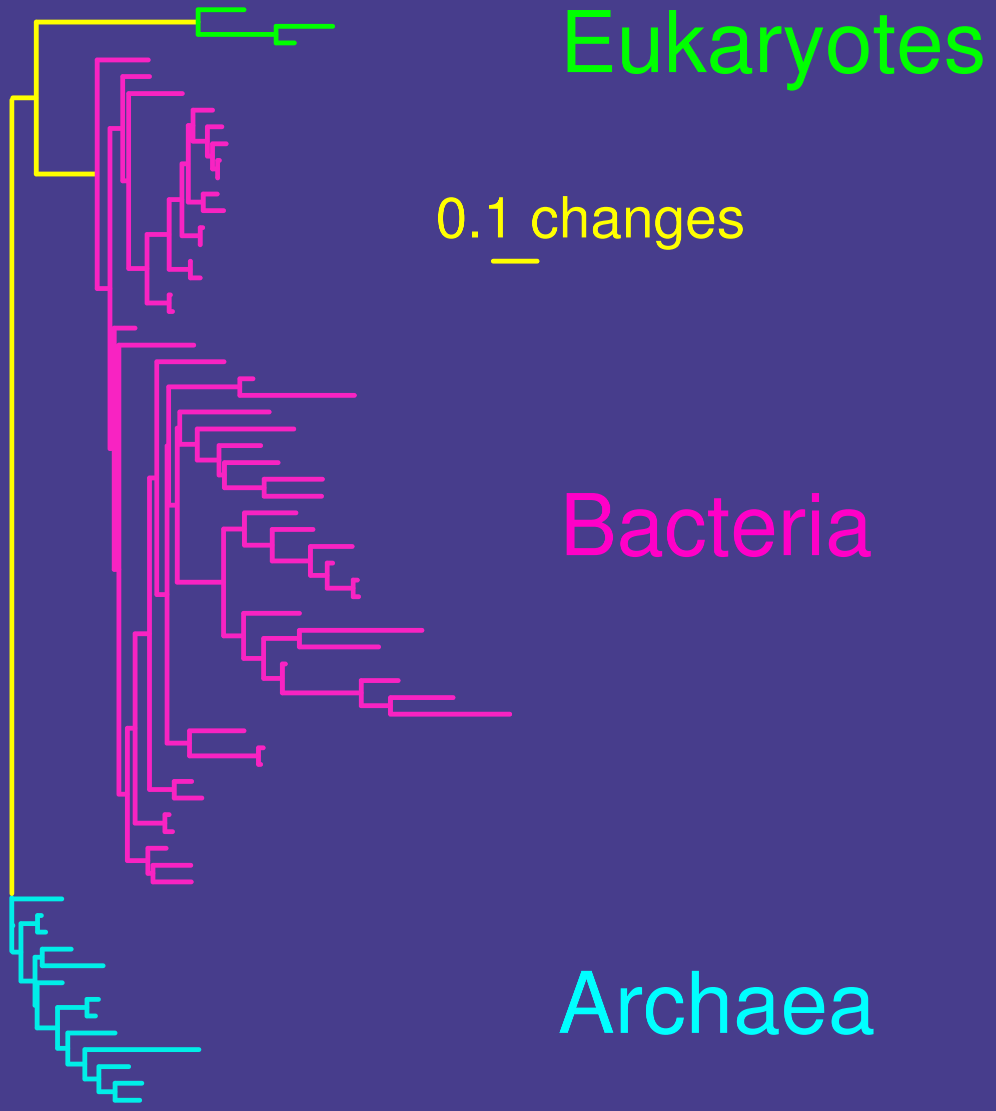
Lateral transfer rates?

Species	multiple genes	$0 \rightarrow 1$
E. coli	5.21×10^{-4}	0.27
A. fulgidus & B. subtilis	6.79×10^{-8}	0.40

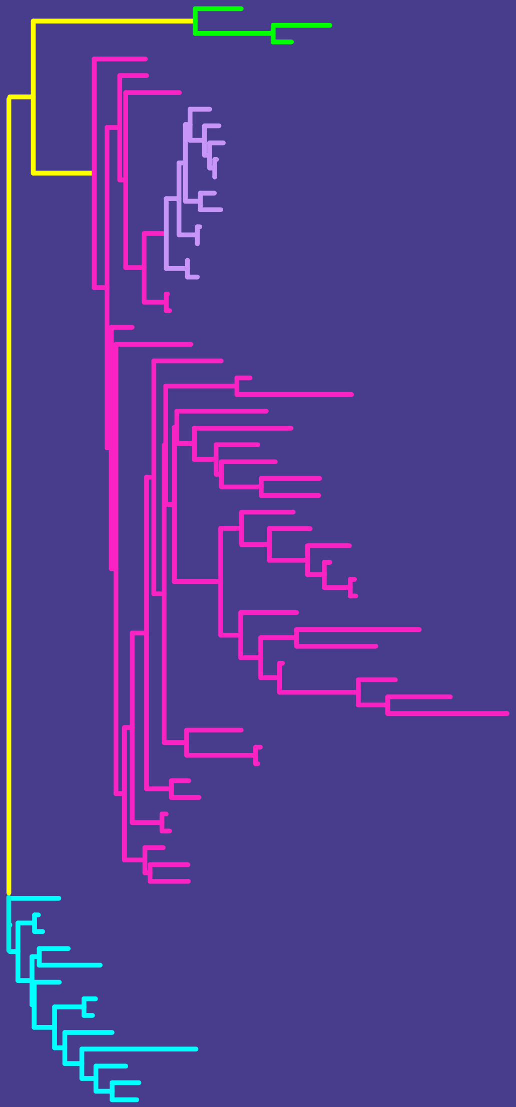
Birth-death phylogeny



Blocks phylogeny



Blocks phylogeny



Eukaryotes

Parasites/
endosymbionts

Bacteria

Archaea

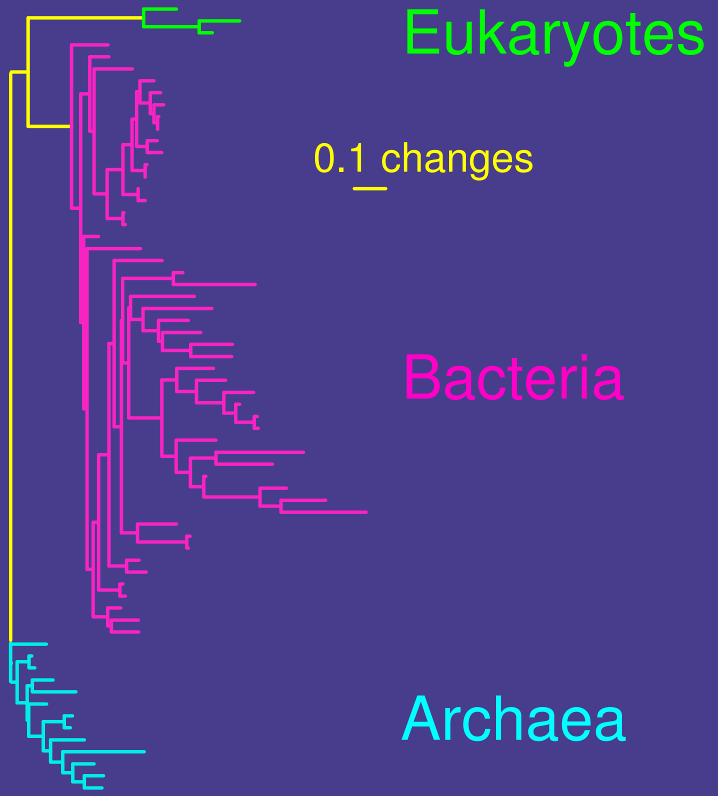
Residence times

- Expected time from origin of a gene (innovation, duplication or transfer) to loss from the genome
- = sum over all states i [(probability we enter state i as a new gene is created) $\times$ (expected time to lose a gene created in state i)]
- *E. coli*: mean 0.60, median 0.33 events
- *A. fulgidus/B. subtilis*: mean 0.48, median 0.34 events
- Between ancestors of bacteria and archaea: 0.19 events

Summary

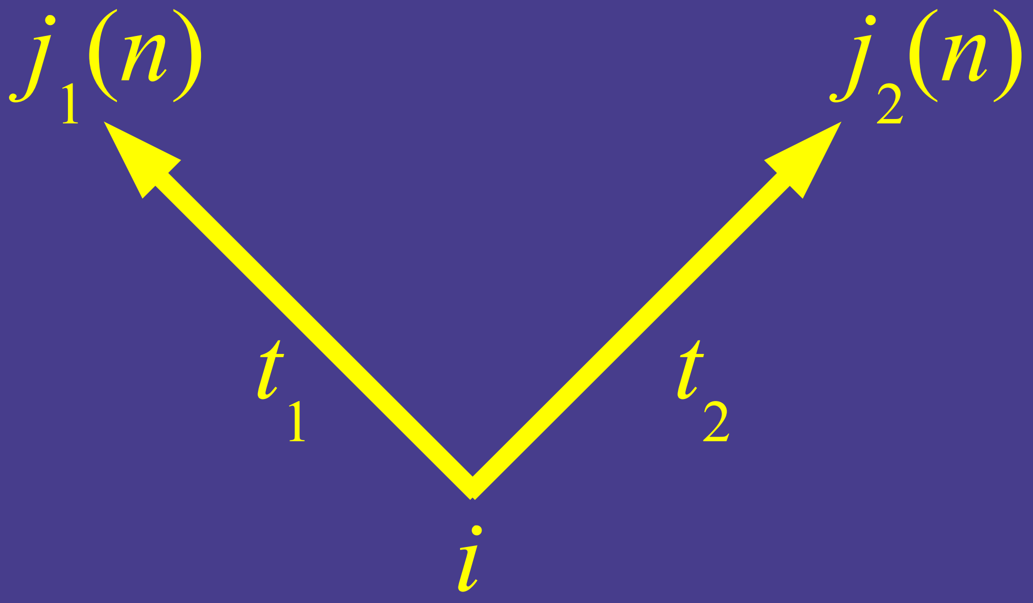
- Models that allow multi-gene events work better than birth-death models
- No evidence for frequent transfers of multiple genes from the same family
- May be a high rate of lateral transfers of single genes
- If we want to use single genes, we should focus on the ones with long residence times

The End



http://www.mathstat.dal.ca/~matts/

Likelihood calculation



$$L = \prod_{n=1}^{N} \sum_{i=0}^{k} \pi_i P\left(i, j_1(n) | t_1\right) P\left(i, j_2(n) | t_2\right)$$

Likelihood calculation

$$L = \prod_{n=1}^{N} \sum_{i=0}^{k} \pi_i P\left(i, j_1(n) | t_1\right) P\left(i, j_2(n) | t_2\right)$$

Likelihood calculation

$$L = \prod_{n=1}^{N} \sum_{i=0}^{k} \pi_i P\left(i, j_1(n) | t_1\right) P\left(i, j_2(n) | t_2\right)$$

- $P(i, j_.(n)|t_.)$ from exponential of rate matrix

Likelihood calculation

$$L = \prod_{n=1}^{N} \sum_{i=0}^{k} \pi_i P\left(i, j_1(n) | t_1\right) P\left(i, j_2(n) | t_2\right)$$

- $P(i, j_.(n)|t_.)$ from exponential of rate matrix
- π_i from stationary probabilities of rate matrix

Likelihood calculation

$$L = \prod_{n=1}^{N} \sum_{i=0}^{k} \pi_i P\left(i, j_1(n) | t_1\right) P\left(i, j_2(n) | t_2\right)$$

- $P(i, j_.(n)|t_.)$ from exponential of rate matrix
- π_i from stationary probabilities of rate matrix
- Sum over possible root states i

Likelihood calculation

$$L = \prod_{n=1}^{N} \sum_{i=0}^{k} \pi_i P\left(i, j_1(n) | t_1\right) P\left(i, j_2(n) | t_2\right)$$

- $P(i, j_.(n)|t_.)$ from exponential of rate matrix
- π_i from stationary probabilities of rate matrix
- Sum over possible root states i
- Product over all gene families n

Residence times

$$E(r) = \sum_{i=0}^{k} \beta_i r_i$$

where β_i is the probability that we enter state i as a gene appears in the genome, and r_i is the expected time until a gene is deleted, given that we were in state i when it appeared in the genome.

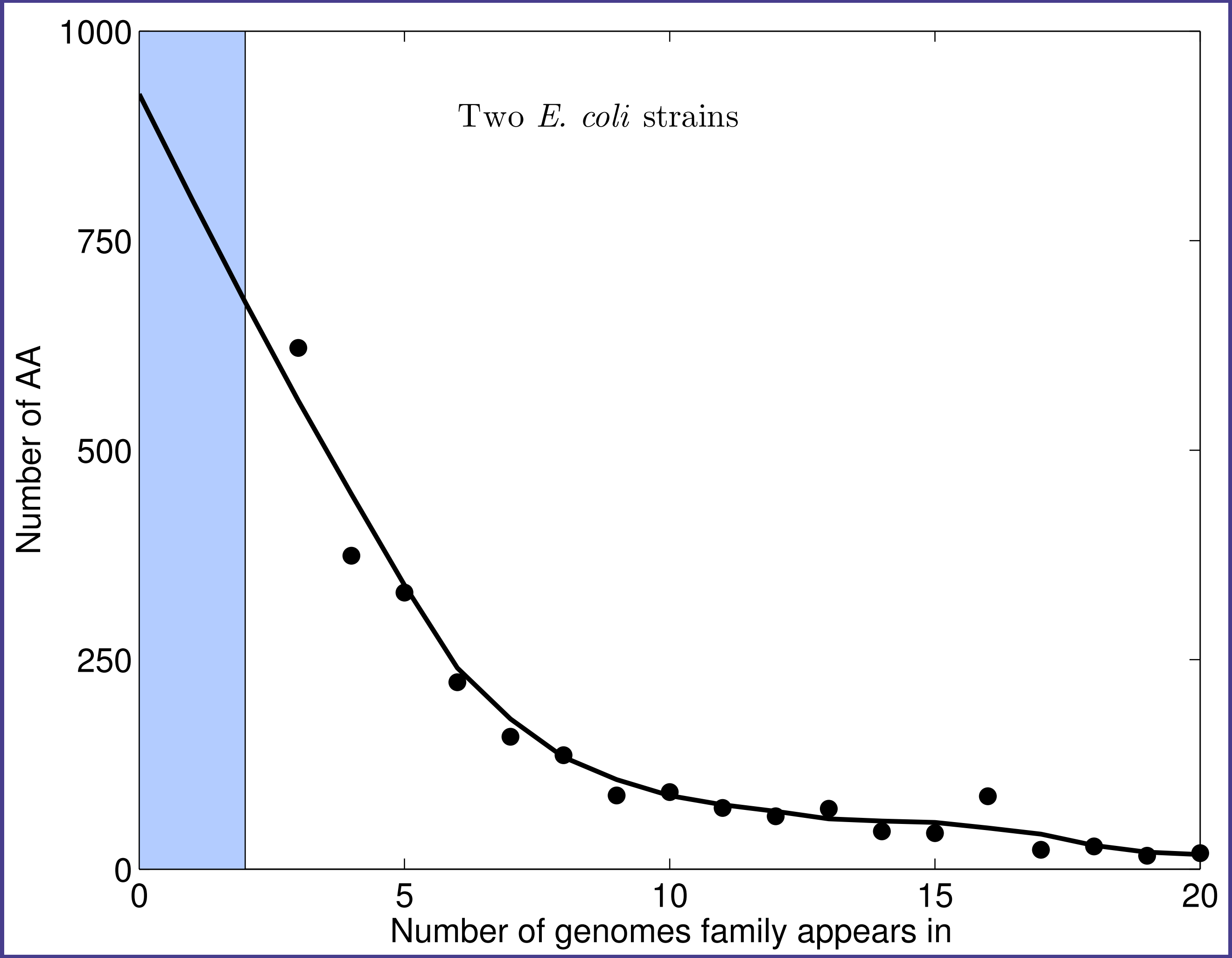
Residence times

At steady state,

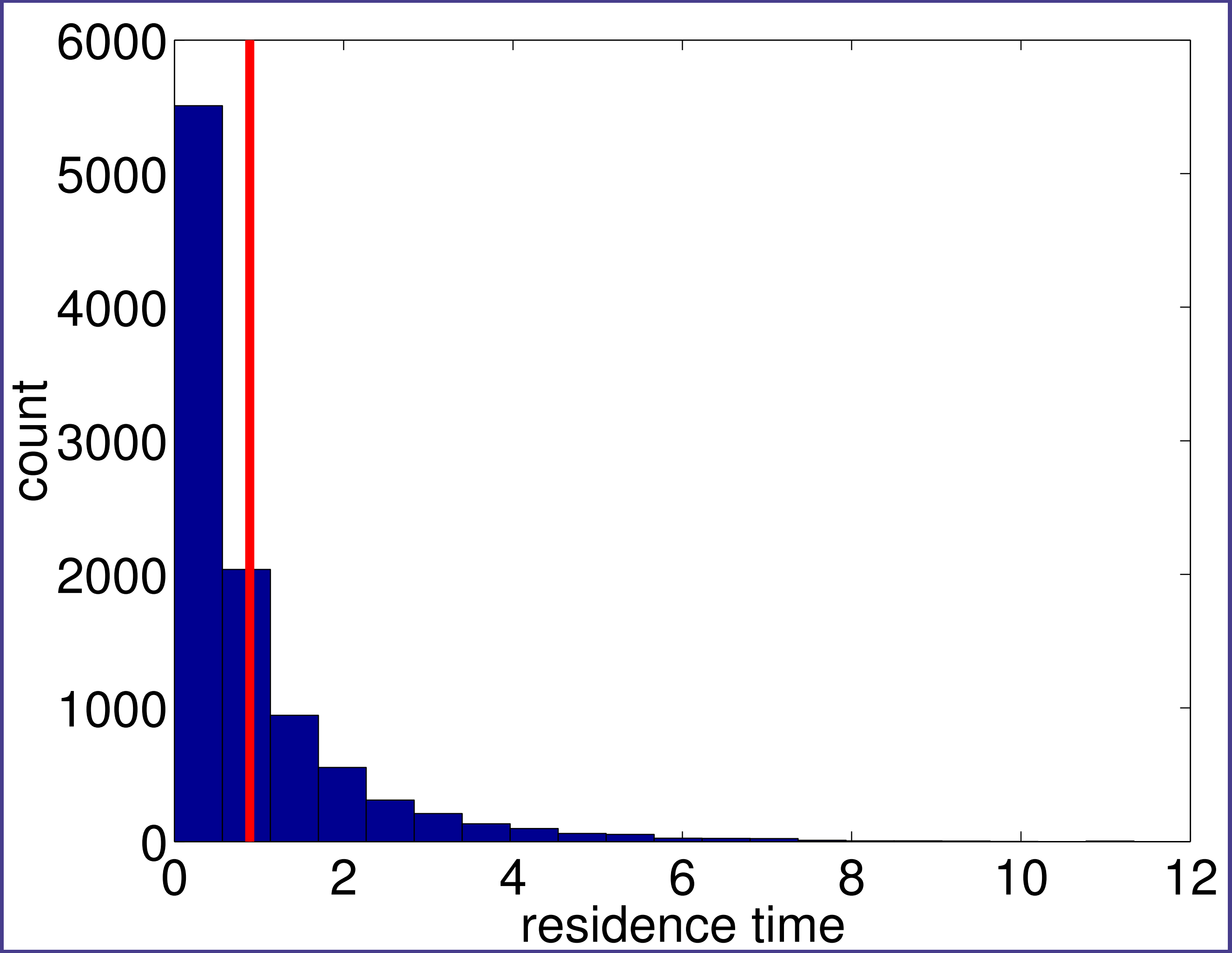
$$\beta_i = \sum_{j<i} q_{ji}\pi_j(i-j) / \sum_i \sum_{j<i} q_{ji}\pi_j(i-j)$$

The numerator is the sum of steady-state rates of flow into state i that add new genes, weighted by the number of genes $i-j$ each flow adds. The denominator normalizes the β_i to probabilities.

Unobservable data by extrapolation



Residence time distribution


count
residence time