Correlation between composition and site-specific evolutionary rate and implications for phylogenetic inference

nonhomogeneous RNA models

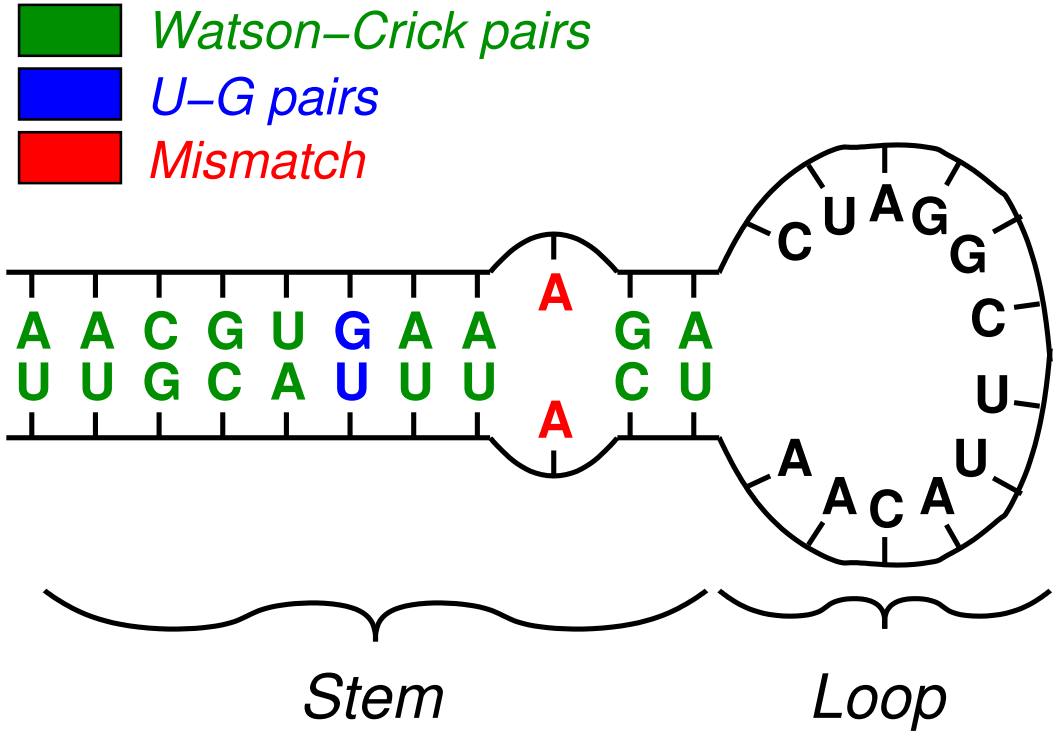
Vivek Gowri-Shankar, Magnus Rattray

vivek.gowri-shankar@cs.man.ac.uk, magnus.rattray@cs.man.ac.uk

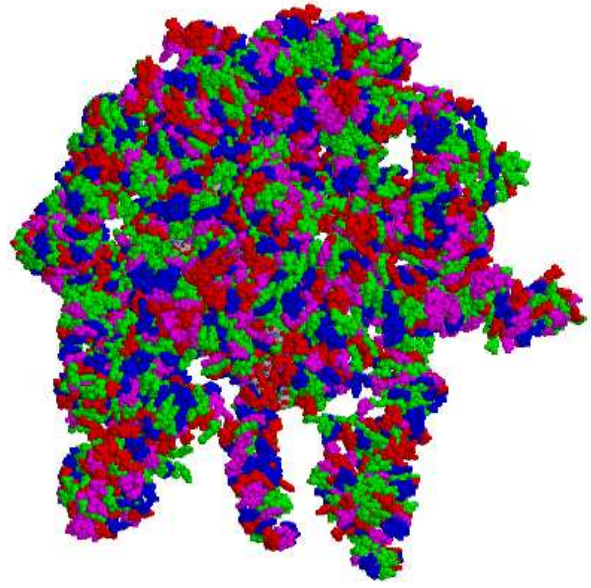
School of Computer Science - University of Manchester

Background: RNA genes

- Single-stranded chains of nucleotides (A,C,G,U)
- Some complementary regions (A:U, G:C pairs) $\Rightarrow$ folding



Secondary structure

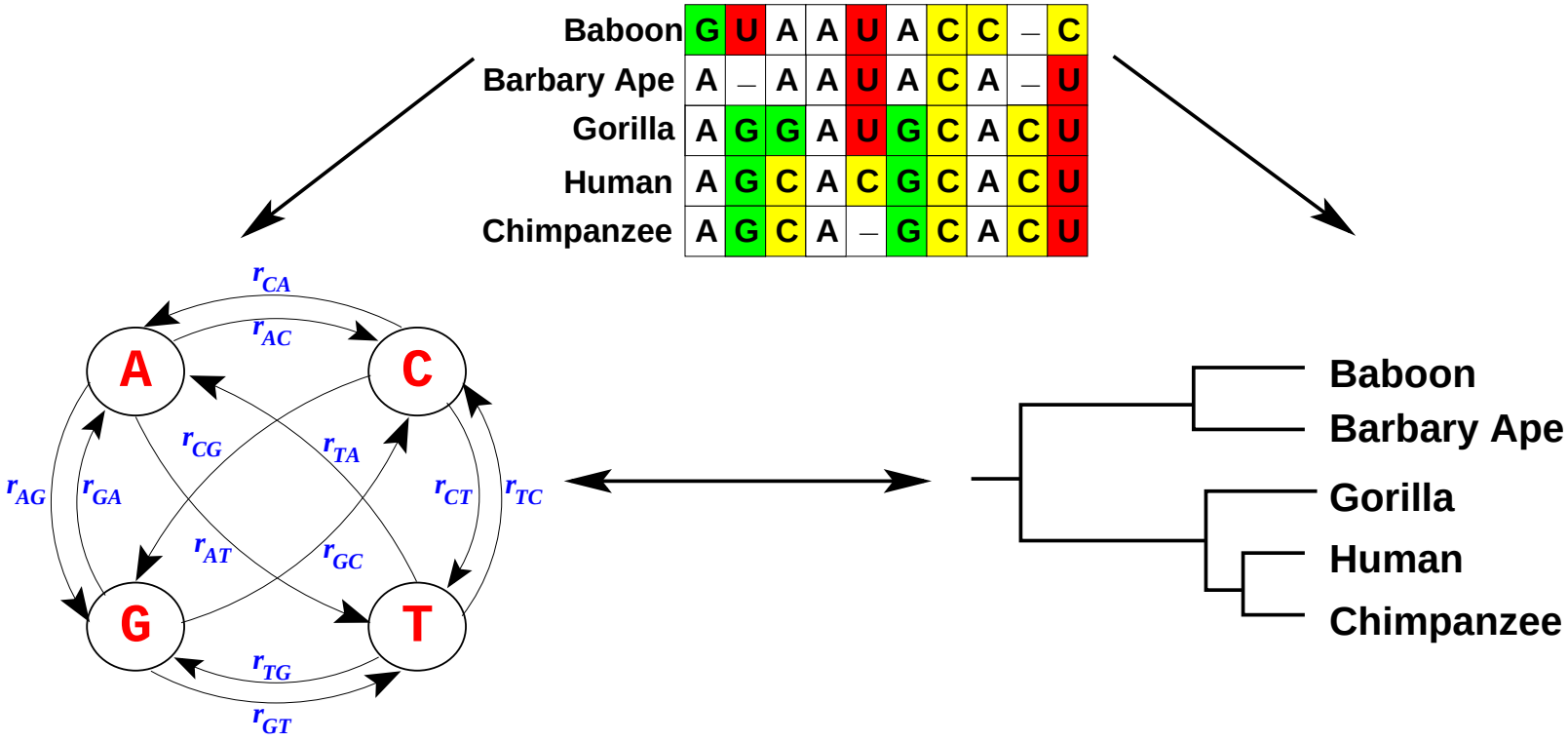


23S tertiary structure

Talk outline

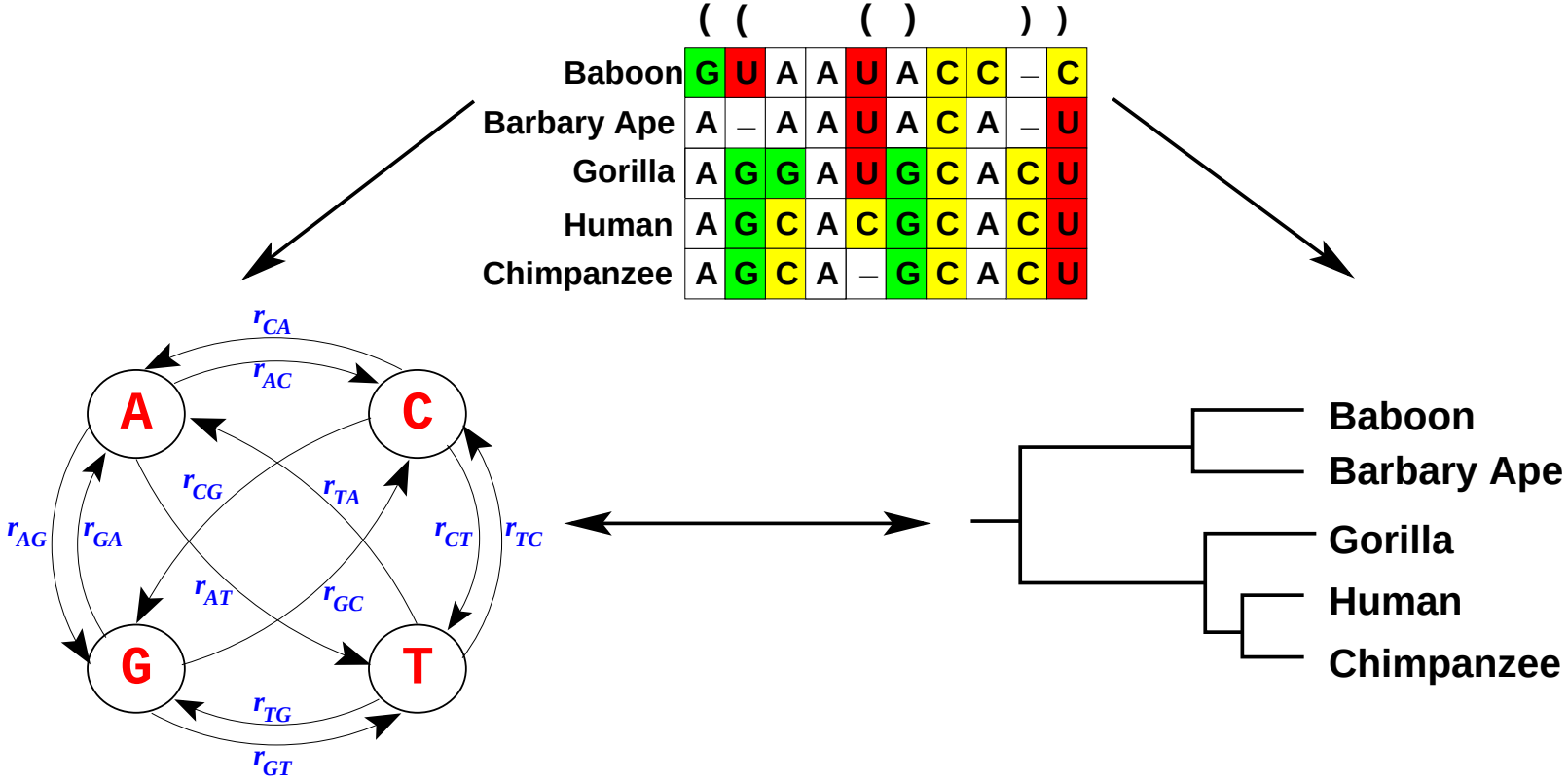
- Phylogenetic inference with RNA genes
- Compositional heterogeneity across sites
- Consequences on phylogenetic inferences
- A space-time model for compositional heterogeneity
- Results related to the thermophily of LUCA

Statistical molecular phylogenetics



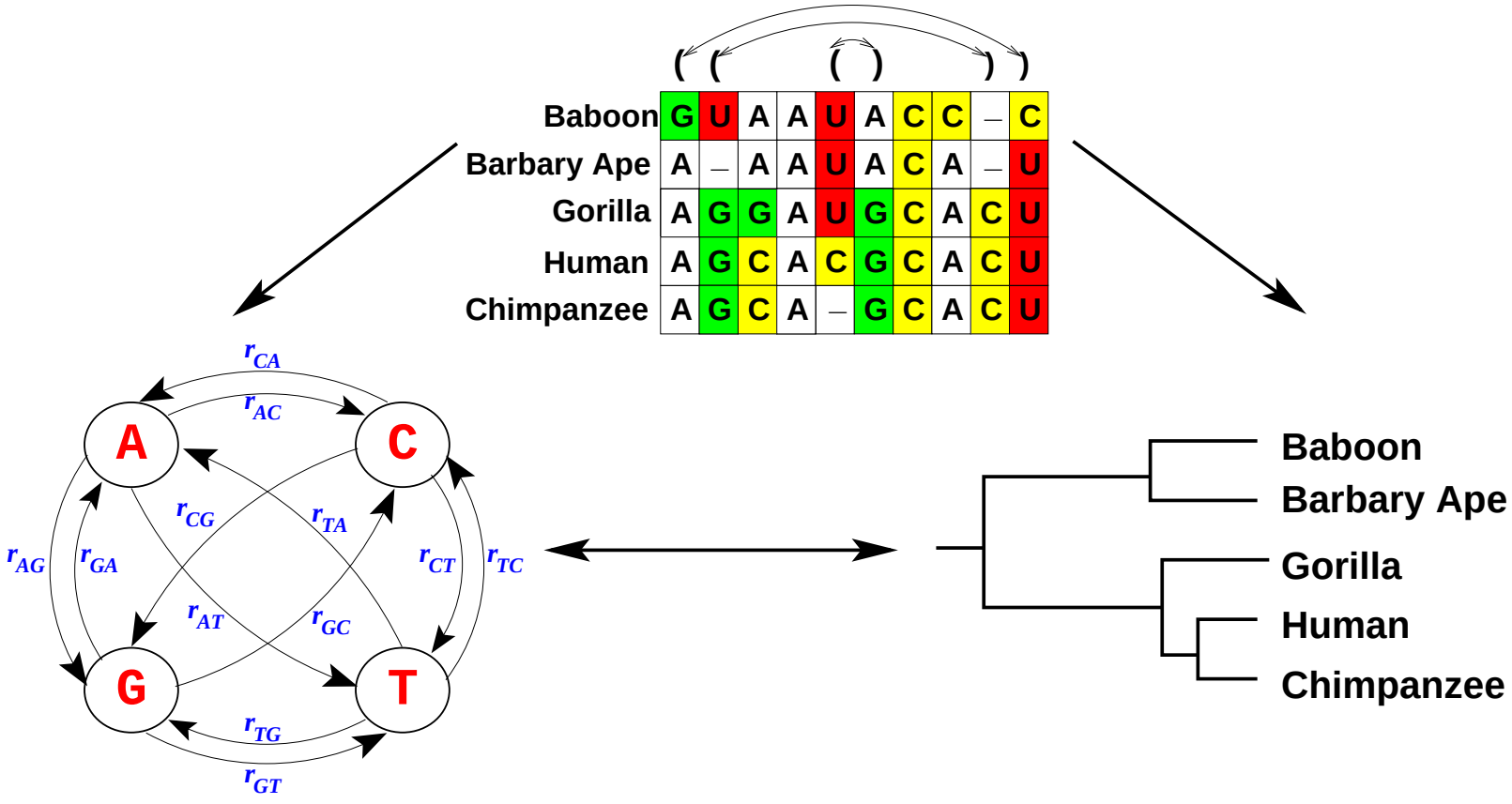
- alignment → substitution parameters + tree
- time-homogeneity: process constant over time and lineages
- nucleotides evolve independently with the same process

Statistical molecular phylogenetics



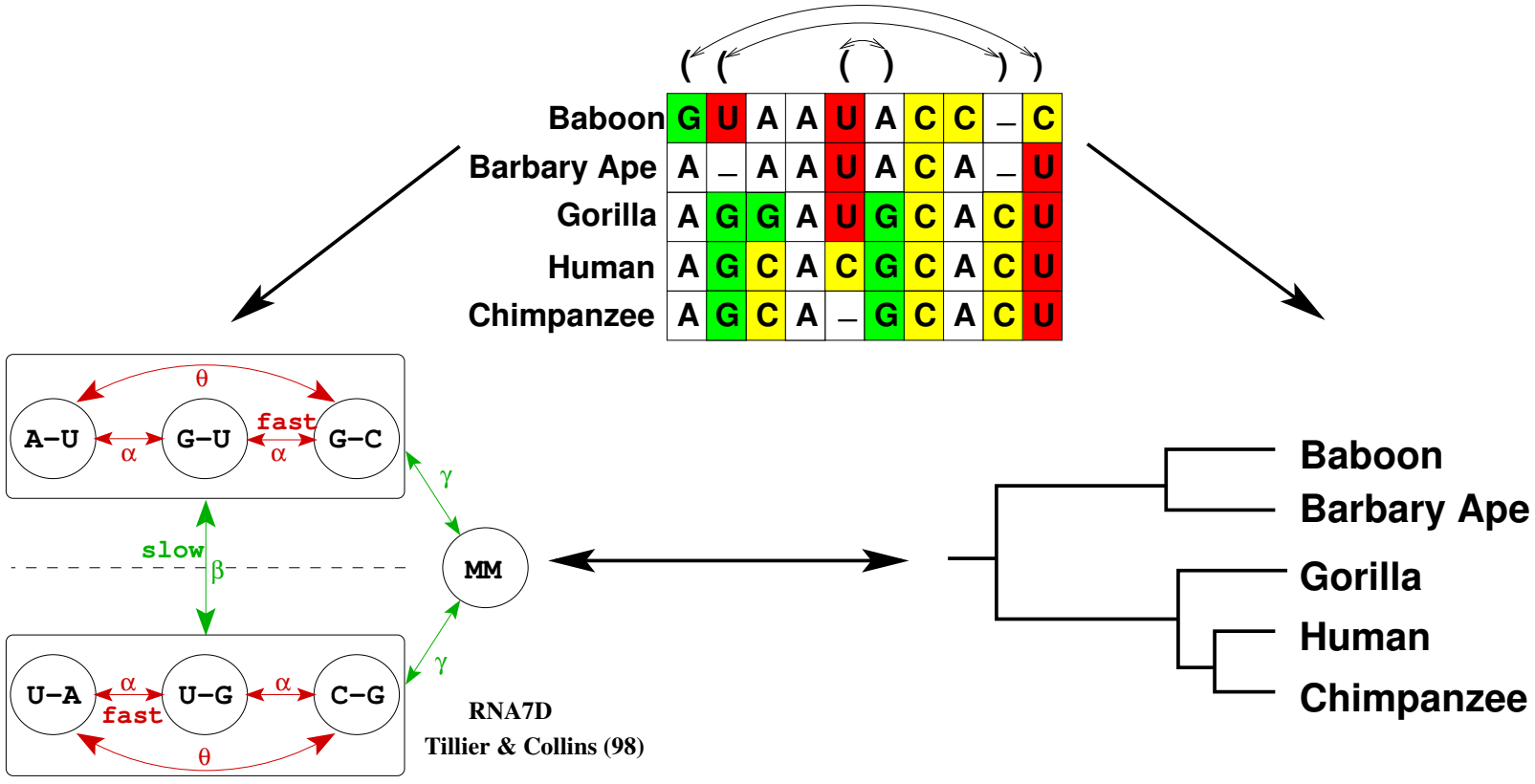
- alignment → substitution parameters + tree
- time-homogeneity: process constant over time and lineages
- nucleotides evolve independently with the same process

Statistical molecular phylogenetics



- alignment → substitution parameters + tree
- time-homogeneity: process constant over time and lineages
- nucleotides evolve independently with the same process

Statistical molecular phylogenetics

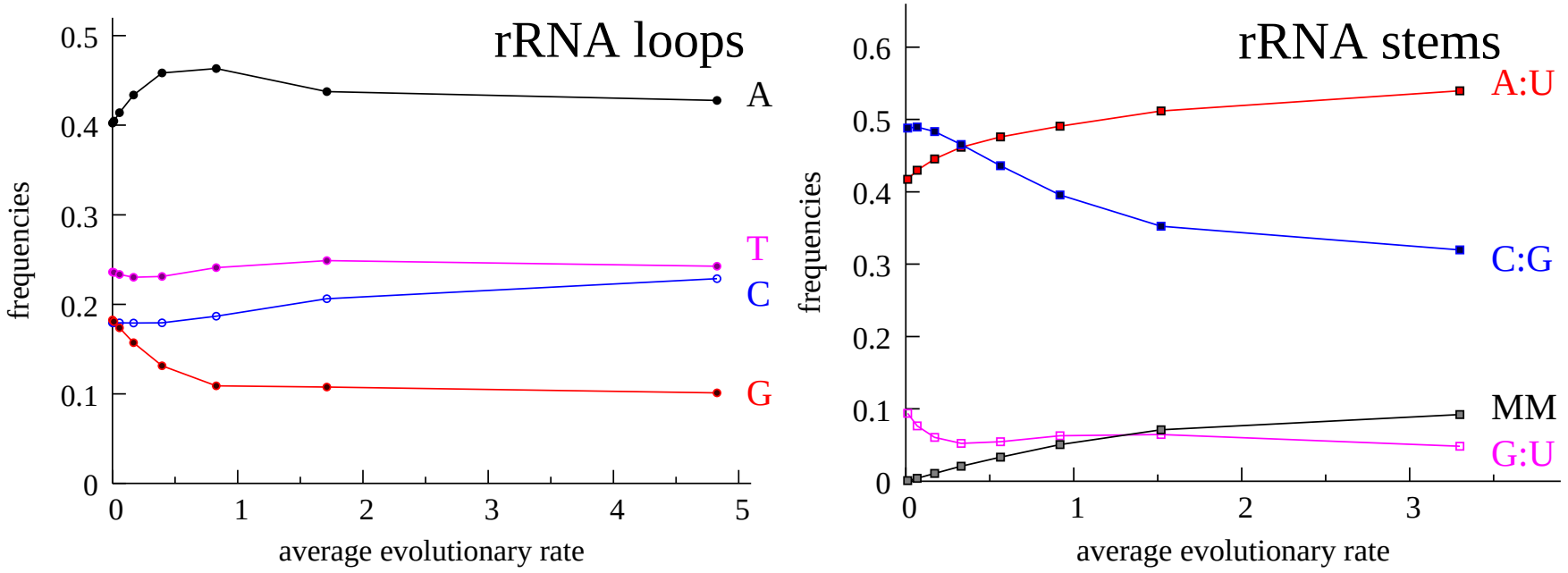


- alignment → substitution parameters + tree
- time-homogeneity: process constant over time and lineages
- nucleotides evolve ~~independently~~ with the same process

Compositional variation with RNA genes

- compositional trends with real RNA sequences:

  site-specific composition and rates are correlated



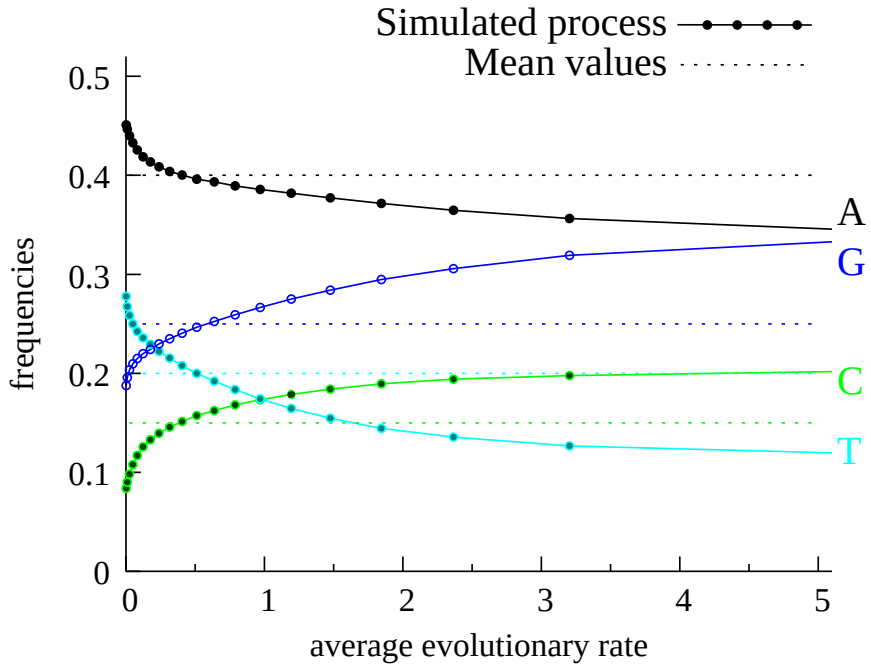
mitochondrial rRNAs of 13 primates + 3 outgroup species

- Loops and stems were used simultaneously in a mixed phylogenetic inference analysis with two different substitution models (a 4-state for loops and 7-state for stems)

Effects on parameter estimates

- Frequency estimates are biased towards composition of fast-evolving sites

	real π	inferred π
π_A	.40	.387
π_C	.25	.273
π_G	.15	.163
π_T	.20	.177



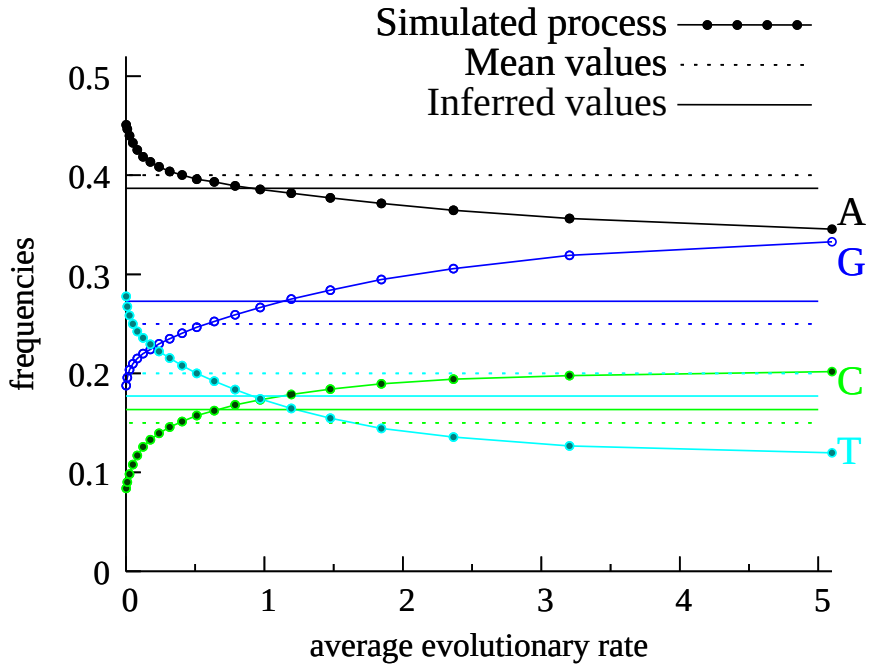
composition profiles

- Branch lengths are slightly underestimated
- Support values are affected

Effects on parameter estimates

- Frequency estimates are biased towards composition of fast-evolving sites

	real π	inferred π
π_A	.40	.387
π_C	.25	.273
π_G	.15	.163
π_T	.20	.177

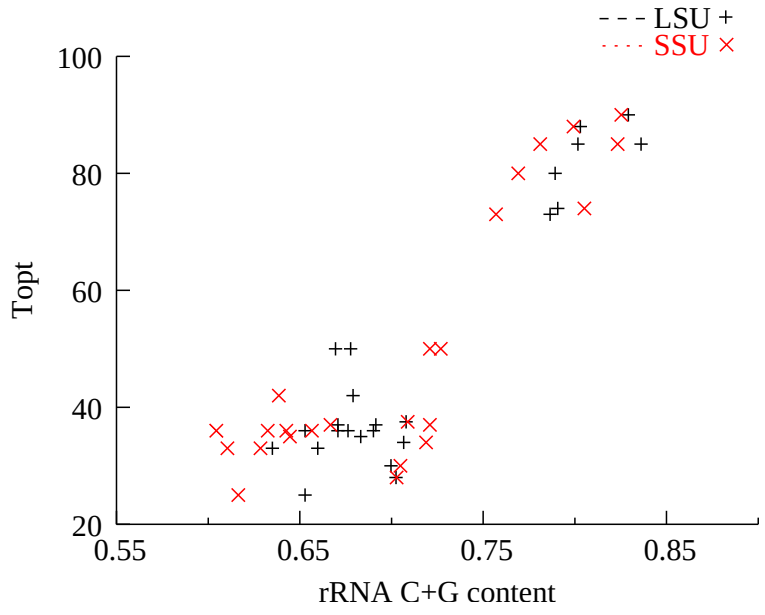


composition profiles

- Branch lengths are slightly underestimated
- Support values are affected

Compositional variation across species

- Compositional bias among species



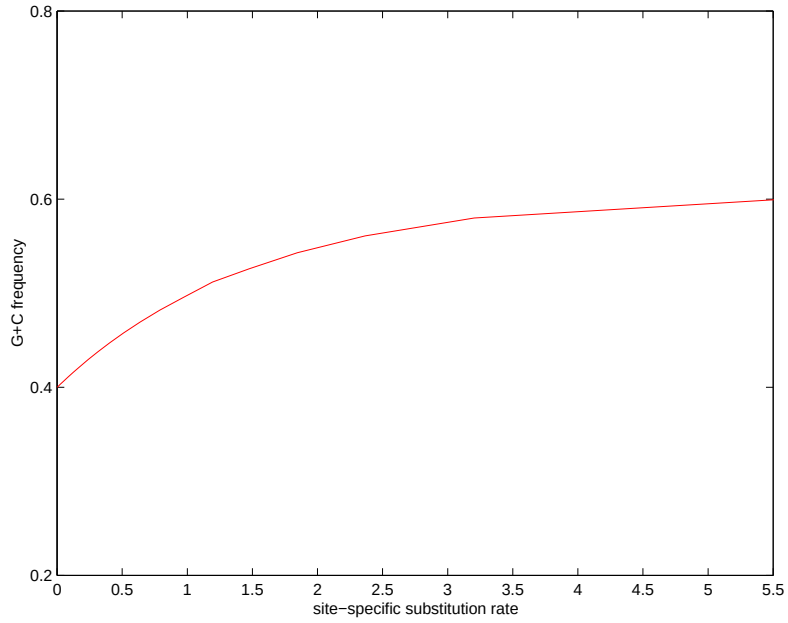
Correlation between rRNA G+C content (in stems) and prokaryotes Optimal Growth Temperature

- The substitution process is not constant in time

A nonhyperthermophilic origin of life?

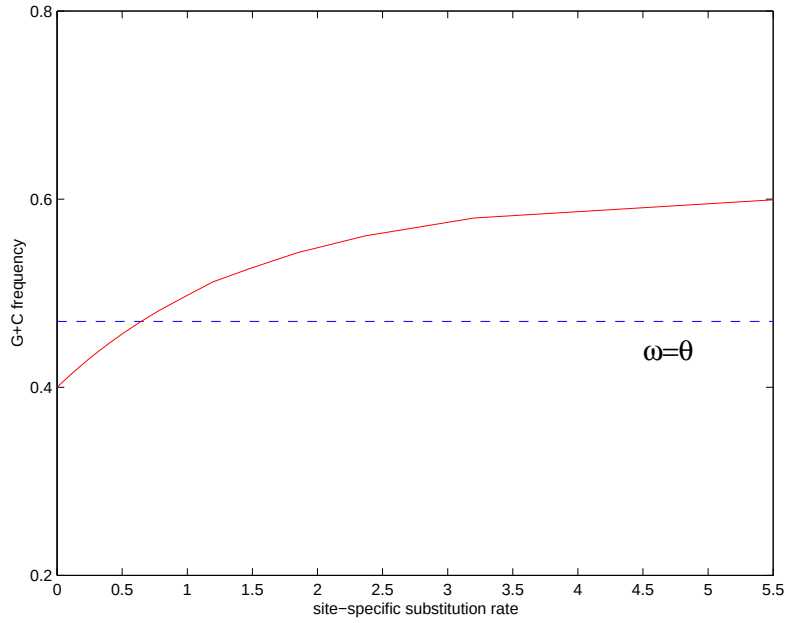
- Time-heterogeneous methods were designed (Galtier and Gouy, 1998).
- The G+C content of the rRNA genes of LUCA was estimated by Galtier et al. (Science, 99).
- A low G+C content in ancestral rRNAs did not support a hyperthermophilic origin of life.
- But time-heterogeneous methods are extremely sensitive to the violation of spatial homogeneity.

Bias with nonhomogeneous methods



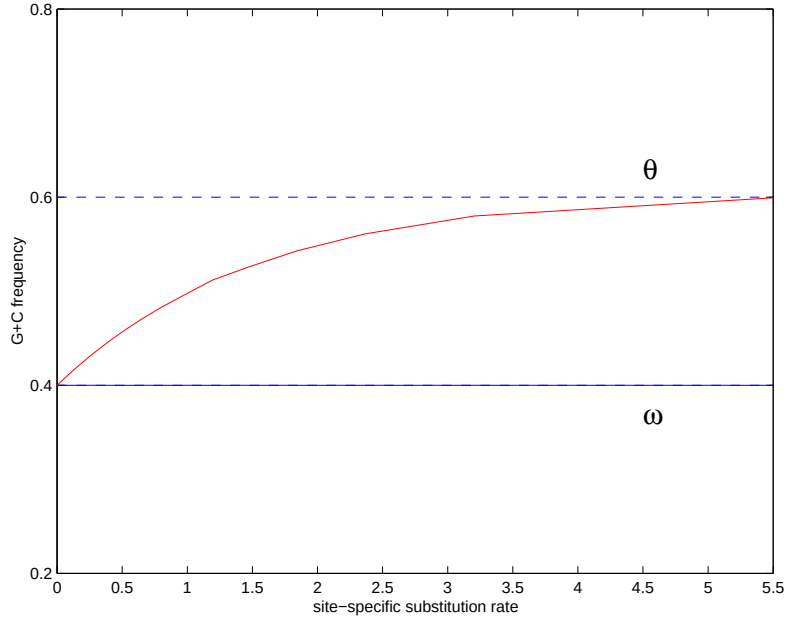
frequency profile for sequences generated by a stationary time-homogeneous model heterogeneous across sites

Bias with nonhomogeneous methods



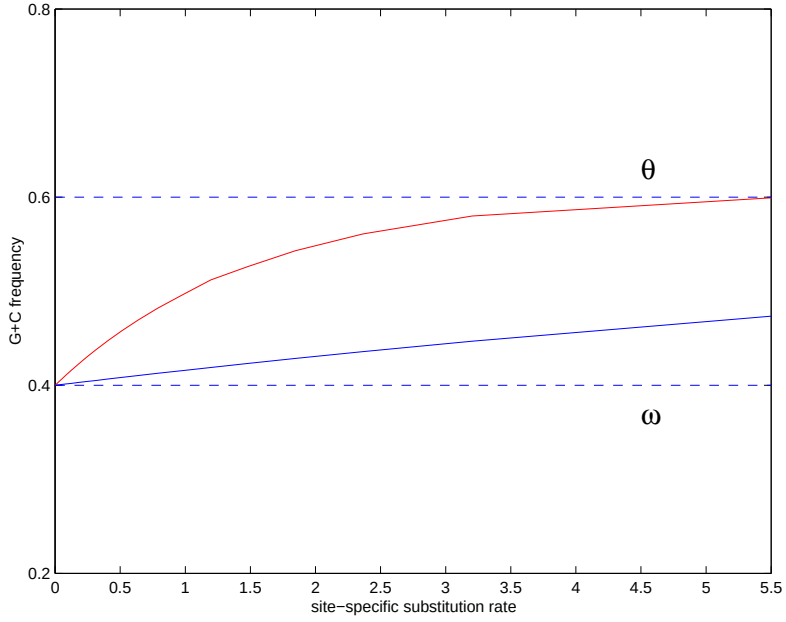
the expected output of a time-heterogeneous model that accounts for rate-heterogeneity across sites but not for compositional heterogeneity across sites

Bias with nonhomogeneous methods



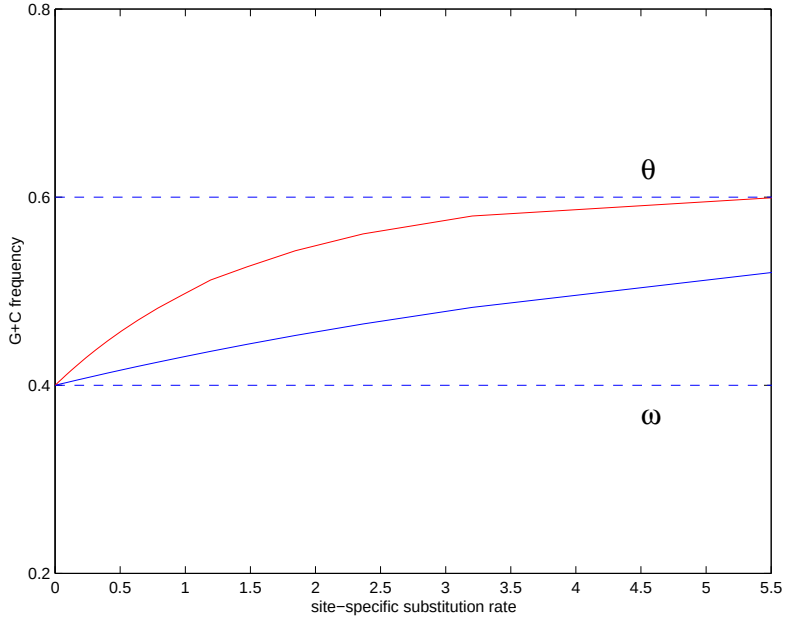
a possible output

Bias with nonhomogeneous methods



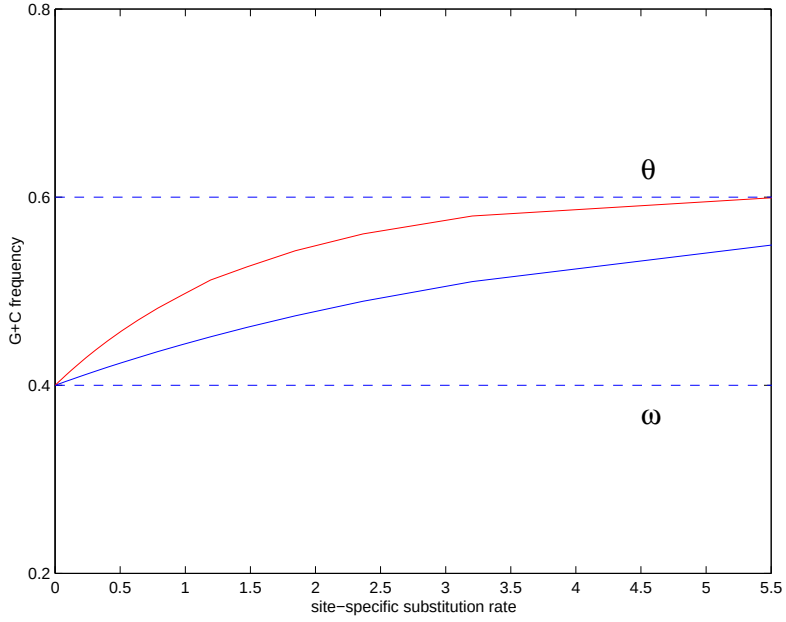
t=0.05

Bias with nonhomogeneous methods



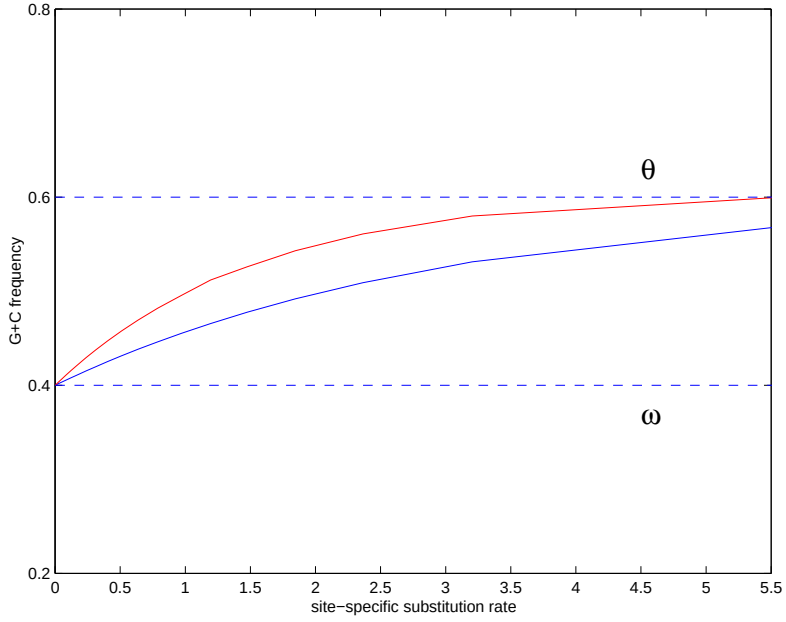
t=0.10

Bias with nonhomogeneous methods



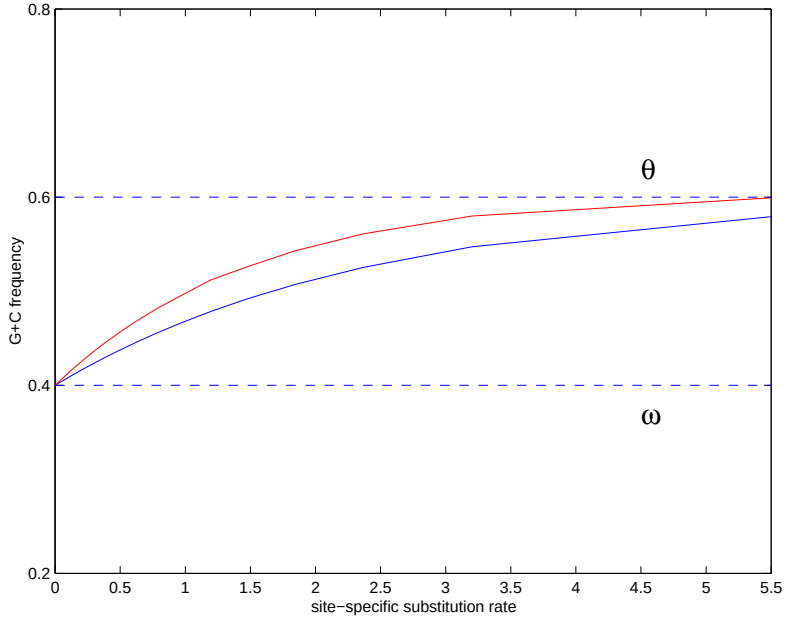
t=0.15

Bias with nonhomogeneous methods



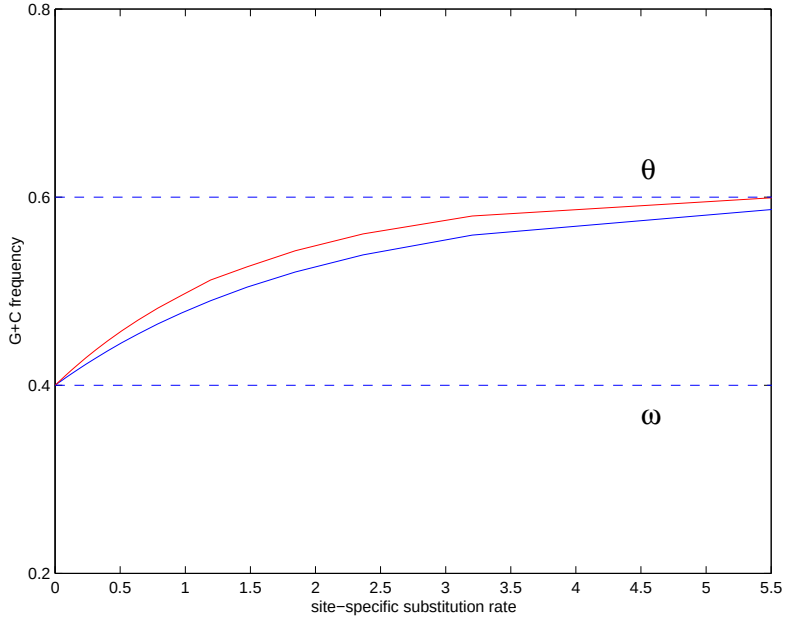
t=0.20

Bias with nonhomogeneous methods



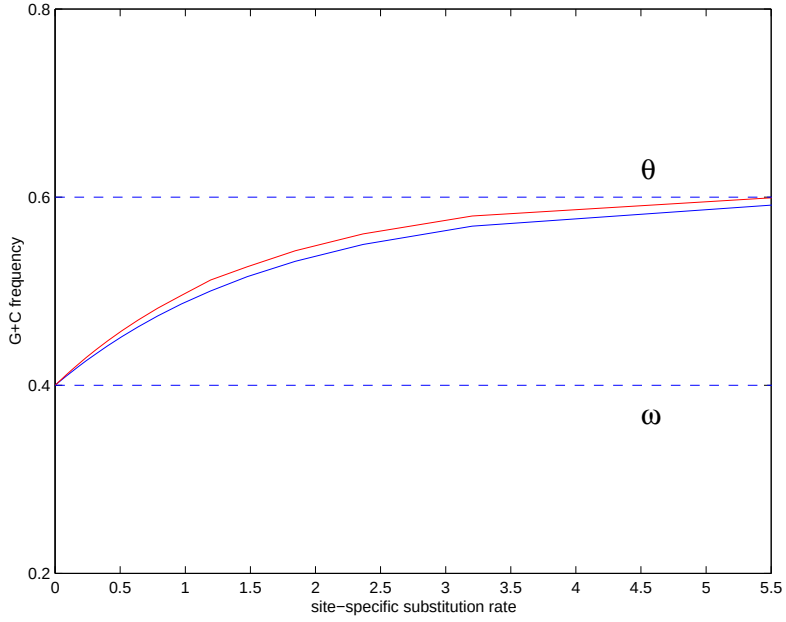
t=0.25

Bias with nonhomogeneous methods



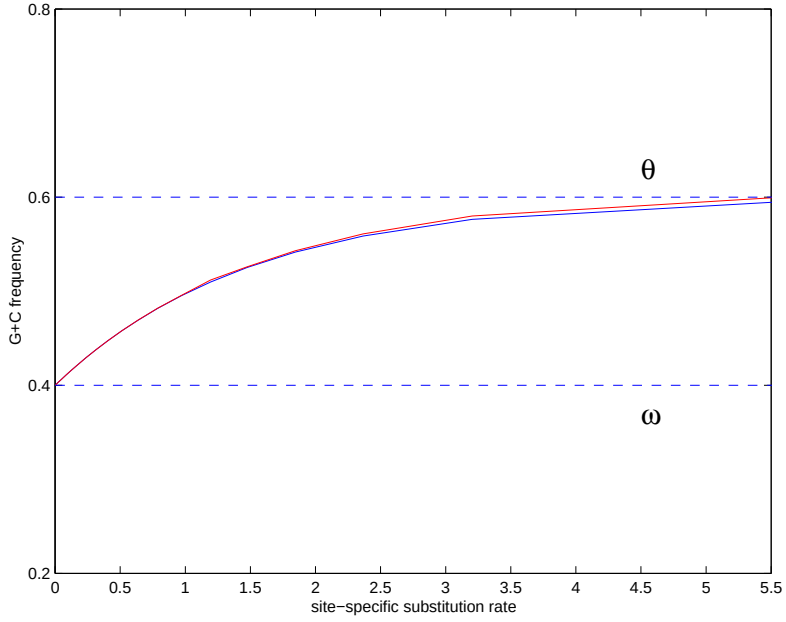
t=0.30

Bias with nonhomogeneous methods



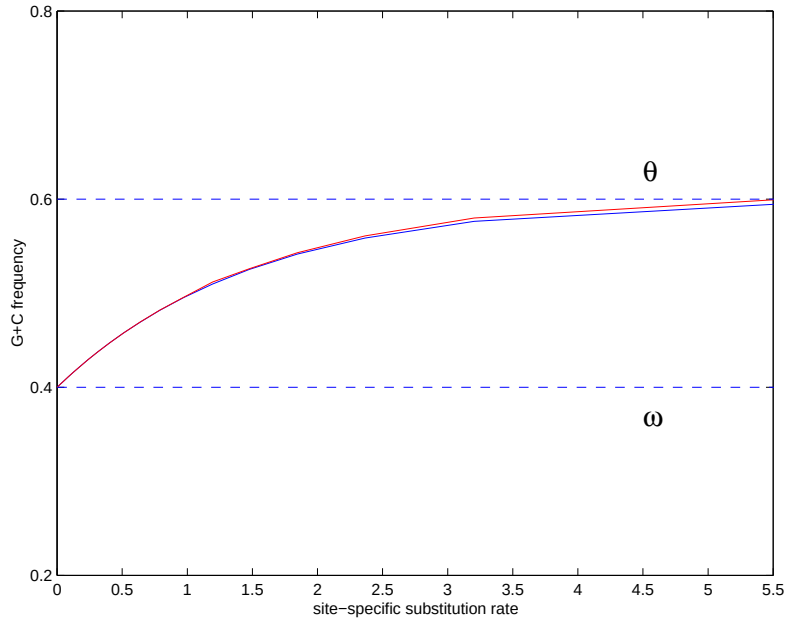
t=0.35

Bias with nonhomogeneous methods



t=0.40

Bias with nonhomogeneous methods

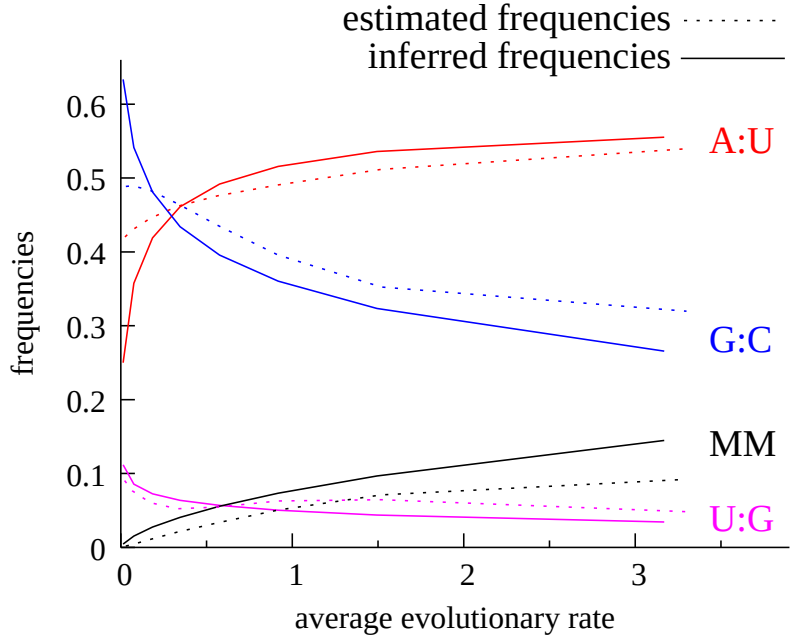


Ancestral frequency estimates are biased towards frequencies observed at slow evolving sites.

Galtier et al. probably underestimated the ancestral G+C content in their study.

A space-time heterogeneous model

- A mixture model was built to account for spatial compositional variation.
- A different frequency vector is used for each rate category.

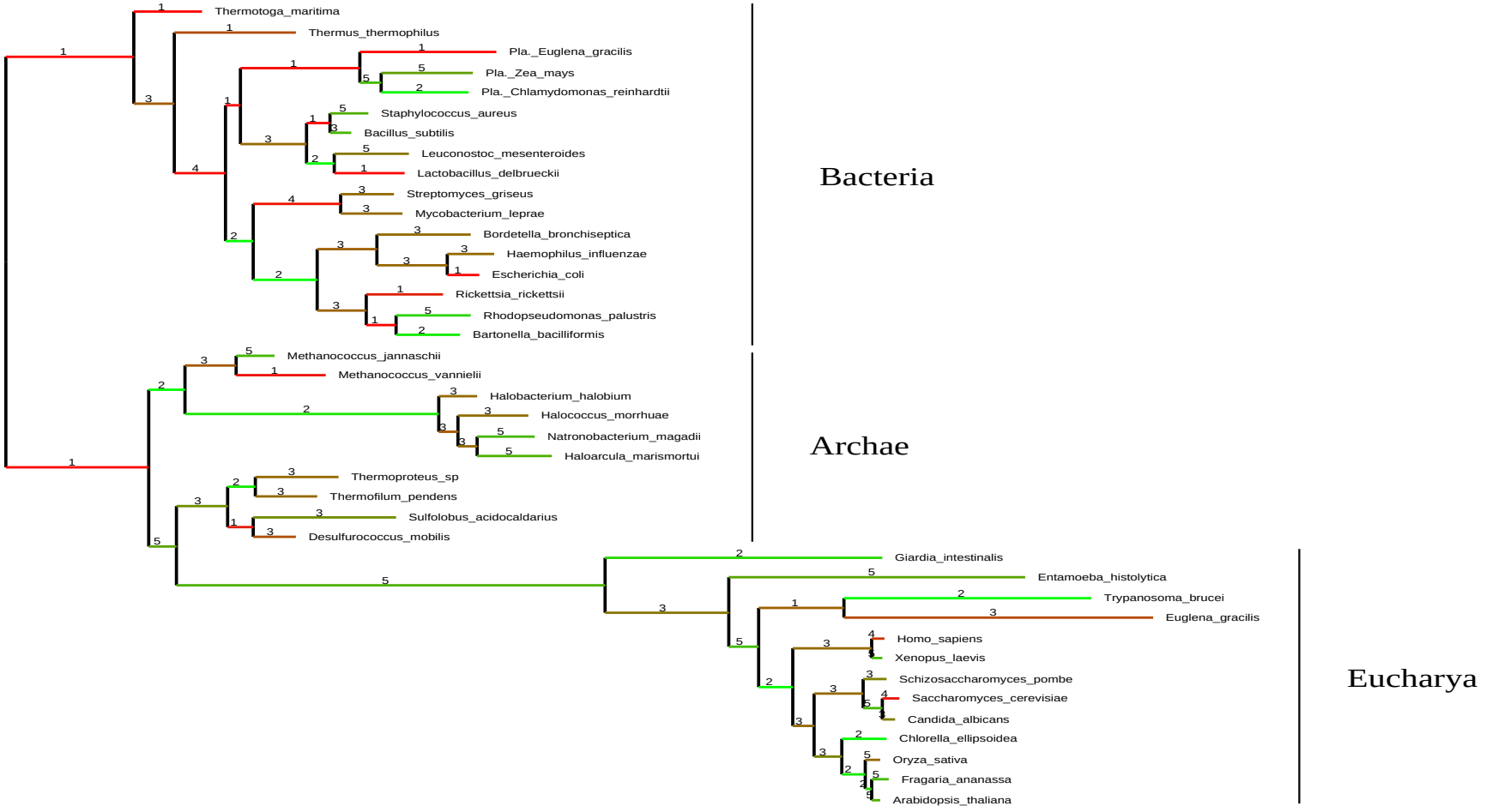


(primates rRNAs dataset shown before).

A space-time heterogeneous model

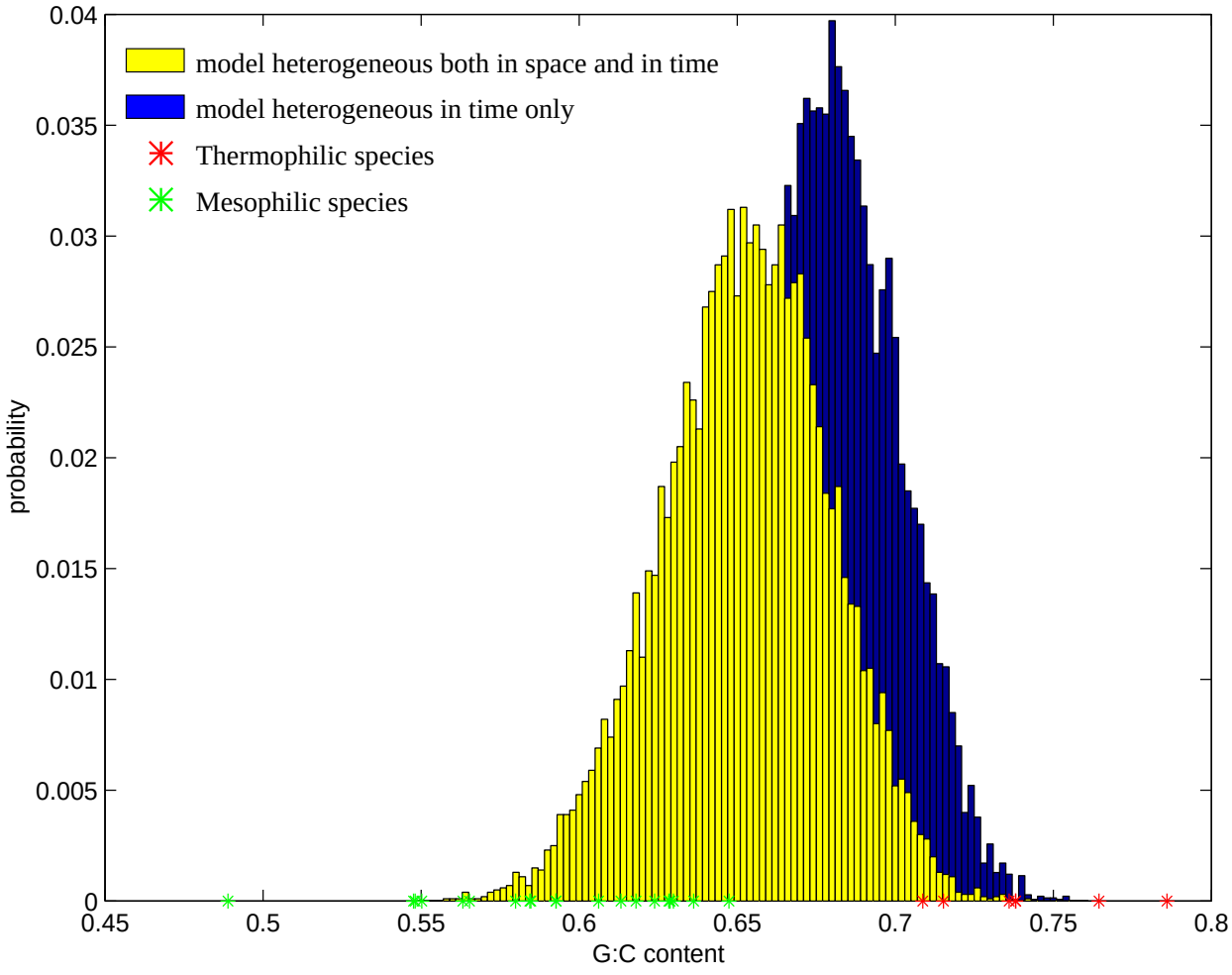
- This mixture model was combined with a time-heterogeneous model.
- Equilibrium frequencies used at each branch are selected among a pool of composition parameters (Foster, 2004).

A glimpse into the past?



Consensus tree obtained with a LSU+SSU dataset (European ribosomal RNA database) analyzed with the space-time heterogeneous model. Branches are colored according to the mean posterior estimates of substitution model parameters.

A nonhyperthermophilic origin of life?



We can compare the inferred ancestral G:C content with the G:C content of contemporary mesophilic and thermophilic prokaryots.

Future work

- Test the performance of the model with synthetic data.
- Increase taxon sampling.
- Substitution models that can **explain** the variation of frequencies across sites.
- Models of RNA secondary structure evolution.
- A phylogenetic software: http://www.bioinf.man.ac.uk/resources/phase