Phylogenies based on gene content

Matthew Spencer

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

Acknowledgements

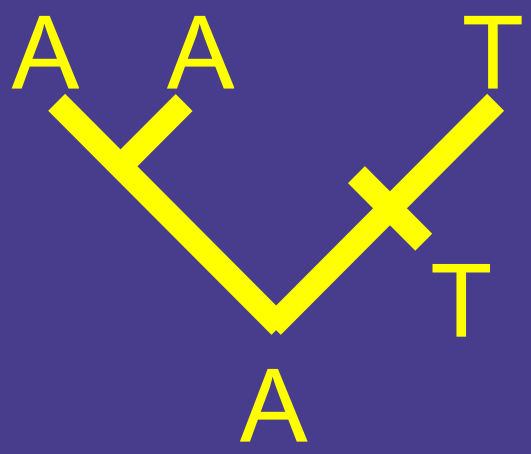
- Andrew Roger
- Ed Susko
- Dalhousie Statistical Evolutionary Bioinformatics group
- Genome Atlantic
- TIGR

Why base phylogeny on gene content?

Inferring deep phylogeny is difficult because

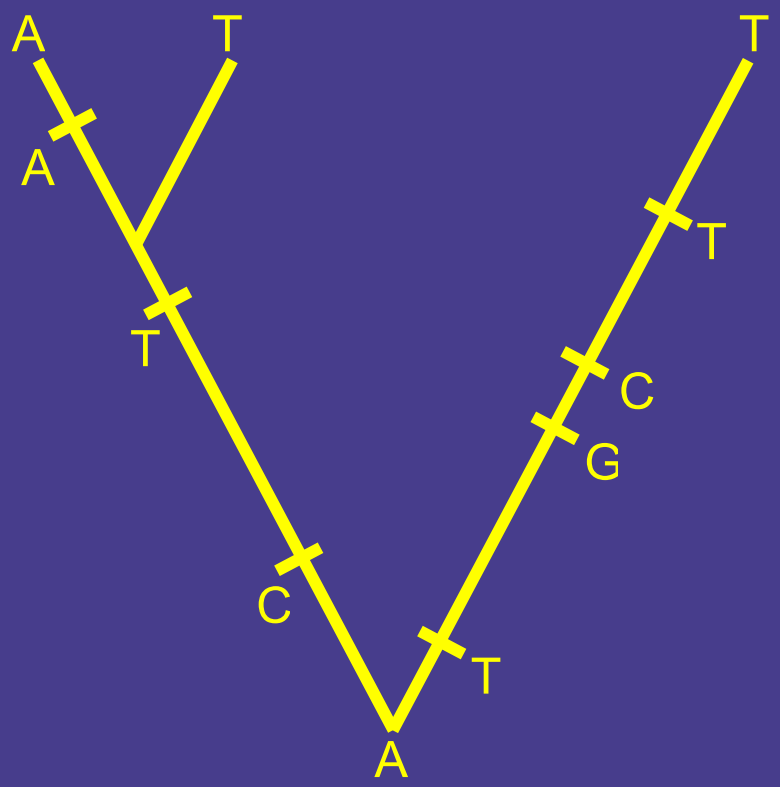
- sequence data are saturated with changes
- genes may be paralogs rather than orthologs
- lateral gene transfers may have been common
- few genes are ubiquitous

Saturation



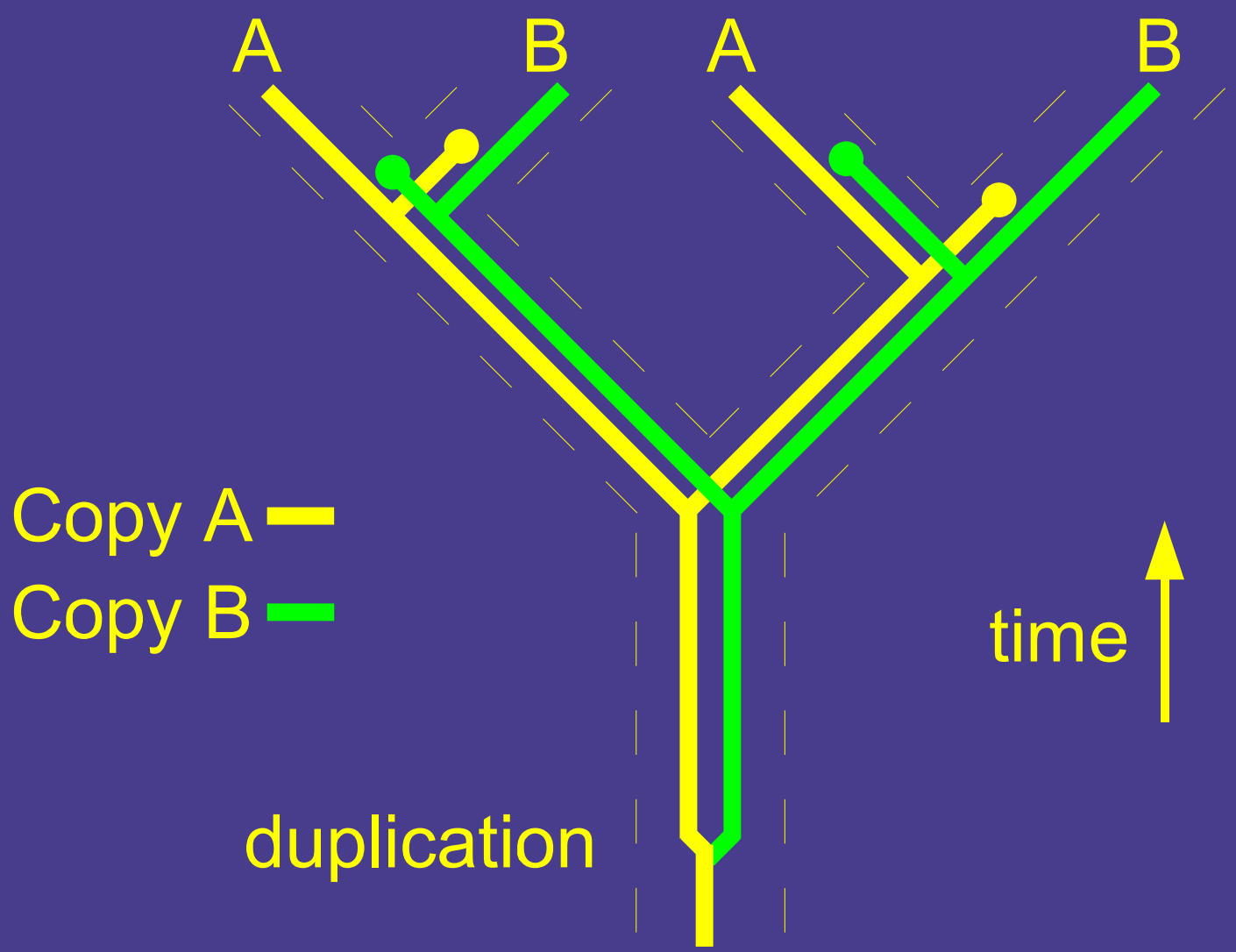
Over small evolutionary distances, closely related species tend to have the same nucleotide/amino acid states

Saturation

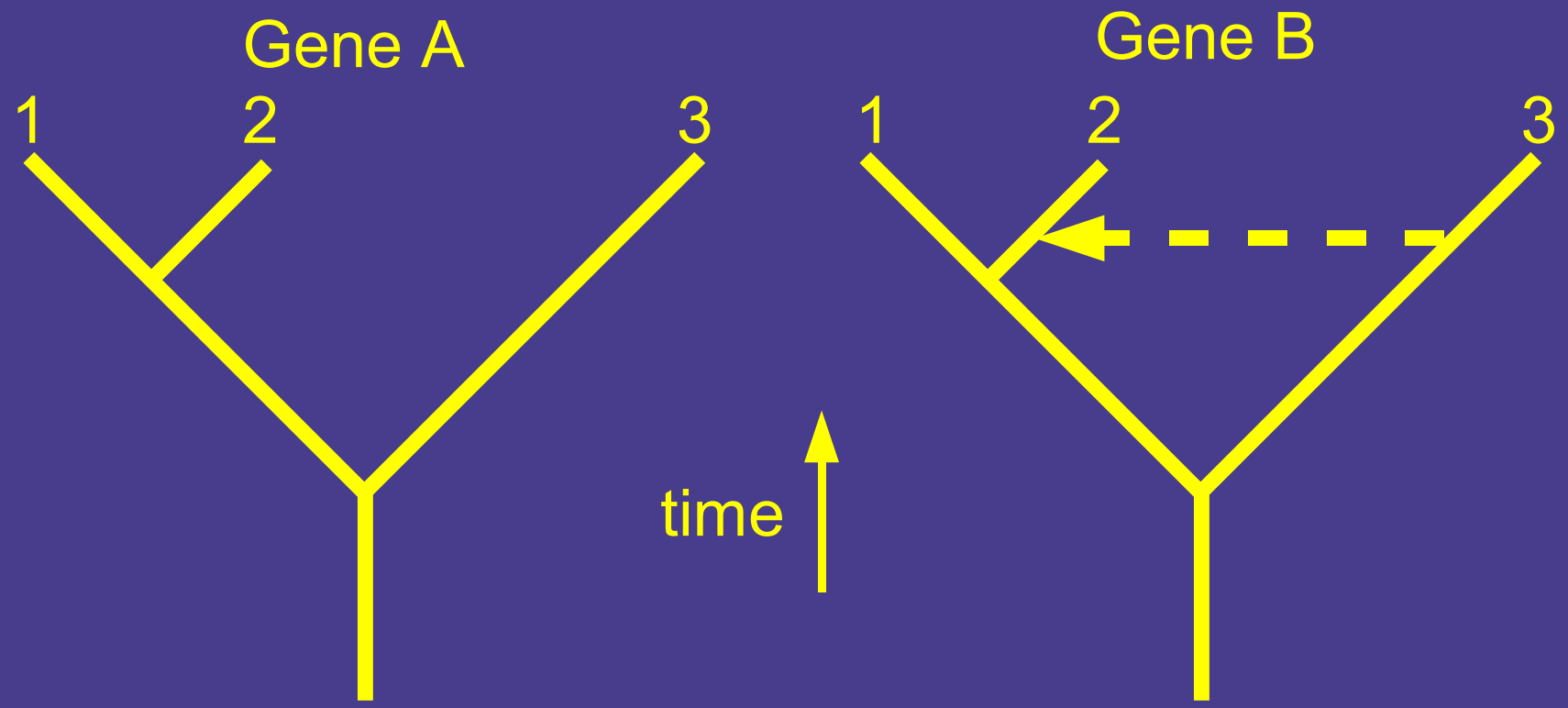


Over larger distances, so many changes have occured that nucleotide/amino acid states are not informative

Paralogy



Lateral gene transfer



Few genes are ubiquitous

- Genes are subject to frequent deletions and duplications
- Out of about 100 genomes sequenced by 2003, there were only about 60 ubiquitous genes
  Koonin (2003) Nature Reviews Microbiology 1:127-136
- Some of these will be saturated
- Others will have been affected by lateral gene transfer

Gene content phylogenetics

- We expect few duplications, deletions and transfers to separate closely related species
- We expect many such events to separate distantly-related species
- Why not use the number of such events as a measure of evolutionary distance?

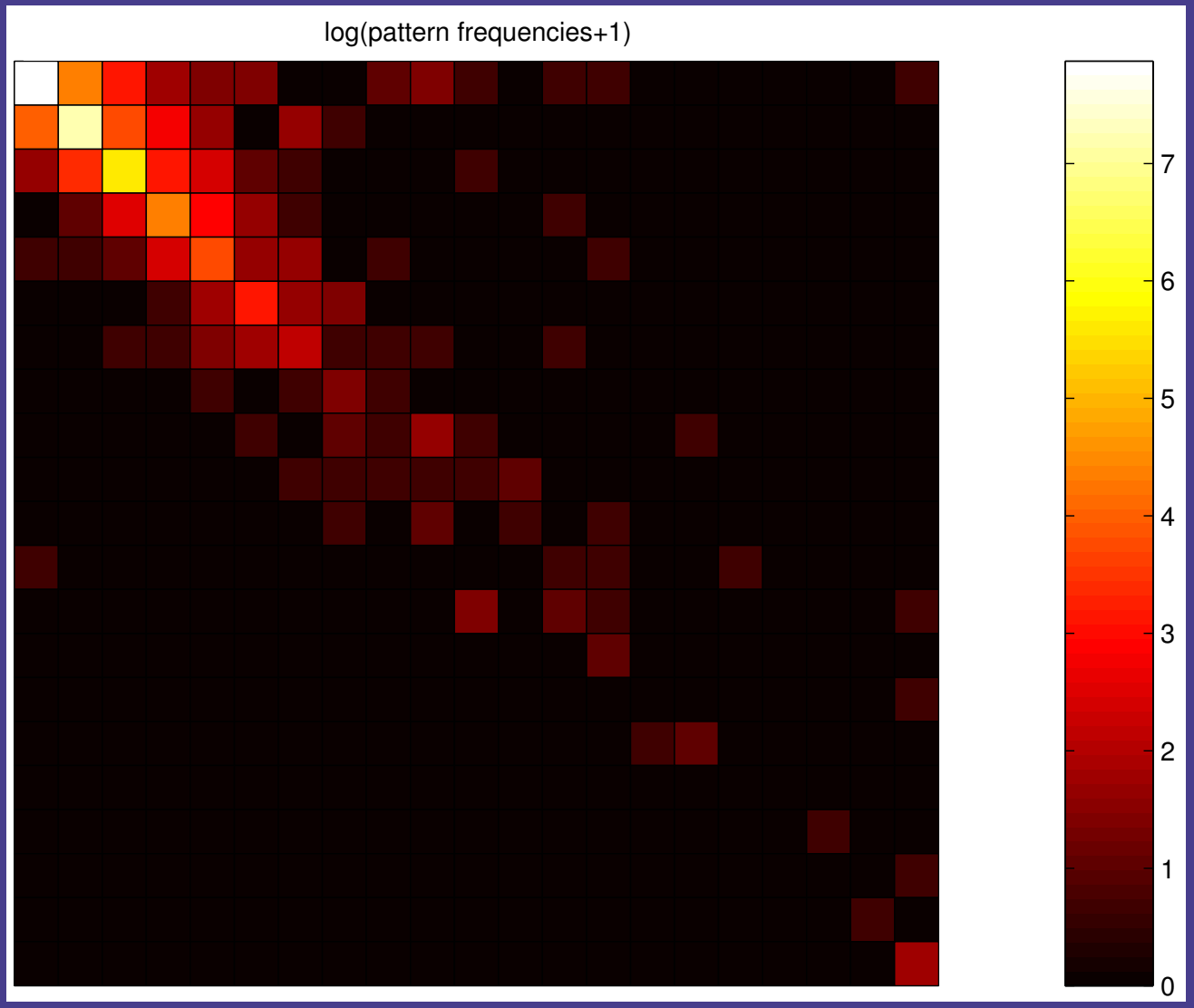
Gene presence/absence data

Presence/absence of gene families in two *E. coli* strains (closely-related)

		0157:H7 EDL933	
		absent (A)	present (P)
K12	absent (A)	2622	120
	present (P)	61	2070

Data from the COG database: http://www.ncbi.nlm.nih.gov/COG/

Gene family size data



Gene presence/absence data

Presence/absence of gene families in an archaeon (*Archaeoglobus fulgidus*) and a bacterium (*Bacillus subtilis*)

		B. subtilis	
		A	P
A. fulgidus	A	2448	1181
	P	654	590

Data from the COG database: http://www.ncbi.nlm.nih.gov/COG/

A naive distance measure

		B. subtilis	
		A	P
A. fulgidus	A	2448	1181
	P	654	590

Data from the COG database: http://www.ncbi.nlm.nih.gov/COG/

Total 4873 families

Distance= # differences / # families

A naive distance measure

- Two *E. coli* strains: distance 0.04
- *A. fulgidus* and *B. subtilis*: distance 0.38
- We could do this for all pairs of species in a database
- Methods like Neighbor-Joining (NJ) and Least-Squares (LS) can be used to estimate phylogenetic trees from pairwise distances

First problem: unobservable data

If a gene family is absent from our database, how do we know it exists?

		B. subtilis	
		A	P
A. fulgidus	A	2448+?	1181
	P	654	590

- Unobservable data mostly affect AA
- The denominator in our distance measure will be wrong

Second problem: multiple changes

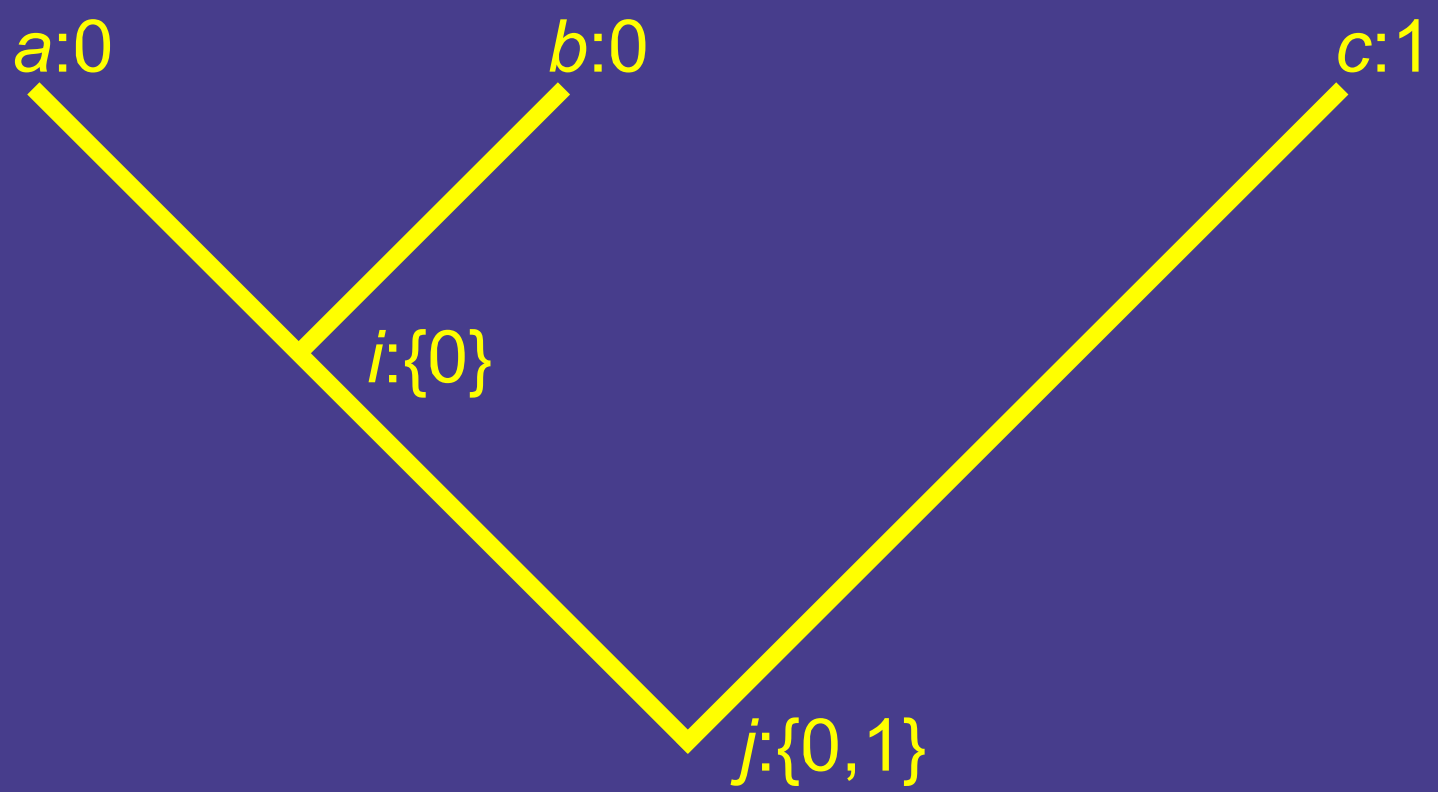
	# changes	
Pattern	True	Observed
$A \rightarrow A$	0	0
$A \rightarrow P$	1	1
$A \rightarrow P \rightarrow A$	2	0

We sometimes underestimate, but never overestimate

Methods

- parsimony
- naive distances
- SHOT
- paralinear distances
- Huson and Steel's method
- Gu and Zhang's method
- models with multi-gene events.

Parsimony



Parsimony

- One of the first methods to be applied to gene content data Fitz-Gibbon and House 1999, Nucleic Acids Research 27: 4218-4222, Montague and Hutchinson 2000, PNAS 97: 5334-5339.
- Most unobservable data are gene families that are absent everywhere. This implies no changes on the tree, so these data will have no effect on parsimony
- No attempt to deal with multiple changes. This leads to well-known problems with long branch attraction

Parsimony

How should we weight gene gains and losses?

- If gains were rare but losses were common, minimizing the number of gains should be more important than minimizing the number of losses
- Parsimony can't decide on the appropriate weights

Parsimony

- External criteria such as the plausibility of ancestral metabolic pathways could be used, but are somewhat subjective. Mirkin et al 2003, BMC Evolutionary Biology 3(2). Boussau et al 2004, PNAS 101:9722-9727.
- If the transition from absence to presence can occur only once, but multiple losses can occur, we could use Dollo parsimony. But lateral gene transfer could result in multiple absent to present transitions.

Naive distances

		B. subtilis	
		A	P
A. fulgidus	A	2448	1181
	P	654	590

Data from the COG database: http://www.ncbi.nlm.nih.gov/COG/

- We can imagine many different naive distances for these data
- We really want evolutionary distances based on a model for gene content

A model for gene content

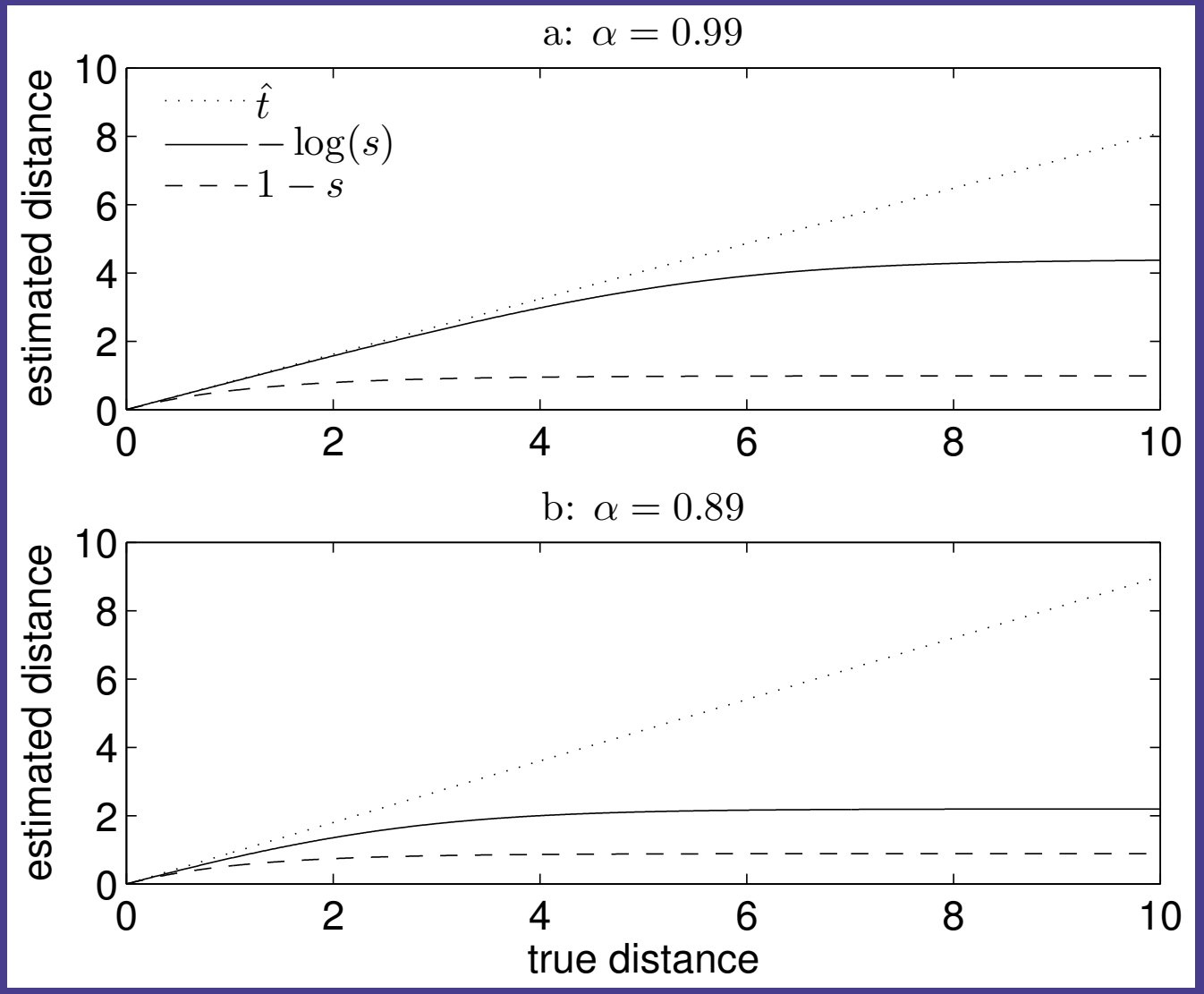
- Rate of change of gene presence =
  rate of change from absent to present $\times$
  probability of absence
  - rate of change from present to absent $\times$
  probability of presence
- Assume the same rates for all genes

SHOT distances

- Based on the number of shared orthologs (PP pattern)
- Normalized by a function of genome size to get a similarity measure s
- Distance=$-\log(s)$
- Approximate evolutionary distance when the distance is small and most genes are absent from most genomes

Korbel et al 2002 Trends in Genetics 18:158-162

SHOT distances



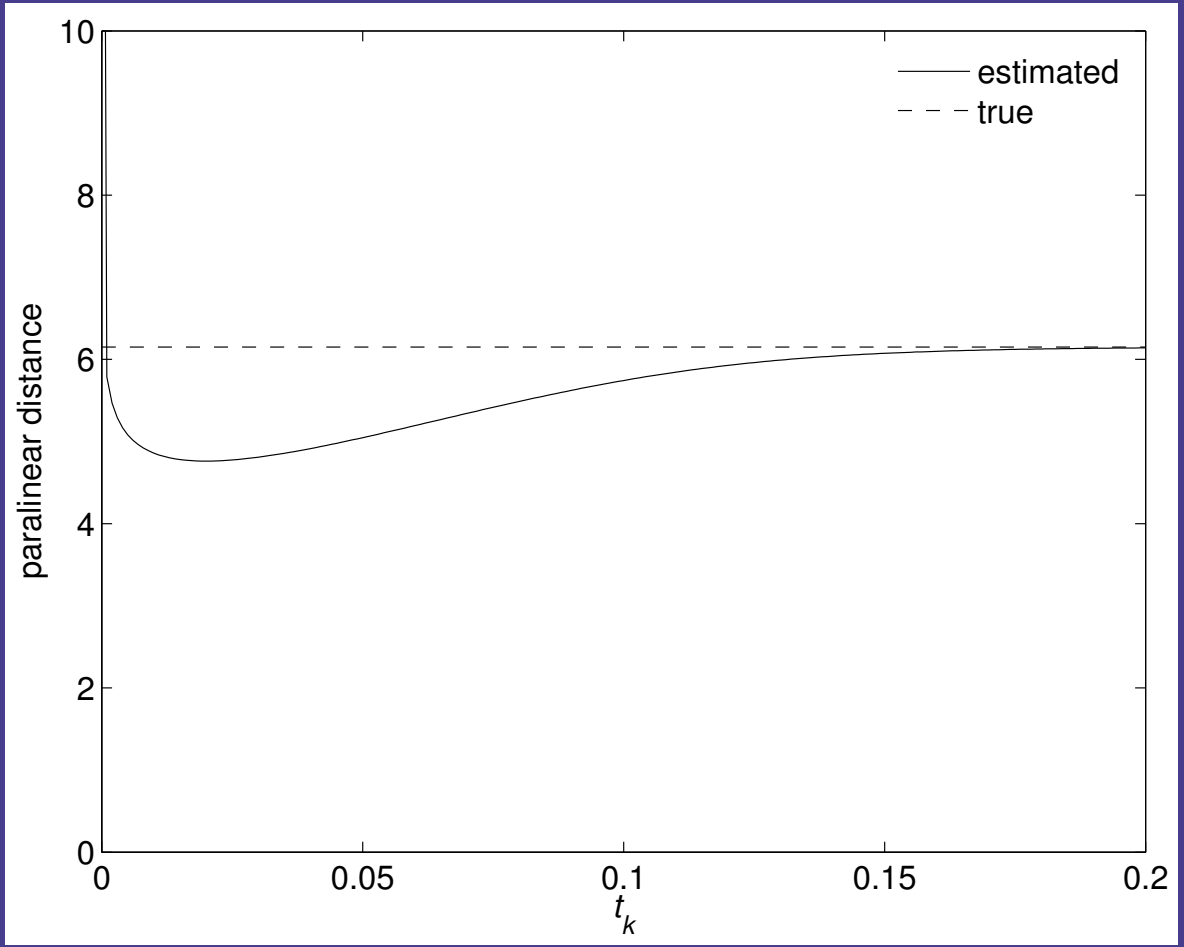
Paralinear distances

- Composition bias in sequence-based phylogenetics: Organisms with similar nucleotide or amino acid composition tend to be grouped together
- Genome size varies. *Mycoplasma genitalium*: 362 gene families in the COG database. *Pseudomonas aeruginosa*: 2243.
- Paralinear (logdet) distances can sometimes deal with sequence composition biases.

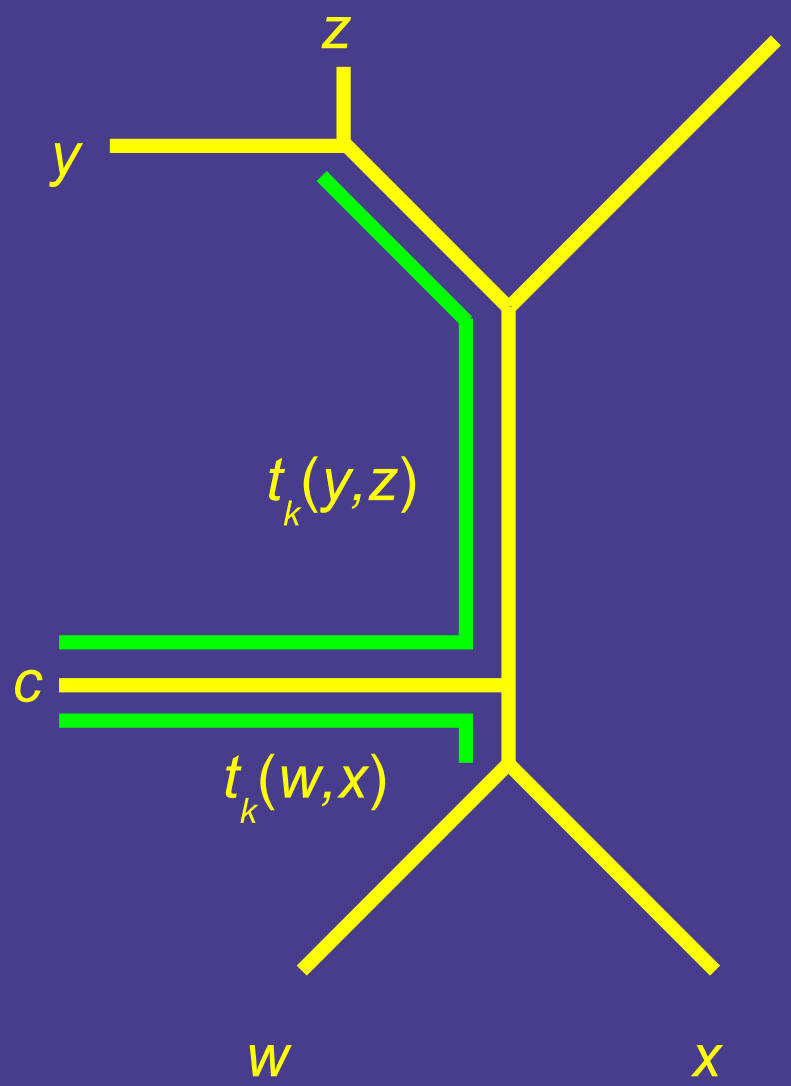
Lake and Rivera 2004, Molecular Biology and Evolution 21: 681-690. Rivera and Lake 2004, Nature 431: 152-155.

Conditioning genome

Use only those genes present in a third genome.



Conditioning genome



Huson and Steel's model

- Rate of change of genome size =
  gene birth rate
  - gene death rate $\times$ number of genes
- Rate of gene birth independent of number of genes: assume duplications are unimportant?
- Huson and Steel derive an evolutionary distance from this
- Doesn't need any correction for unobservable data

Huson and Steel 2004, Bioinformatics 20: 2044-2049

Gu and Zhang's method

- Rate of change of probability of having n genes in family =
  birth rate $\times(n-1)\times$ probability of having $n-1$ genes
  + death rate $\times(n+1)\times$ probability of having $n+1$ genes
  - (birth rate+death rate) $\times(n)\times$ probability of having n genes
- Assume we can delete or duplicate one gene at a time

Gu and Zhang's method

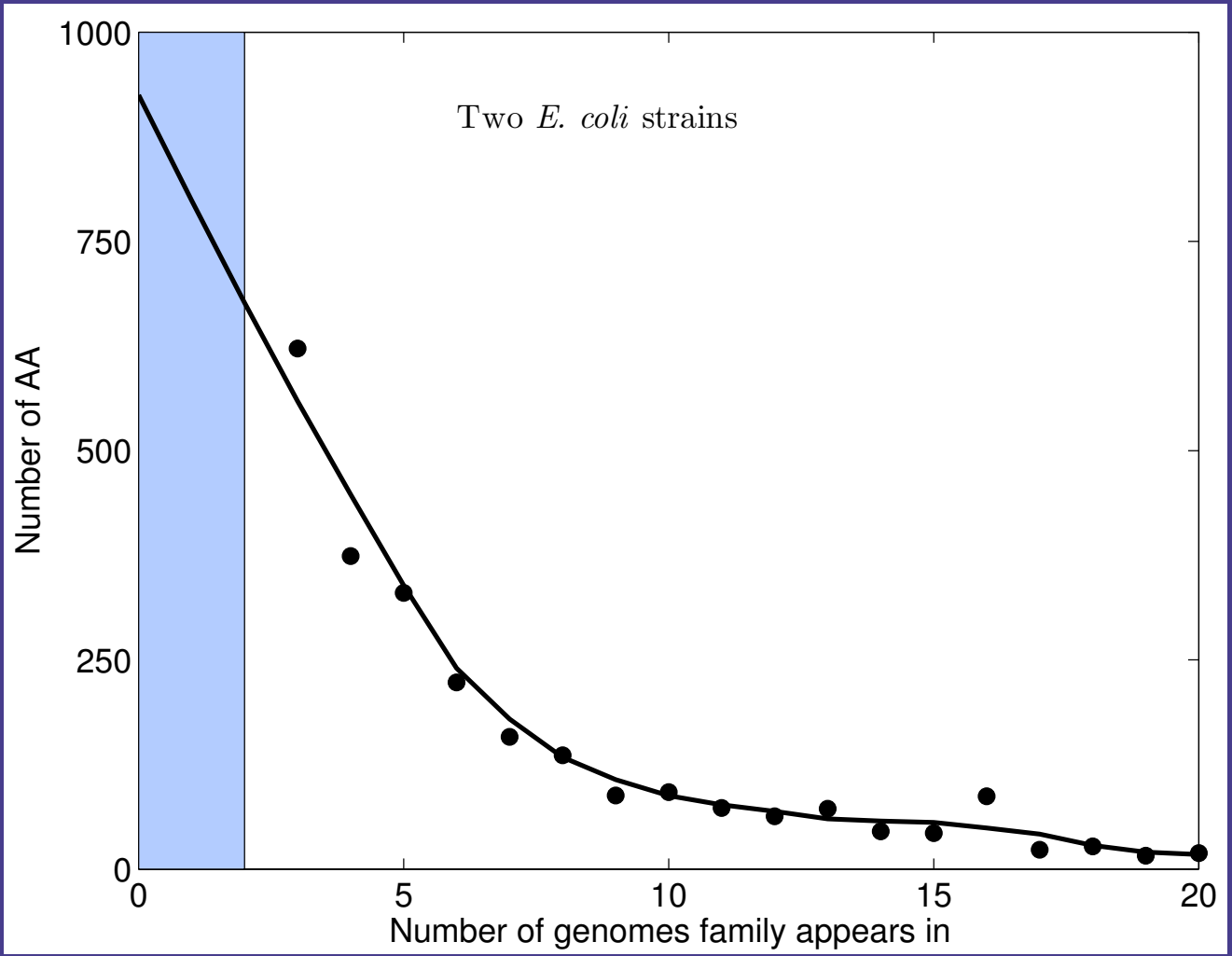
- We can't calculate evolutionary distances for this model from presence/absence data alone
- Gu and Zhang used absence/1 member of family/> 1 member of family
- They used likelihood conditional on the data being observable to estimate evolutionary distances (assuming AA is the only unobservable pattern)
- Applied to 35 genomes from the COG database

Gu and Zhang 2004, Molecular Biology and Evolution 21: 1401-1408

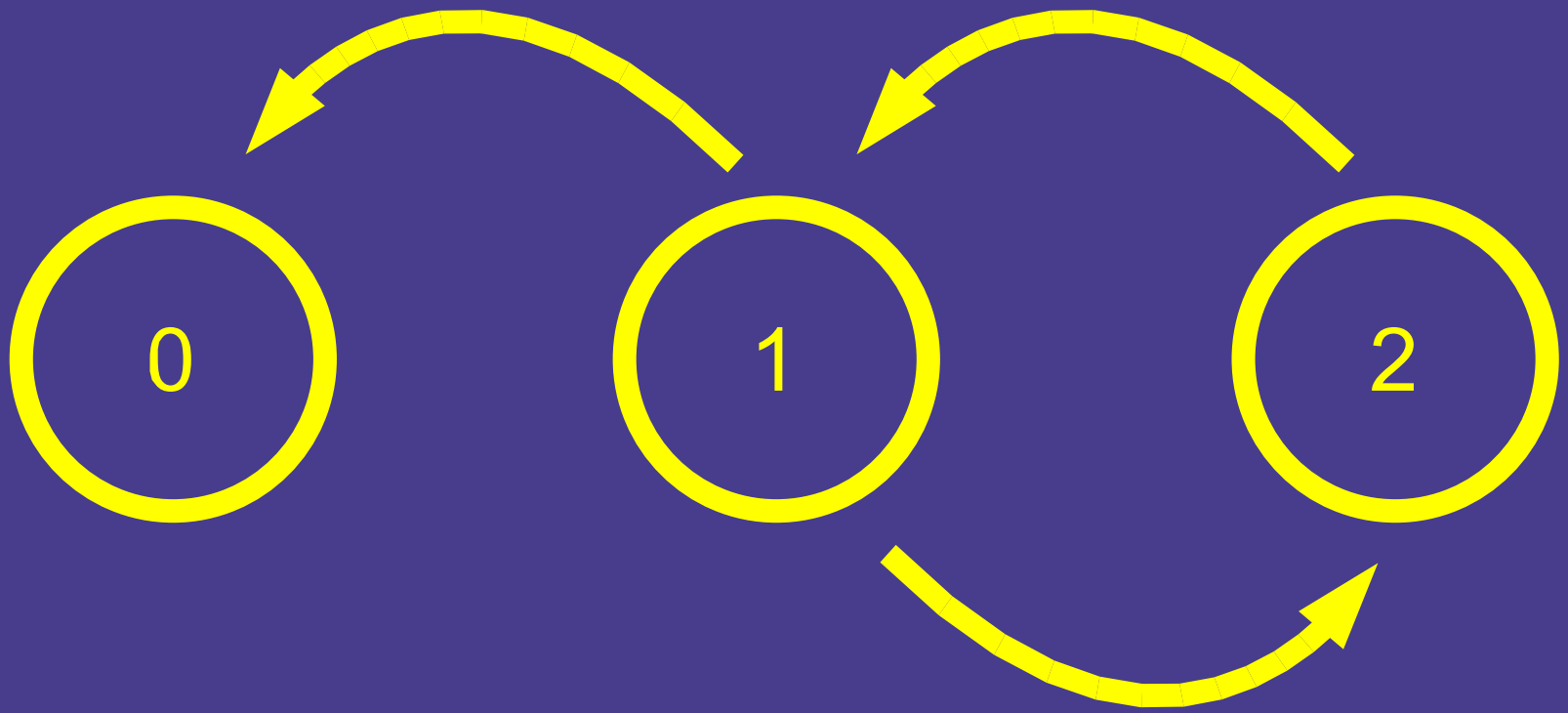
Possible improvements

- A gene family must be in at least 3 genomes to appear in the COG database, so pairwise data can't really tell us about unobservable patterns. AA is also the commonest observed pattern, so we don't want to discard it.
- Allow more than one gene to be deleted or duplicated at a time
- Allow lateral gene transfer and innovation (otherwise a gene family that is lost is lost forever)

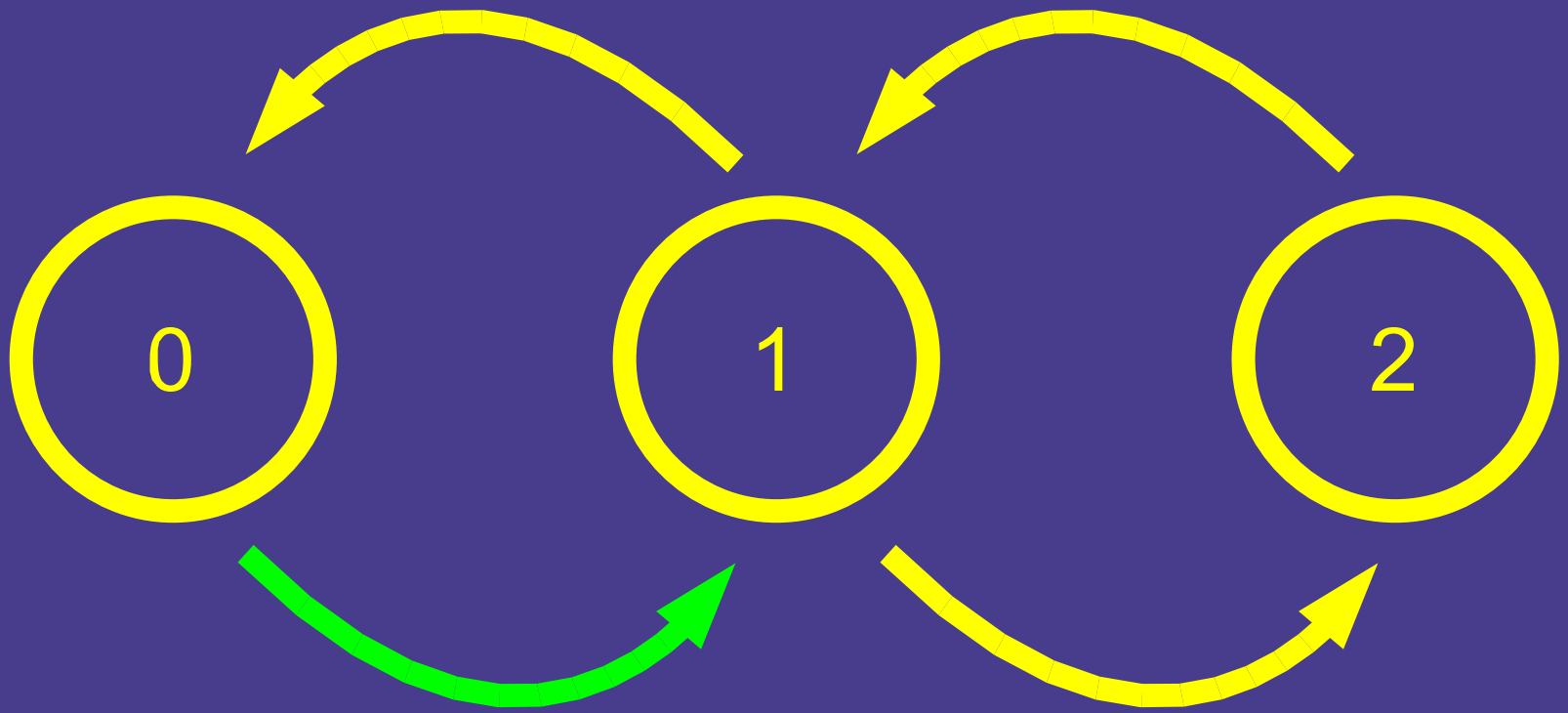
Unobservable data by extrapolation



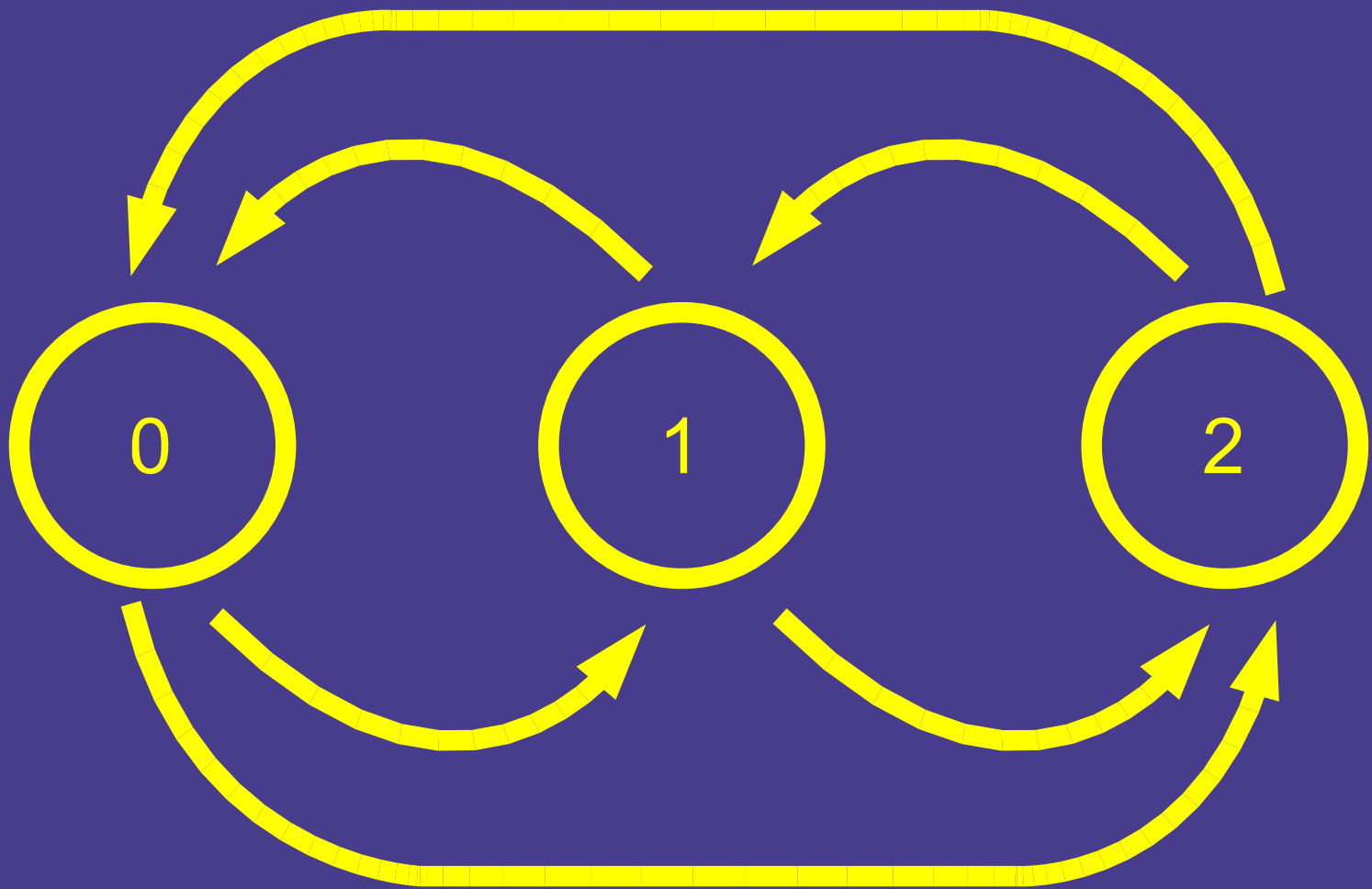
Single-gene events



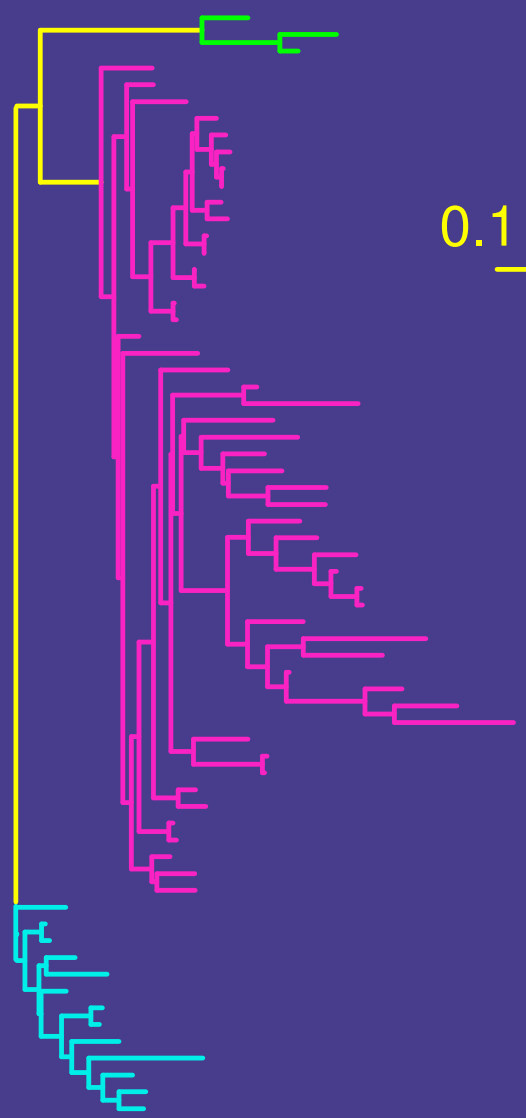
Single-gene events



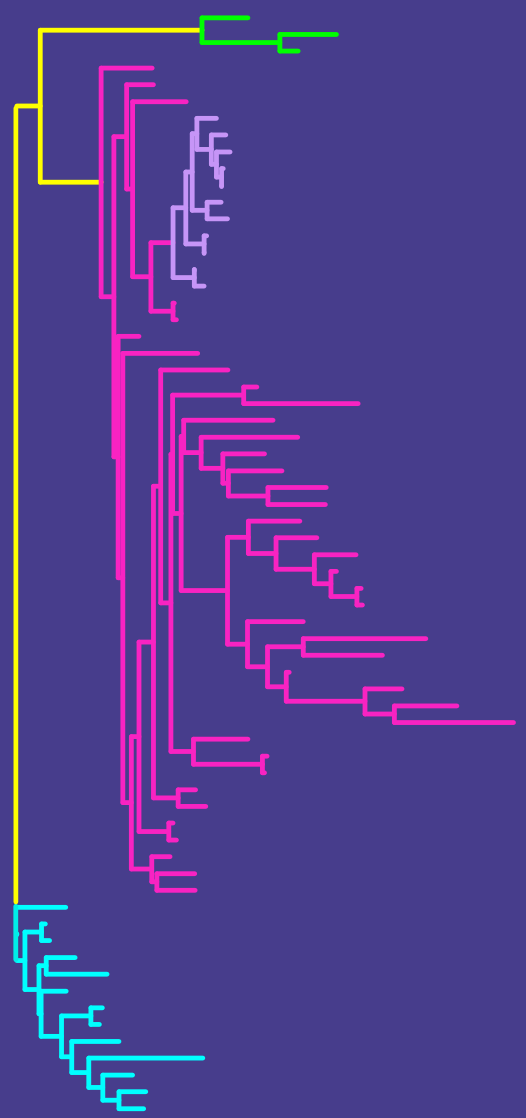
Multi-gene events



Microbial gene content phylogeny



Microbial gene content phylogeny



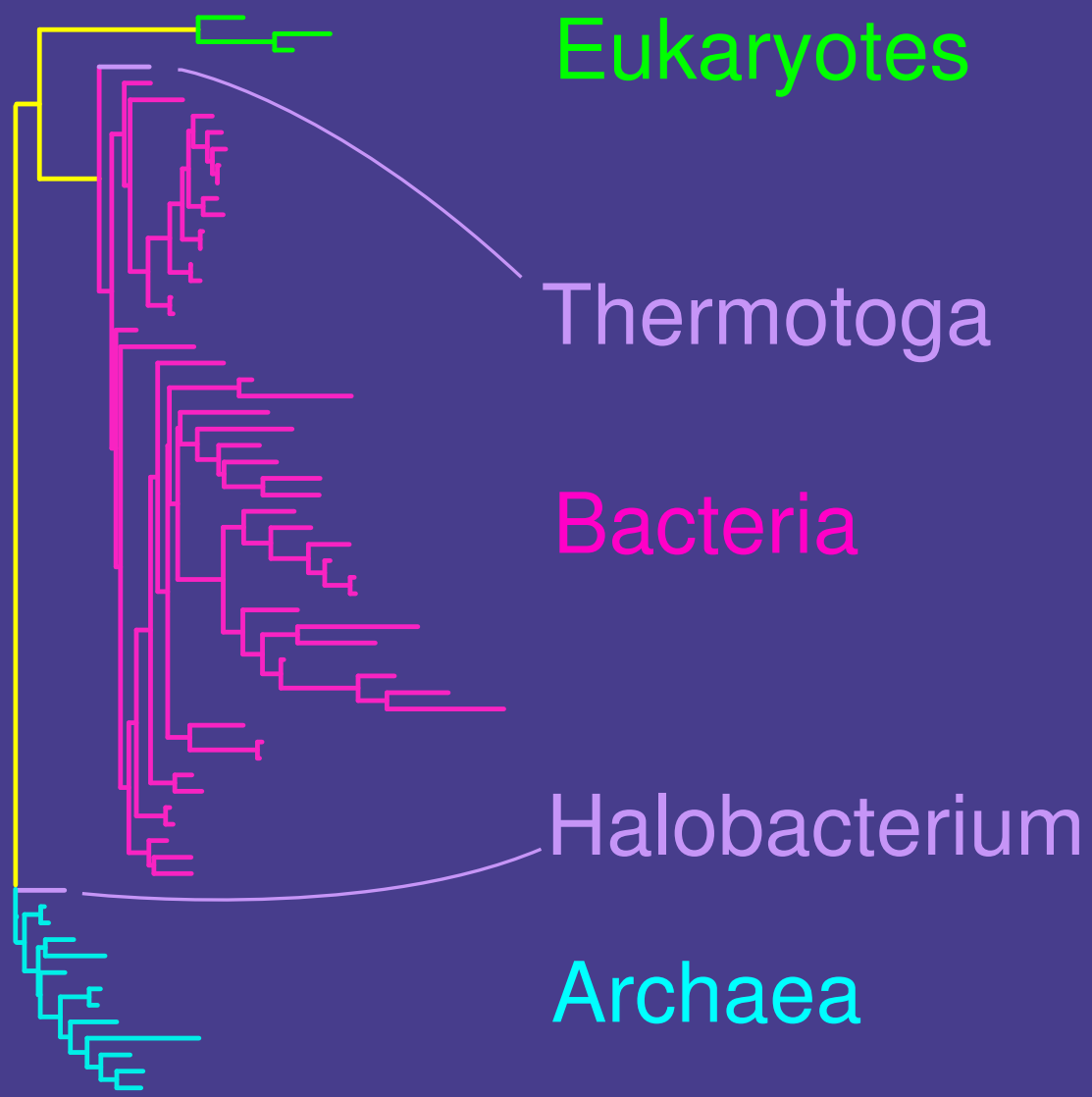
Eukaryotes

Parasites/
endosymbionts

Bacteria

Archaea

Microbial gene content phylogeny



What can we learn about biology?

We looked at rate estimates for *E. coli* and for *A. fulgidus* & *B. subtilis*

- The estimated rate of transfers of multiple genes in the same family was not significantly greater than zero
- Transitions from 0 to 1 genes may be mostly lateral transfers, and had a rate about 1/5 of the rate of deletions of entire gene families
- A single gene might persist in the genome for about the time that separates *A. fulgidus* from *B. subtilis*

Recommendations

- Don't use parsimony. We don't know how to weight gains and losses, and there are problems with multiple changes
- Don't use naive distances. They aren't good measures of evolutionary distances
- SHOT distances are OK for a quick analysis
- Paralinear distances can deal with genome size variation, but their properties aren't yet well understood for gene content
- More sophisticated models can be used to learn more about biology

Areas for future work

- Better models of gene content: rate variation across gene families and species?
- Full maximum likelihood? So far, this has only been done for small numbers of species

  Zhang & Gu 2004, Statistical Applications in Genetics & Molecular Biology 3:31
- Network methods to locate areas of extensive lateral gene transfer?

More information:

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