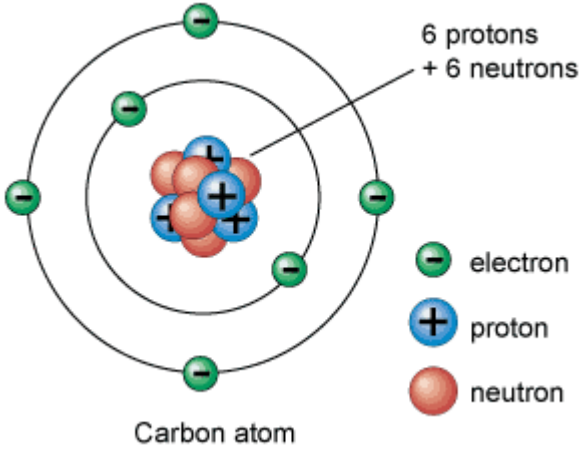
WORKSHOP IN ADVANCED BAYESIAN PHYLOGENETICS

NOV 17-21, 2014, ADELAIDE

- WHAT'S IN A MODEL?
- CONTINUOUS-TIME MARKOV MODELS
  - DNA SUBSTITUTIONS
  - AMINO ACIDS
  - PHENOTYPIC CHARACTERS
  - AMONG-SITE RATE VARIATION
- CONTINUOUS CHARACTERS
- (OVER)PARAMETERIZATION

WHAT'S IN A "MODEL"?

WHAT'S IN A "MODEL"?



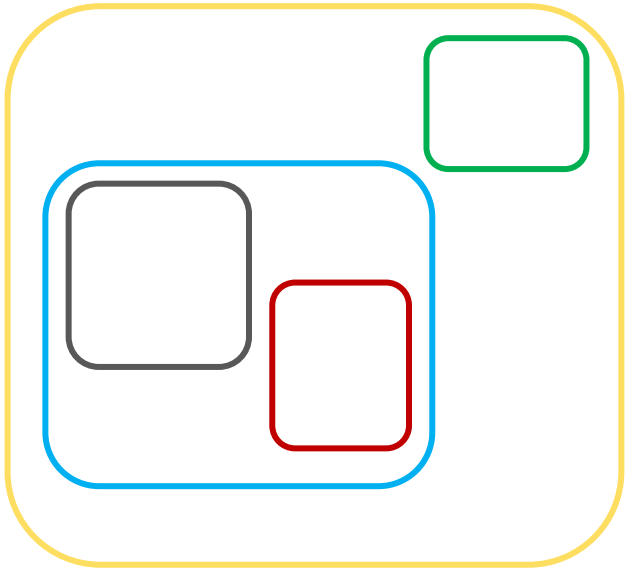
WHAT'S IN A "MODEL"?

stochastic **model**

- stochastic representation of process in nature
- simplification
- models as tools / crutches in biology
- strength and weakness
  - necessary assumptions to make
  - result might be model dependent
  - very powerful!

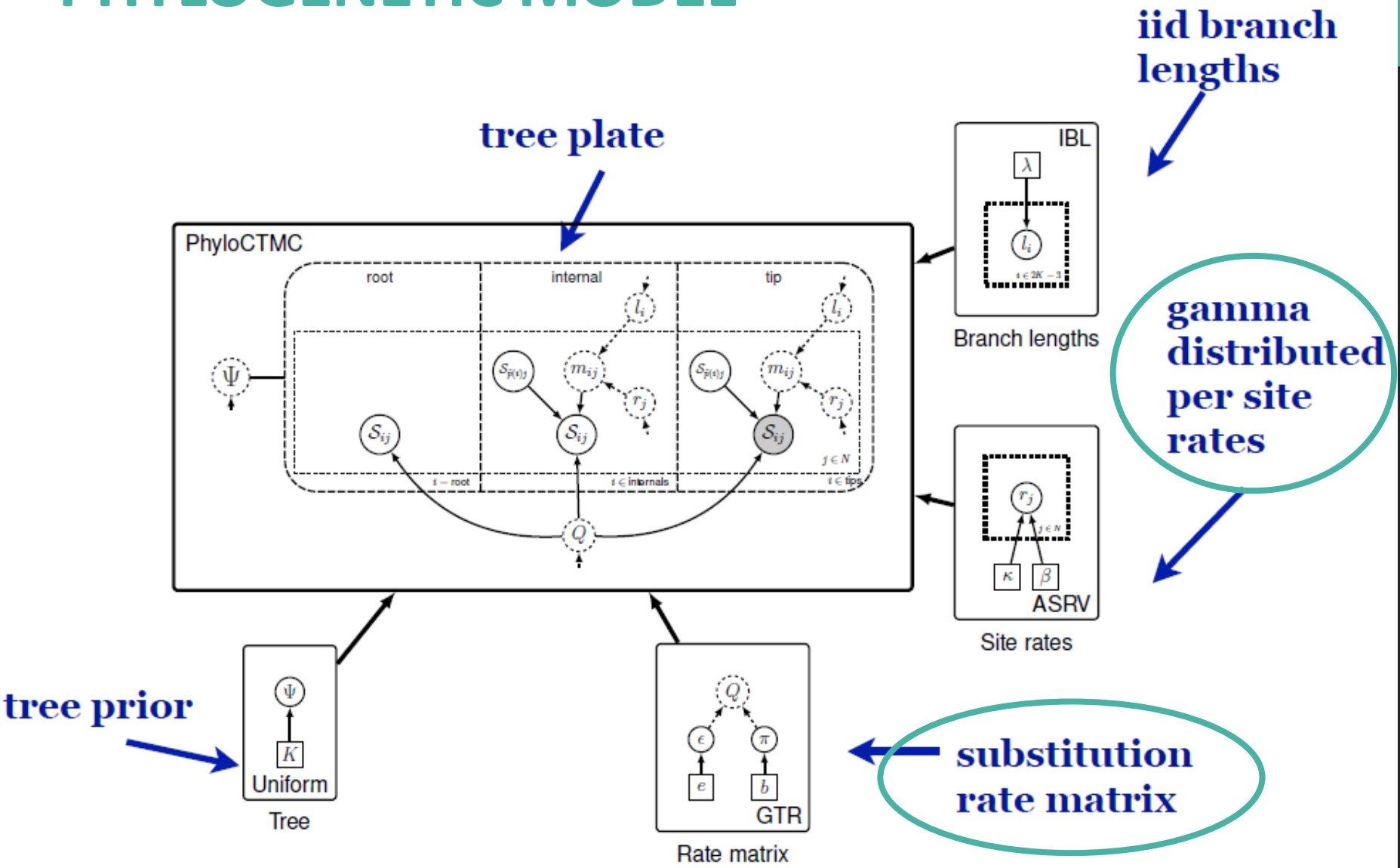
WHAT'S IN A "MODEL"?

stochastic **model**



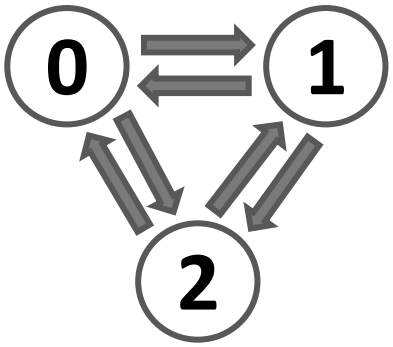
- each model has parameters
  - e.g., DNA substitution model:
    - base frequencies
    - exchangeability rates
  - phylogenetic model: tree
- model-parts can be combined
  - e.g., substitution model with relaxed-clock model
- models are always wrong! (except in simulations)

PHYLOGENETIC MODEL



MARKOV MODELS

general Markov model



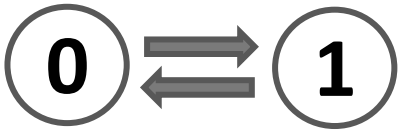
- states, probabilities of change
- memory-less process
- finite state space: e.g. nucleotides
- infinite state space: e.g. phylogenetic MCMC

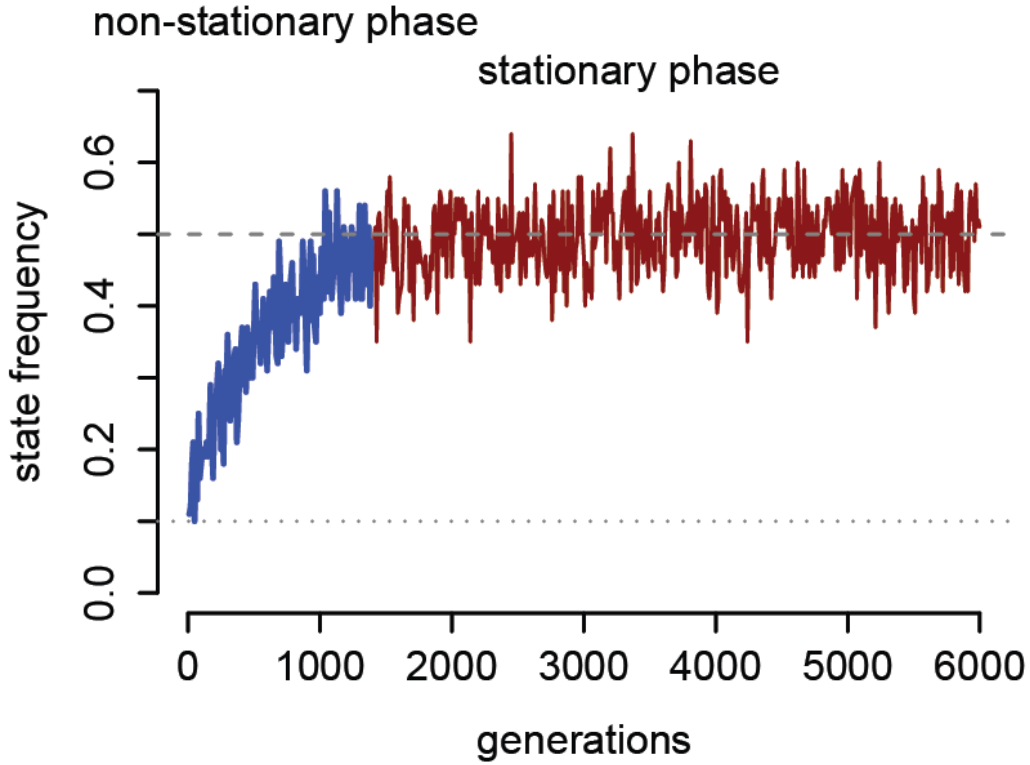
common simplifications

- stationarity
- homogeneity
- time-reversibility

MARKOV MODELS

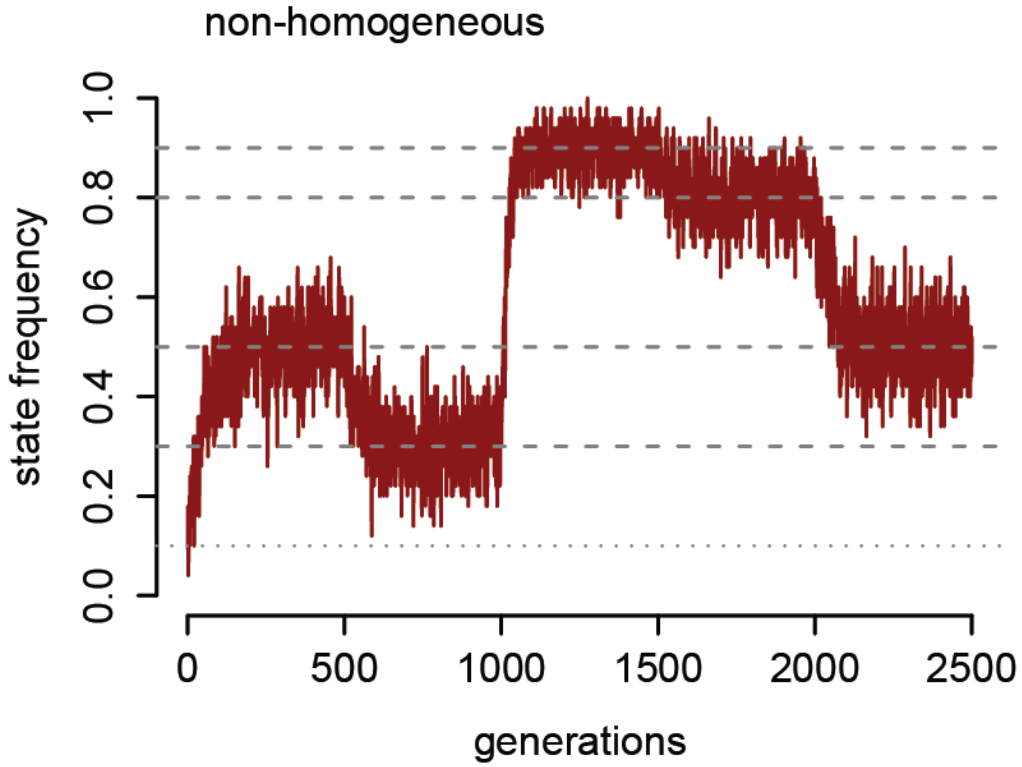
stationarity: process is at equilibrium





MARKOV MODELS

homogeneity: process is homogeneous over time

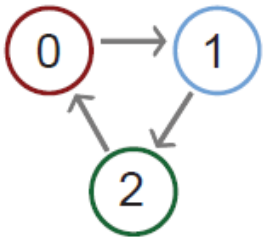


MARKOV MODELS

Time-reversibility: step forward as likely as backward

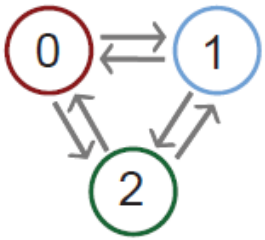
$$\pi_i * p_{ij} = \pi_j * p_{ji}$$

time irreversible

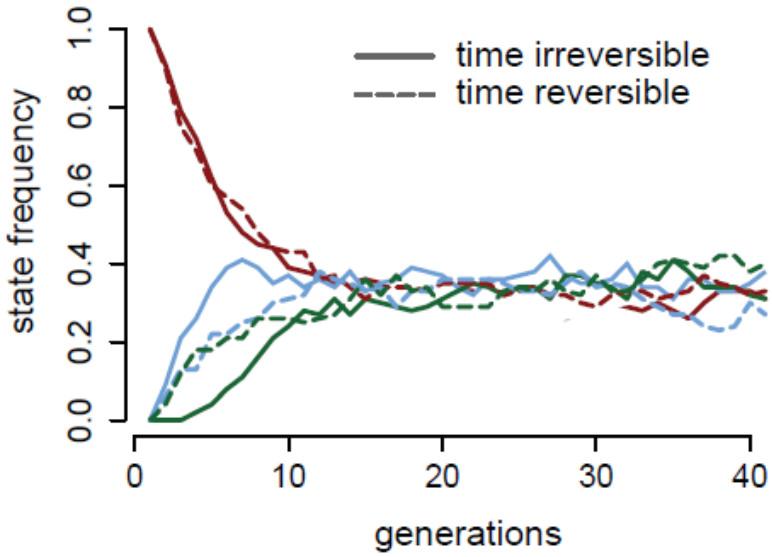


	0	1	2
0	-	r	0
1	0	-	r
2	r	0	-

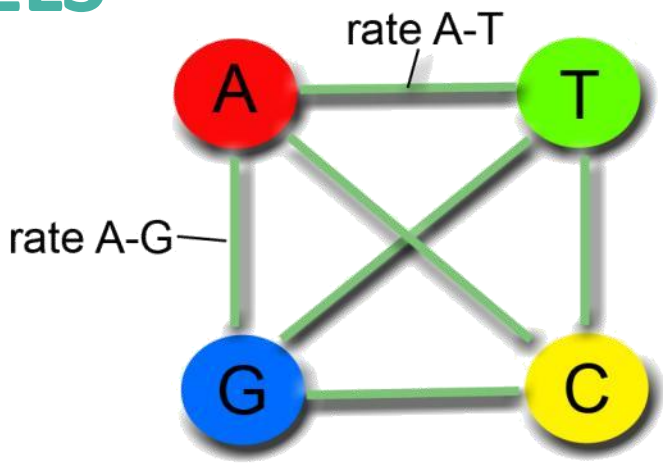
time reversible



	0	1	2
0	-	r	r
1	r	-	r
2	r	r	-

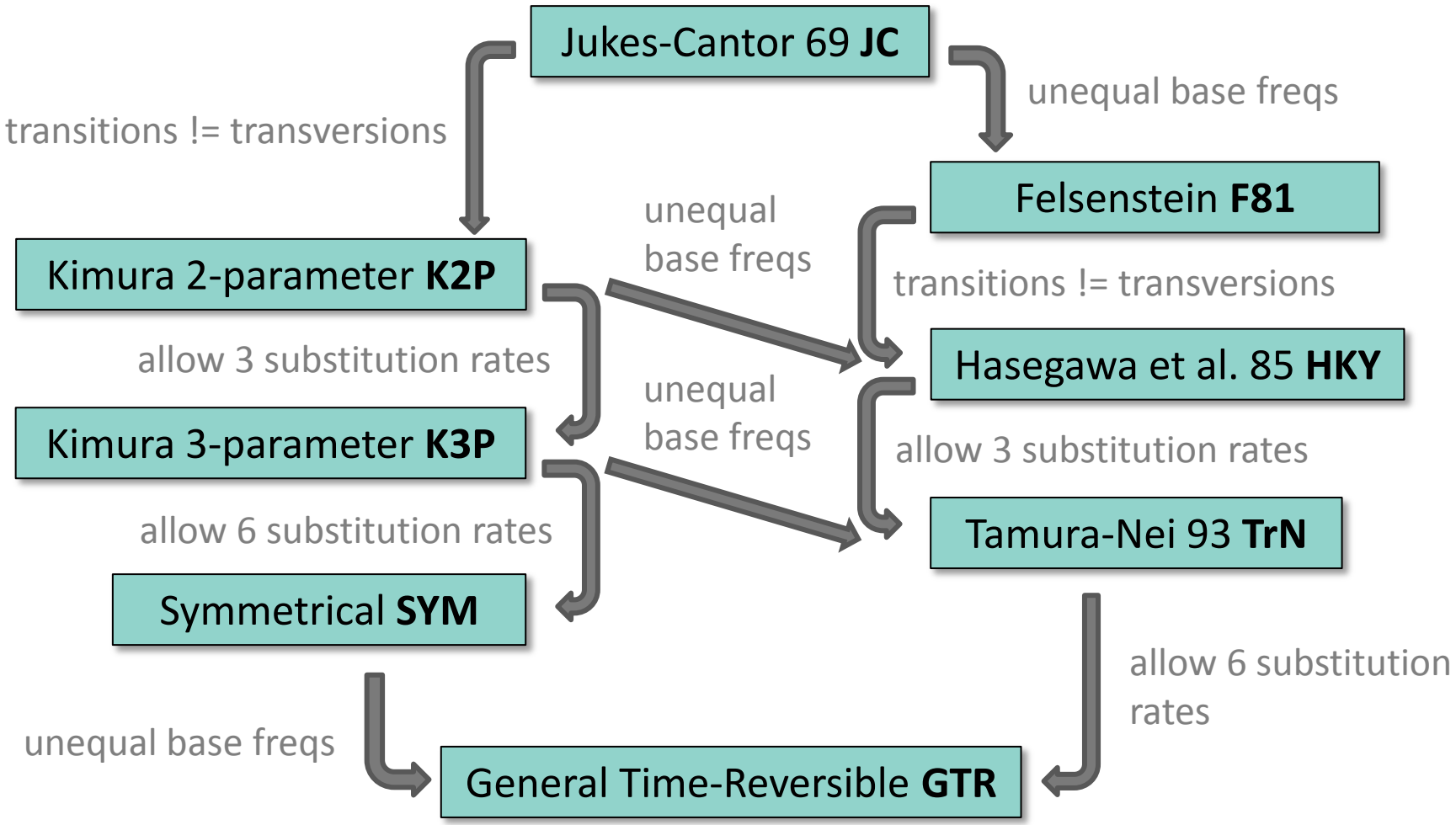


DNA SUBSTITUTION MODELS



- 4-state Markov model
- memoryless stochastic process
- parameters:
  - stationary frequencies of the four nucleotides
  - exchangeability rates of the 6 (12) possible transitions
- usual assumptions: stationarity, homogeneity, time-reversibility (GTR subspace)
- site independence
- (can be combined with among-site rate variation models)

DNA SUBSTITUTION MODELS



DNA SUBSTITUTION MODELS

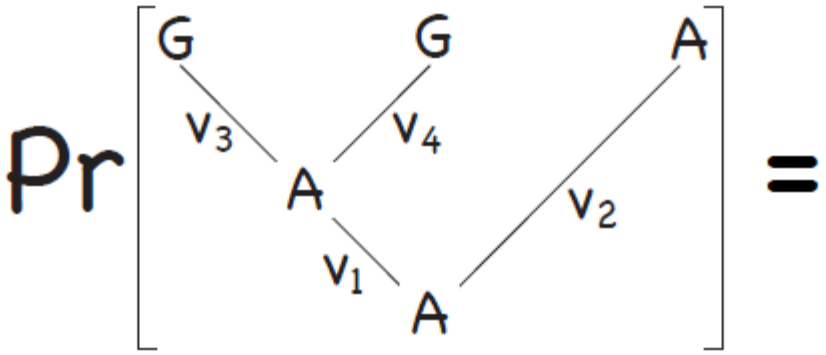
- models from GTR subspace are nested
- number of free parameters of substitution models?
  - JC: 0!
  - K2P: 2
  - K3P: 3
  - SYM: 5
  - F81: 3
  - HKY: 4
  - TrN: 6
  - GTR: 8
- likelihood approaches to test models: jModeltest, MrModeltest, PartitionFinder
- Bayesian: Bayes factor tests, reversible jump over whole model space (MrBayes)

MARKOV MODELS

calculating tree likelihood under Markov model

- continuous-time
- sites are independent
- state frequencies & exchangeability rates -> 'Q matrix'

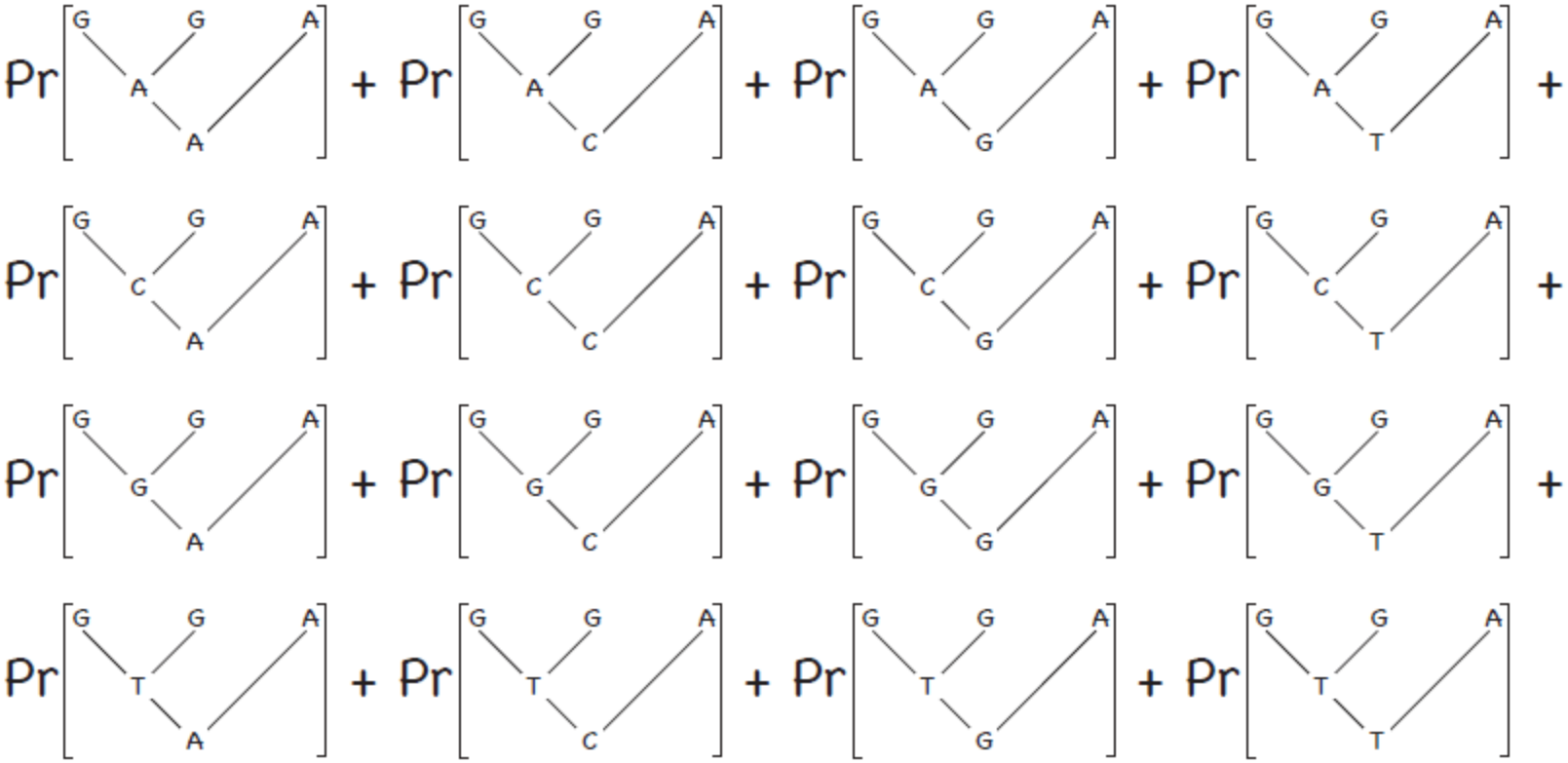
- integrating over all possible ancestral states



$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3) \times p_{AG}(v_4)$$

π_i – Stationary frequencies

$p_{ij}(v)$ – Transition probabilities


Pr G G A A + Pr G G A C + Pr G G A G + Pr G G A T +
Pr G G C A + Pr G G C C + Pr G G C G + Pr G G C T +
Pr G G G A + Pr G G G C + Pr G G G G + Pr G G G T +
Pr G G T A + Pr G G T C + Pr G G T G + Pr G G T T

AMINO-ACID SUBSTITUTION MODELS

parameter space **of AA models**

- 20 different amino acids

→ full GTR model has 19 free parameters for the state frequencies and 189 free parameters for exchangeability rates!

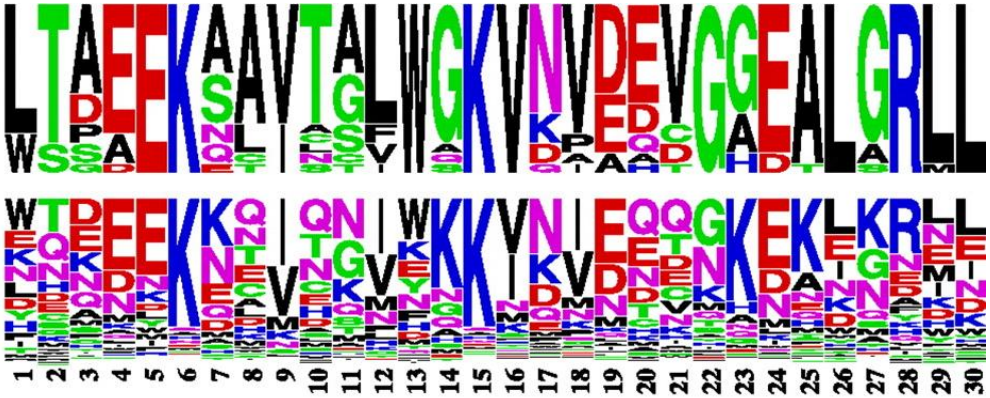
AMINO-ACID SUBSTITUTION MODELS

- fixed-rate models
  - Poisson model: JC for AA (all rates / state freqs equal)
  - empirical models: “trained” on large datasets (Dayhoff, mtRev, mtMam, WAG, Blosum62, etc.)
- variable-rate models
  - Equalin model: rates equal, state freqs estimated
  - GTR model: all 189 + 19 parameters estimated
- mixture models
  - CAT model (Lartillot & Philippe 2004)

AMINO-ACID SUBSTITUTION MODELS

CAT model

- mixture model
- accounting for site-specific amino-acid preferences
- implementation: PhyloBayes



MODELING PHENOTYPIC CHARACTERS

Antennae

22. Male antennae: flagellomeres without lateral projections = 0; flagellomeres with distinct, slender lateral projections many times longer than the base of the flagellomeres, projections not flattened and appressed = 1; flagellomeres with flattened and appressed lateral projections = 2 (unordered).

23. First flagellomeres (Vilhelmsen, 1997a: 3; Ronquist et al., 1999: 14): not broader, and not much longer than the length of any of the following flagellomeres = 0; distinctly broader and much longer than any of the following flagellomeres = 1; distinctly enlargened, distal flagellomeres absent = 2 (ordered).

24. Apical flagellomeres: not conspicuously modified = 0; clubshaped = 1; reduced and with flattened apex = 2 (unordered). The last character state was newly added. Scored as inapplicable when only one flagellomere was present.

25. Multiporous plate sensilla (Basibuyuk & Quicke, 1999): absent = 0; present = 1.

MODELING PHENOTYPIC CHARACTERS

Why?

- use characters for phylogenetic inference
  - morphological tree
  - morphology to place fossils
  - characters from genome „morphology“
- study character evolution
  - reconstruct ancestral states
  - study correlated evolution (comparative method)

MODELING PHENOTYPIC CHARACTERS

Difficulties

- how to encode phenotypic characters?
  - define characters
  - define character states
- model of evolution?
- different character types: discrete and continuous

MODELING PHENOTYPIC CHARACTERS

differences to nucleotides / AA:

- state names are arbitrary
- differing number of states per character
- no constant characters sampled!

  -> account for it in likelihood function

Mk MODEL

-> extend the JC model for an arbitrary number of states: Mk model (Lewis, 2001)

- k= number of states
- equal frequencies of 1/k
- equal rates between them -> transition probabilities = $1/k - 1/k*e^{-kt}$
- can in principle be applied to any type of discrete characters

Mk MODEL

extentions possible:

- relax equal-frequency assumption: allow for asymmetry in stationary state frequencies according to a symetric Dirichlet/Beta distribution
- among-site rate varition, same way as for nucleotides

DISCUSSION MARKOV MODELS

- For DNA sequences, which of the 3 standard assumptions is the violated most often?
  - stationarity
  - homogeneity
  - time-reversibility
  - site independence

DISCUSSION MARKOV MODELS

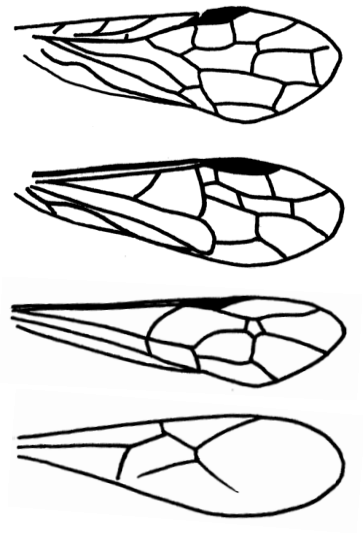
- Is Markov-Chain Monte Carlo (MCMC) algorithm
  - stationary?
  - homogeneous?
  - time-reversible?
- What is the main difference to substitution models?

NON-STANDARD MARKOV MODELS

GTR assumes stationarity & homogeneity

- what if progressive evolution?

→ different state frequencies at the root

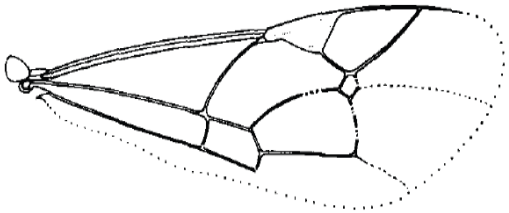


Fredrik Ronquist,
Stockholm

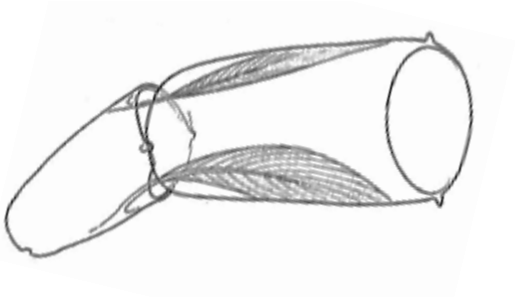
Lars Vilhelmsen,
Copenhagen

HYMENOPTERAN MORPHOLOGY

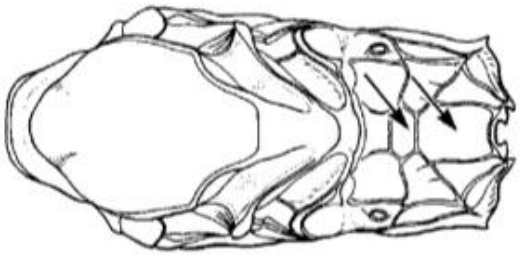
a) wing veins: present or absent [23]

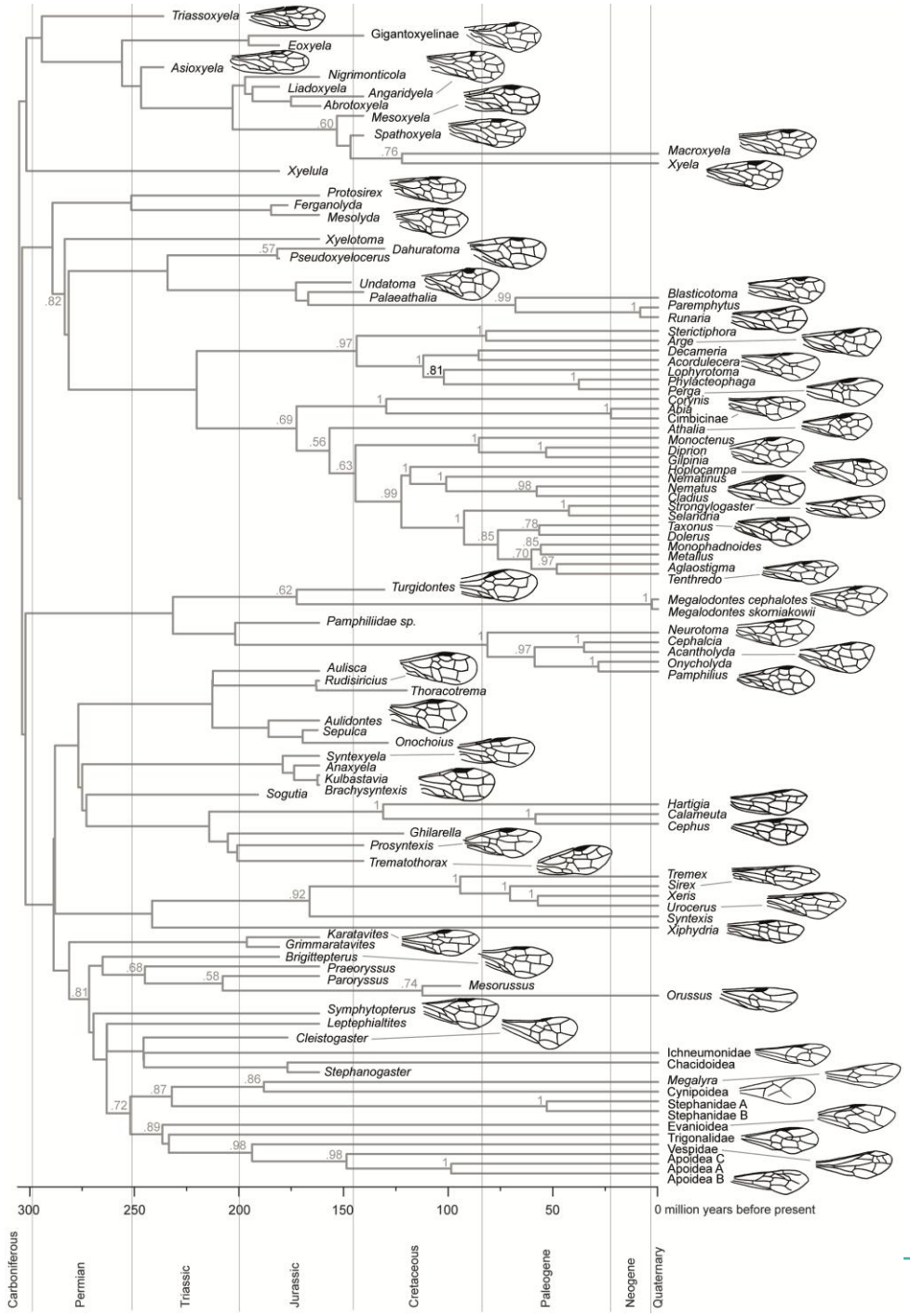


b) muscles: present or absent/fused [56]



c) sclerites: separate or absent/fused [67]

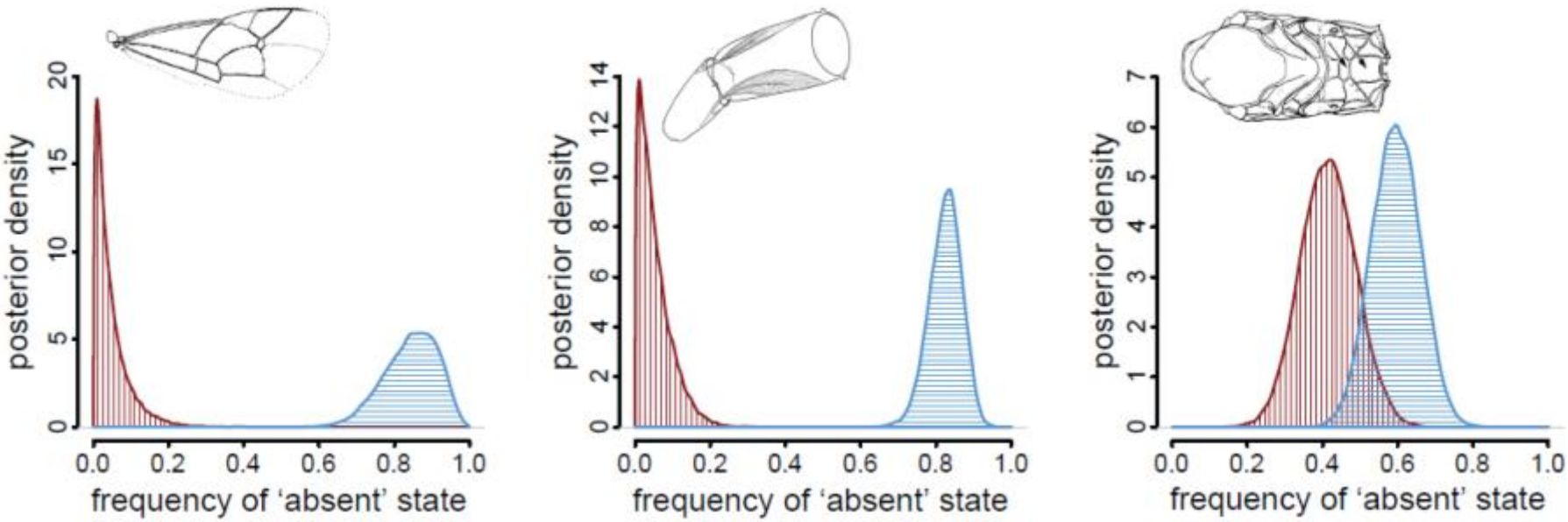




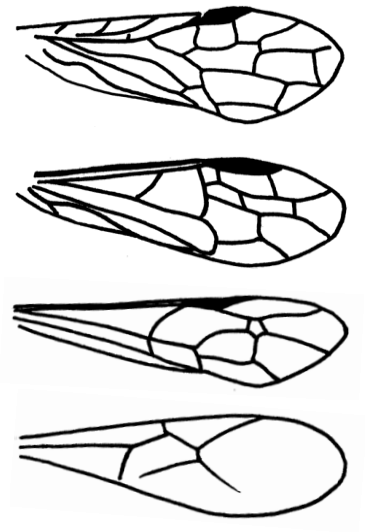
- morphology only
- adding 7 genes
- adding fossil and time information

HYMENOPTERAN MORPHOLOGY

- frequency of state “absent” at root
- frequency of state “absent” at equilibrium



NON-STANDARD MARKOV MODELS



GTR assumes stationarity & homogeneity

- what if progressive evolution?

→ different state frequencies at the root

- what if nucleotide composition differs among terminals?

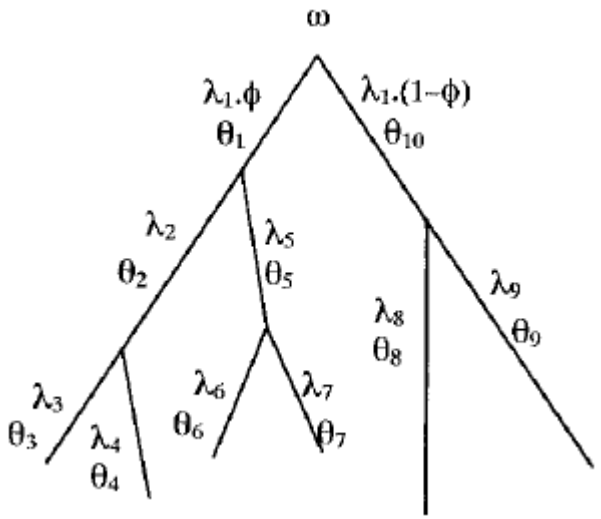
→ non-homogeneous models

NON-STANDARD MARKOV MODELS

GC content changes over the tree

(Galtier & Gouy 1998, MBE)

parameters	symbol	number
ancestral G+C %	ω	1
branch lengths	λ_i	$2n - 3$
root location	ϕ	1
Ts/Tv ratio	κ	1
equilibrium G+C %	θ_i	$2n - 2$
		$4n - 2$

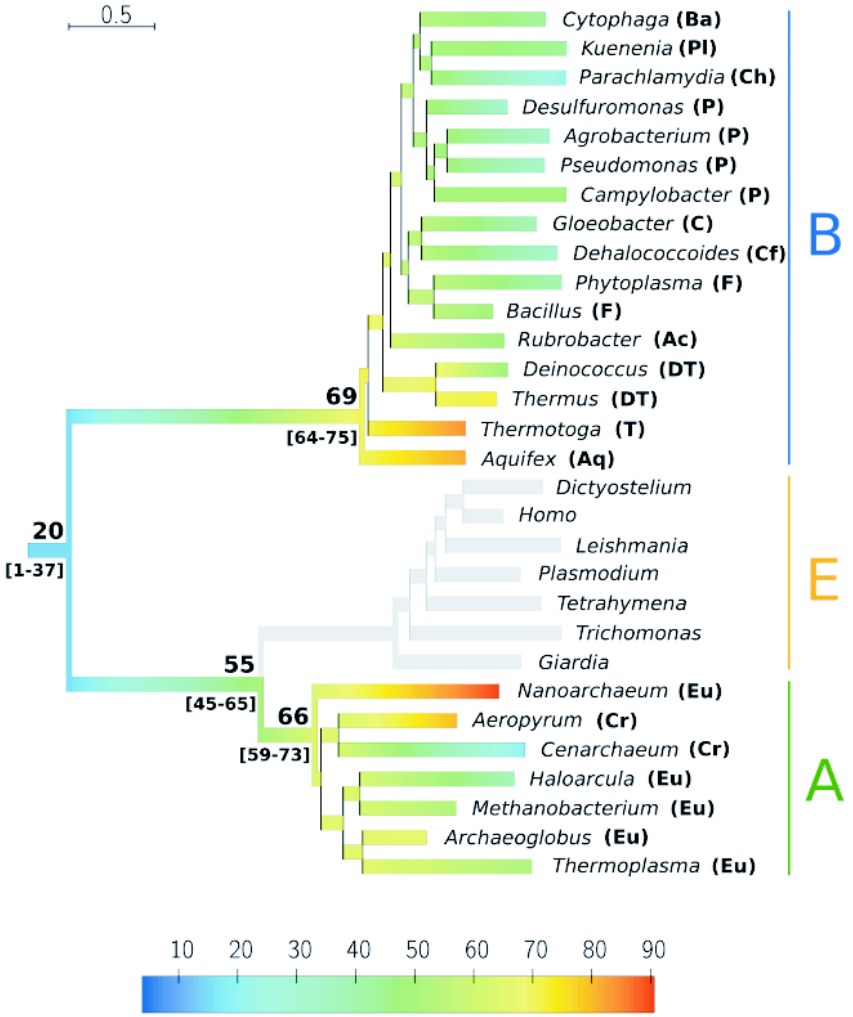


NON-STANDARD MARKOV MODELS

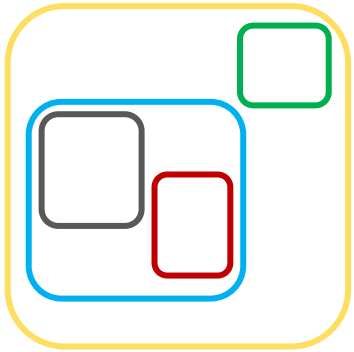
Bacteria: GC content correlates with optimal growth temperature

→inferring environment of the root of the tree of life

(Bousseau ea, 2008)

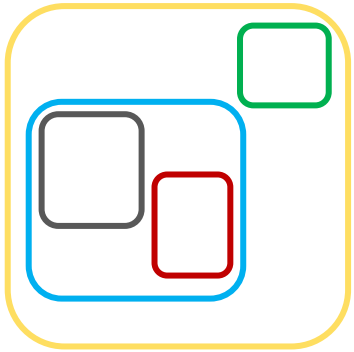


AMONG-SITE RATE VARIATION



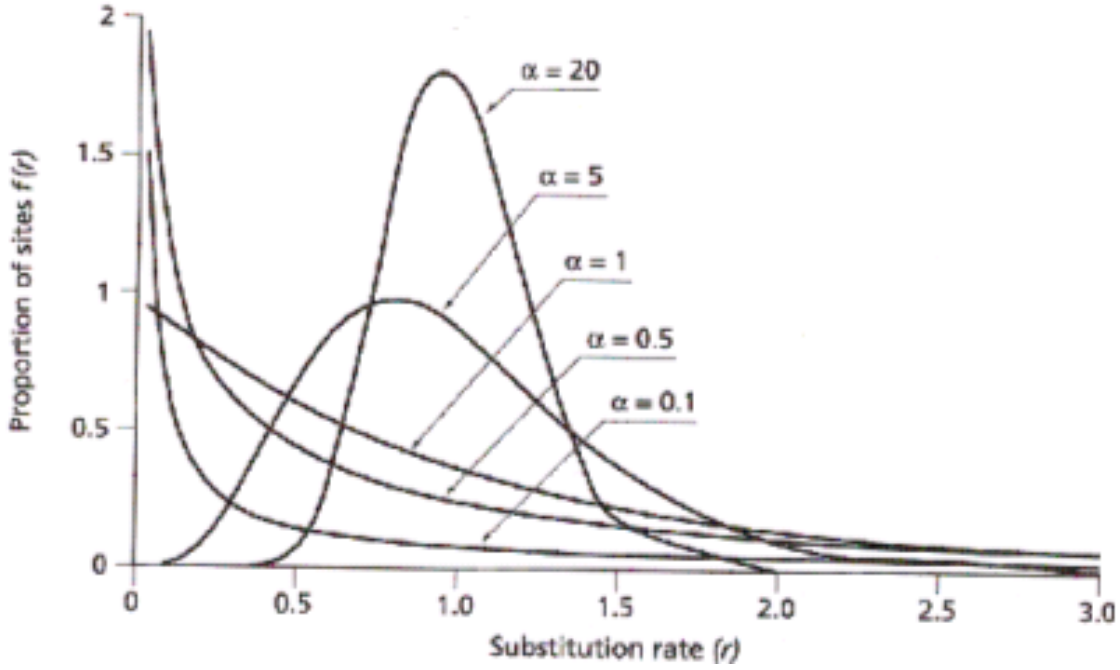
sites evolving at different rates

AMONG-SITE RATE VARIATION



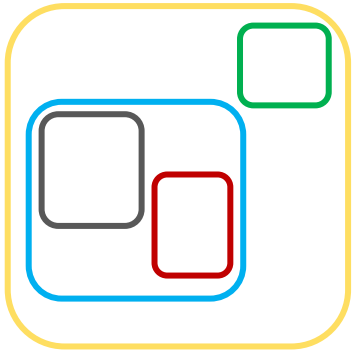
sites evolving at different rates

- discretized gamma-distribution



Yang 1994

AMONG-SITE RATE VARIATION

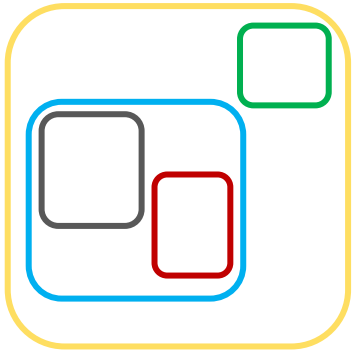


sites evolving at different rates

- discretized gamma-distribution
- proportion of invariant sites
    - potential interaction with gamma!
- mixture models
- partitioning

- usually more important than GTR submodel!

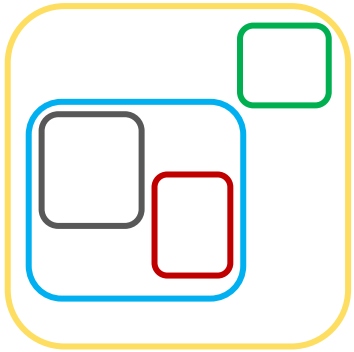
AMONG-SITE RATE VARIATION



partitioning

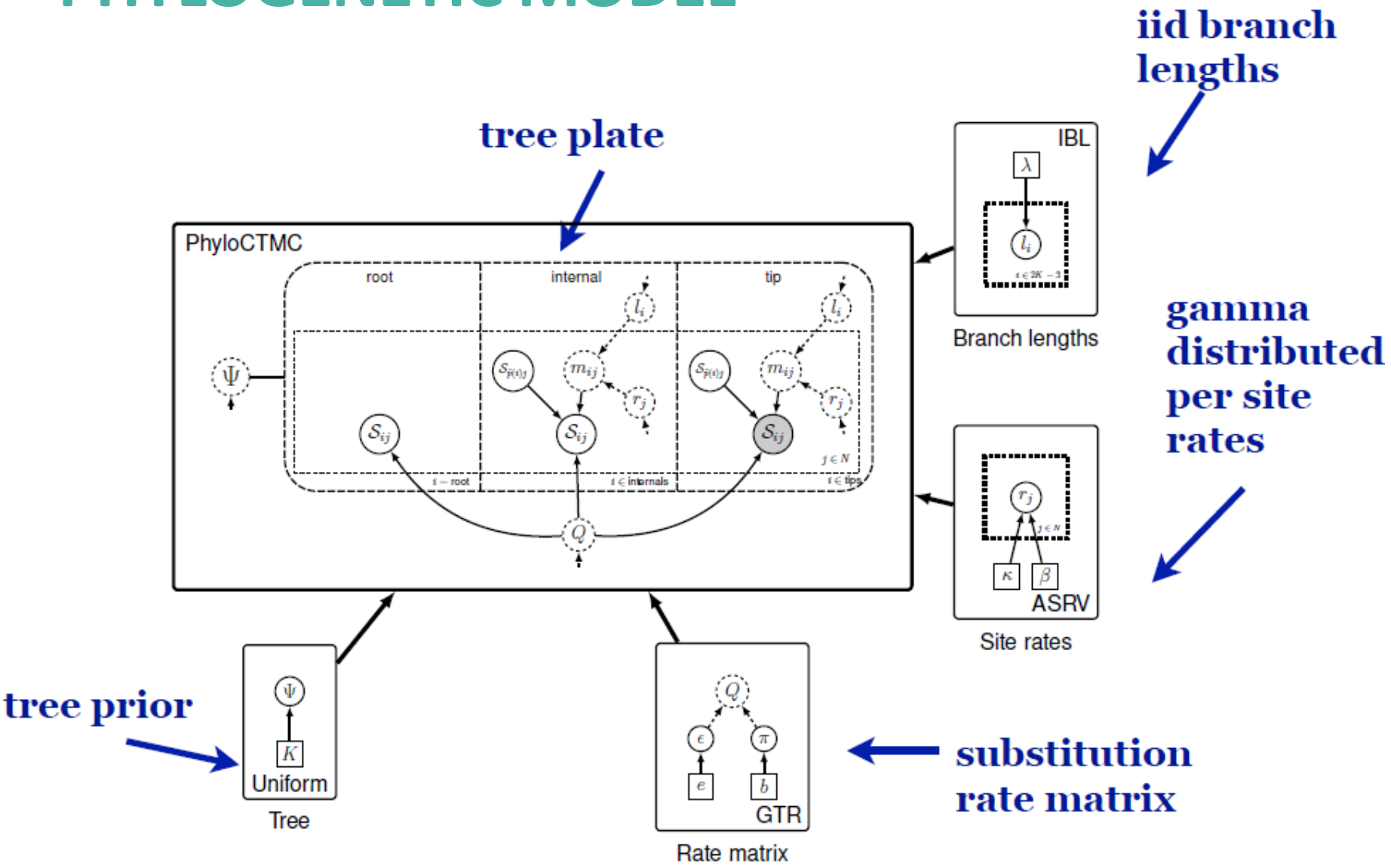
- according to molecular evolution
  - datatype (morphology versus DNA)
  - gene origin (e.g., mtDNA versus nDNA)
  - gene
  - codon position
  - amino acid sequences: functional domains
  - PartitionFinder: trying submodels
- automatized: mixture model

AMONG-SITE RATE VARIATION

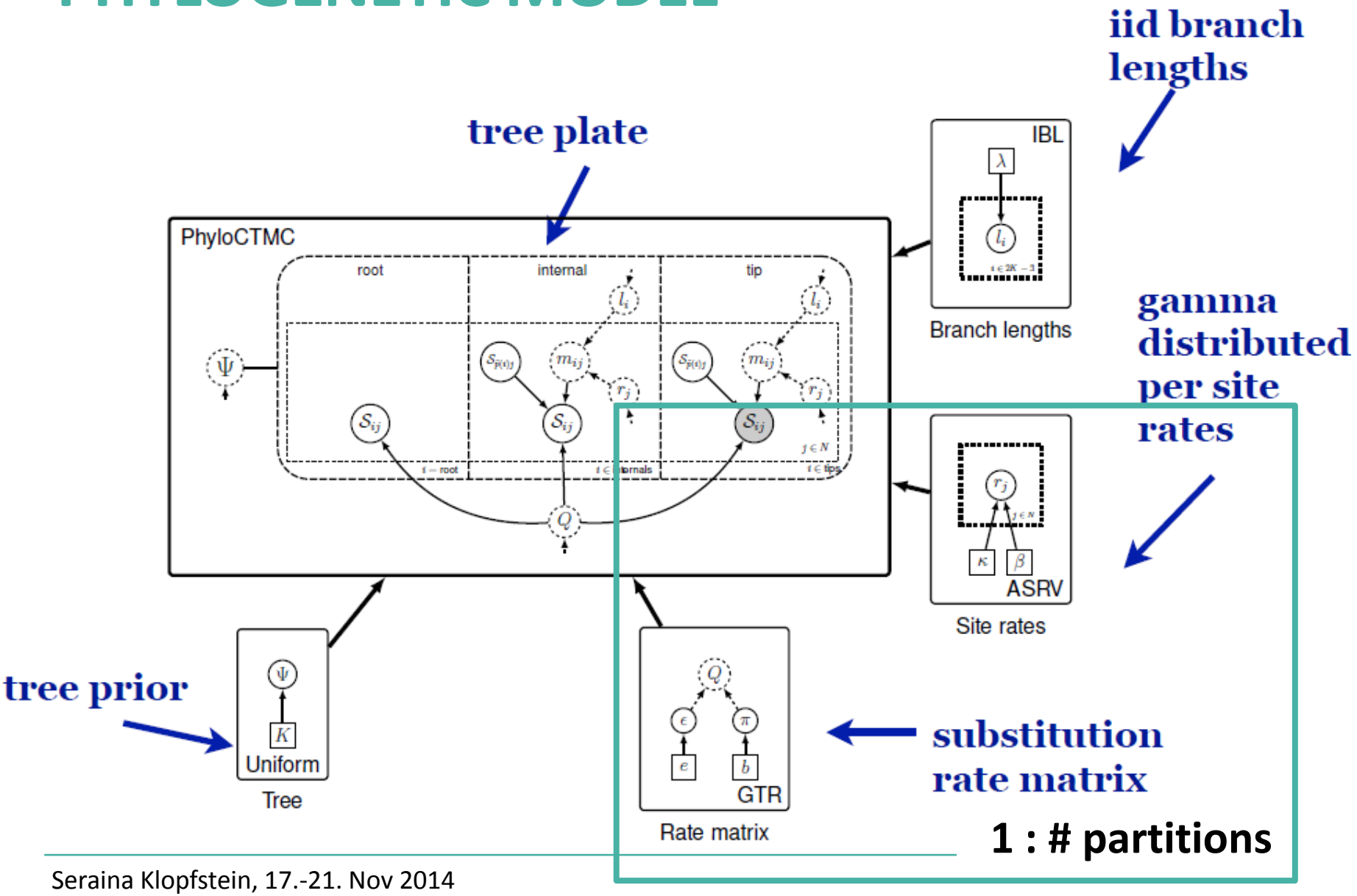


what do we partition?

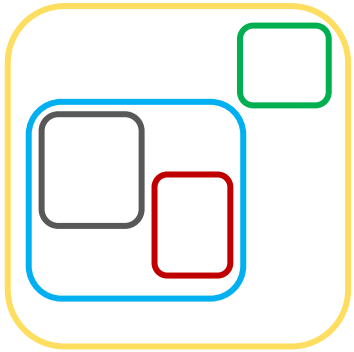
PHYLOGENETIC MODEL



PHYLOGENETIC MODEL



AMONG-SITE RATE VARIATION



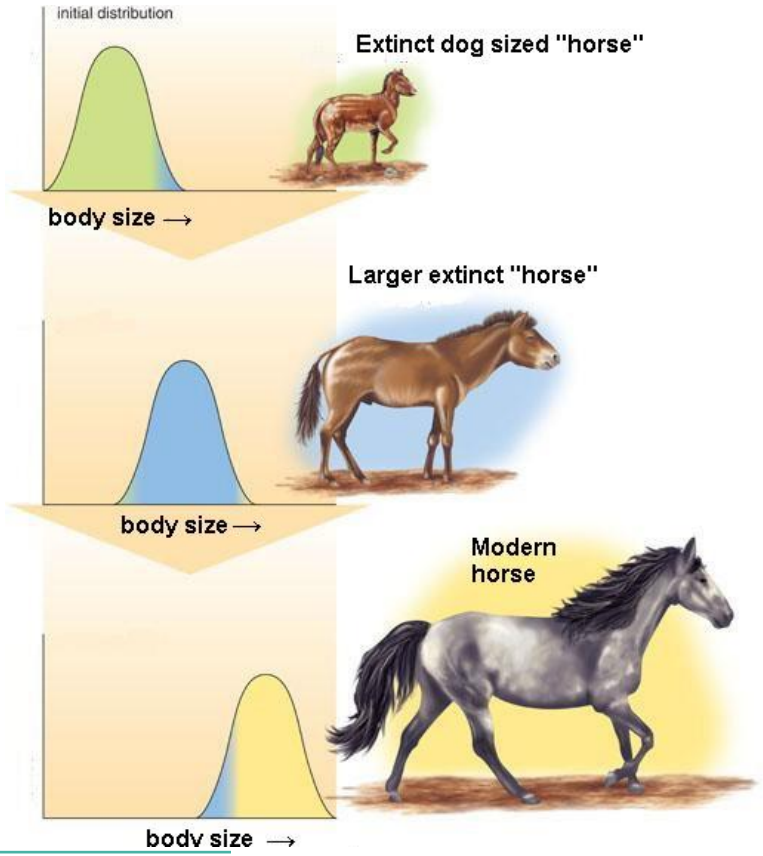
what do we partition?

- substitution model parameters
- among-site rate variation
- base rate

- branch lengths?
- topology?
- relaxed-clock model?

CONTINUOUS CHARACTERS

- body size, brain size, limb length, weight, morphometrics, etc.



CONTINUOUS CHARACTERS

- body size, brain size, limb length, weight, morphometrics, etc.

- study character evolution
- study correlations (comparative method)

- Markov model won't work, unless we discretize the continuous states

→ Brownian motion

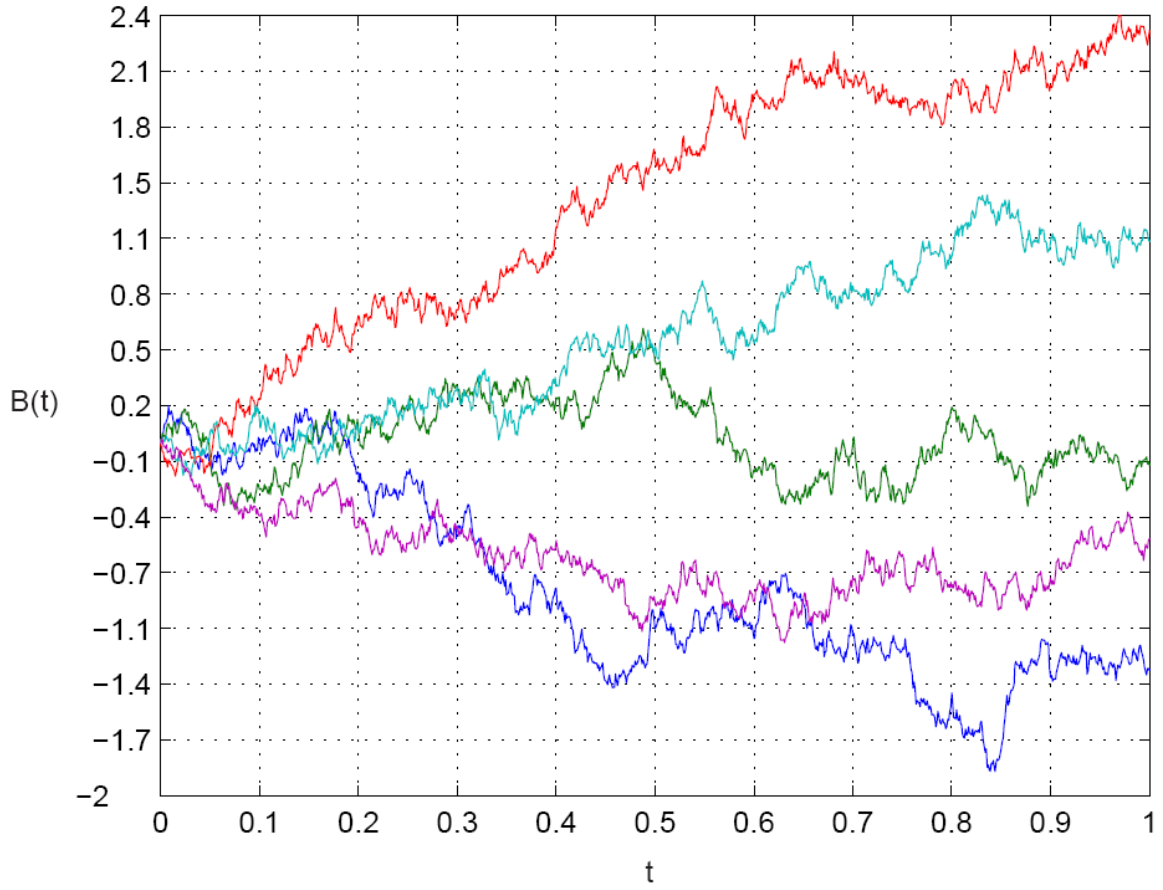
CONTINUOUS CHARACTERS

→ **Brownian motion**

- changes drawn from a normal distribution with mean = previous value and a specified variance
- again memoryless

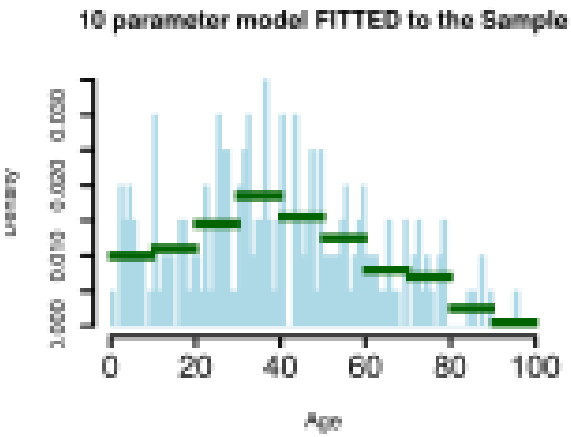
CONTINUOUS CHARACTERS

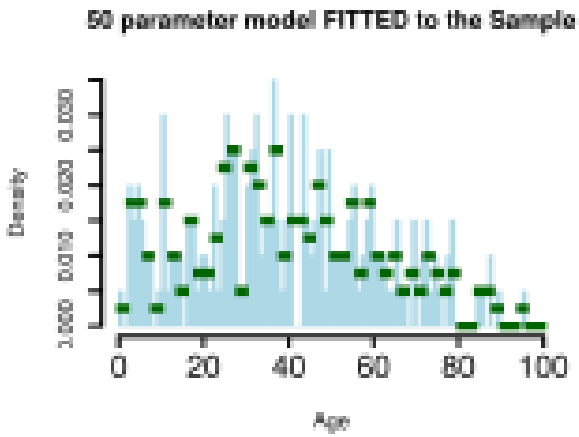
→ Brownian motion



(OVER)PARAMETRIZATION

- relationship between amount of data and number of infered parameters





(OVER)PARAMETRIZATION

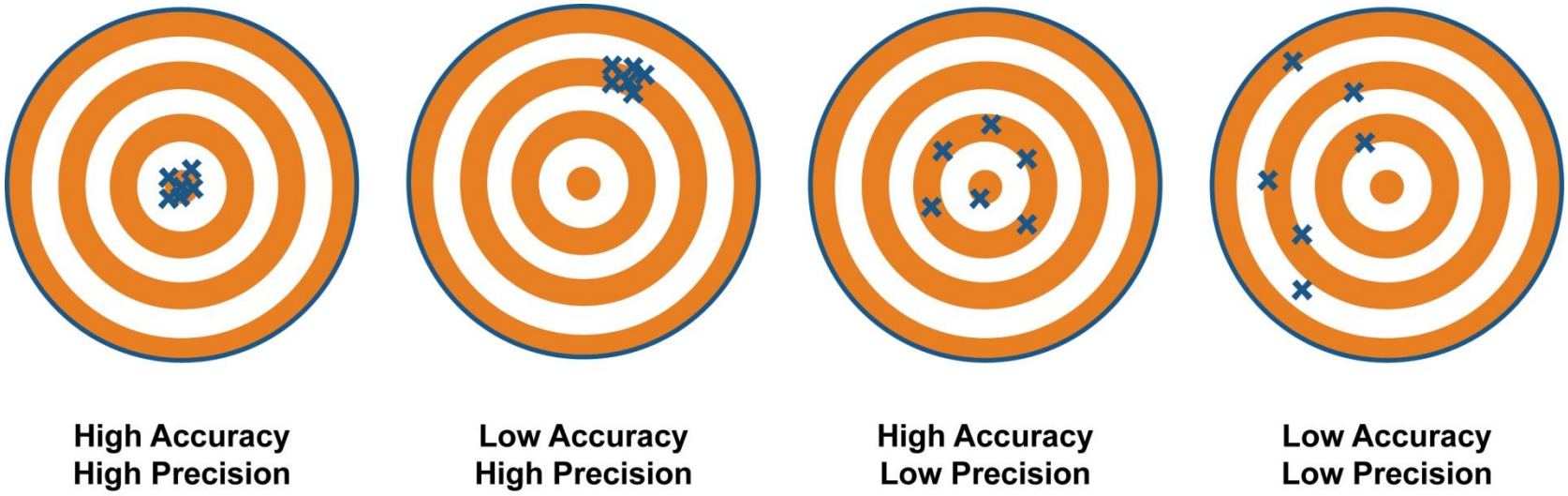
- relationship between amount of data and number of infered parameters
  - example: DNA models
    - how many sites vary per partition?
  - example: morphology models
    - often very few characters -> use simple model!

- model misspecification: failure to incorporate vital characteristic of the process

(OVER)PARAMETRIZATION

distinction between accuracy and precision

- how do those relate to model misspecification and over-parametrization?



SUBSTITUTION MODELS

- which of these model assumptions are “worst” for YOUR current dataset?
  - standard partitioning scheme
  - gamma among-site rate variation
  - homogeneity
- what dataset had the worst ratio in terms of data / model parameters?
- what non-standard data types have you worked with?