Statistical Phylogenetics: An Introduction

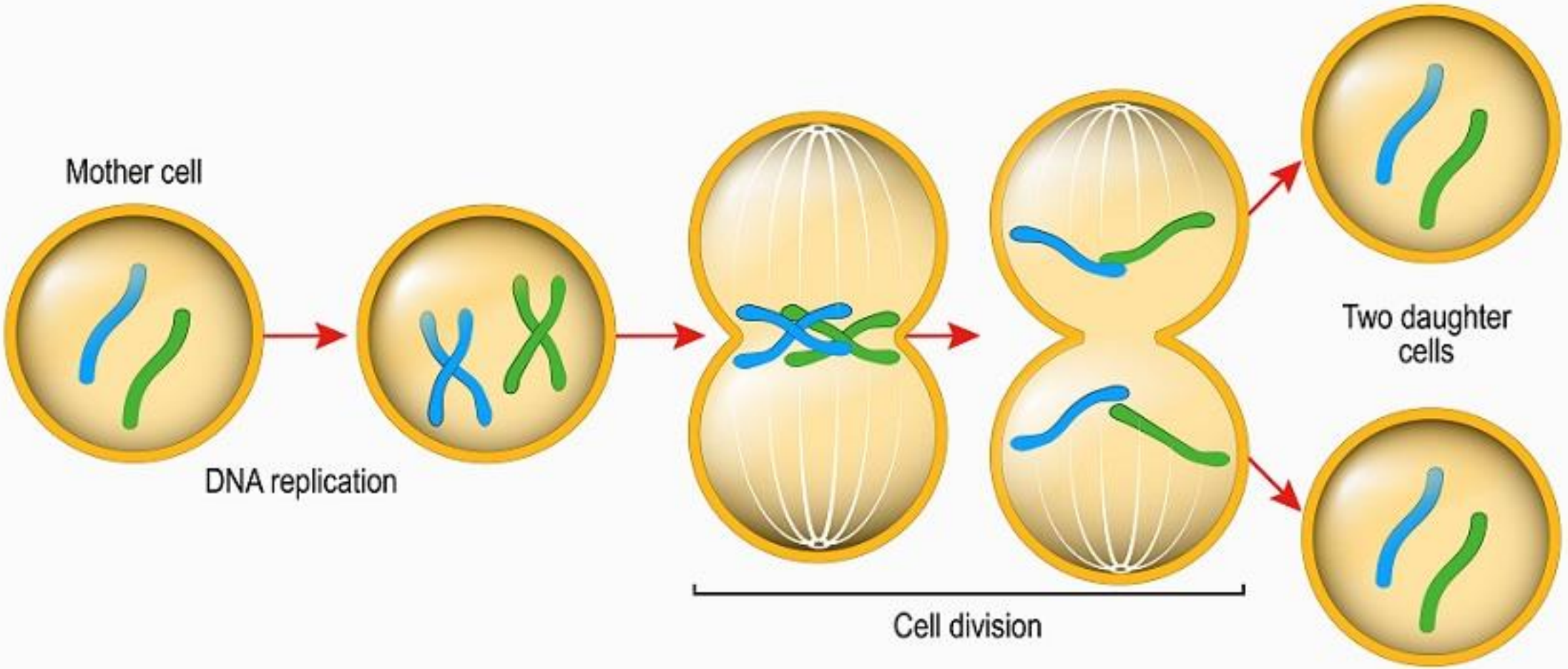
2024-12-04, Miking Workshop

Prof. Fredrik Ronquist

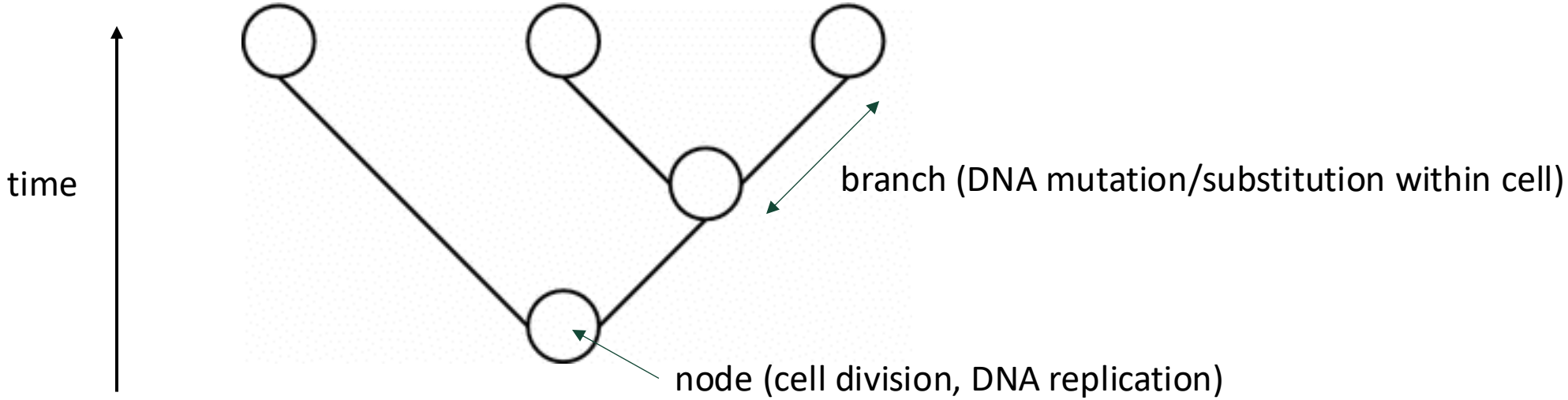
Dept. Bioinformatics and Genetics

NATURHISTORISKA RIKSMUSEET

Cell Division (Mitosis)



Phylogenetic Model: The Basics

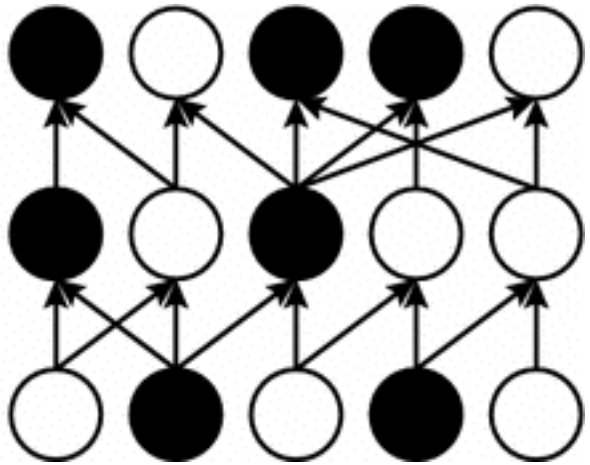


Phylogenetic tree (phylogeny)

Stochastic model:
- Model of branching process
- Model of substitution process

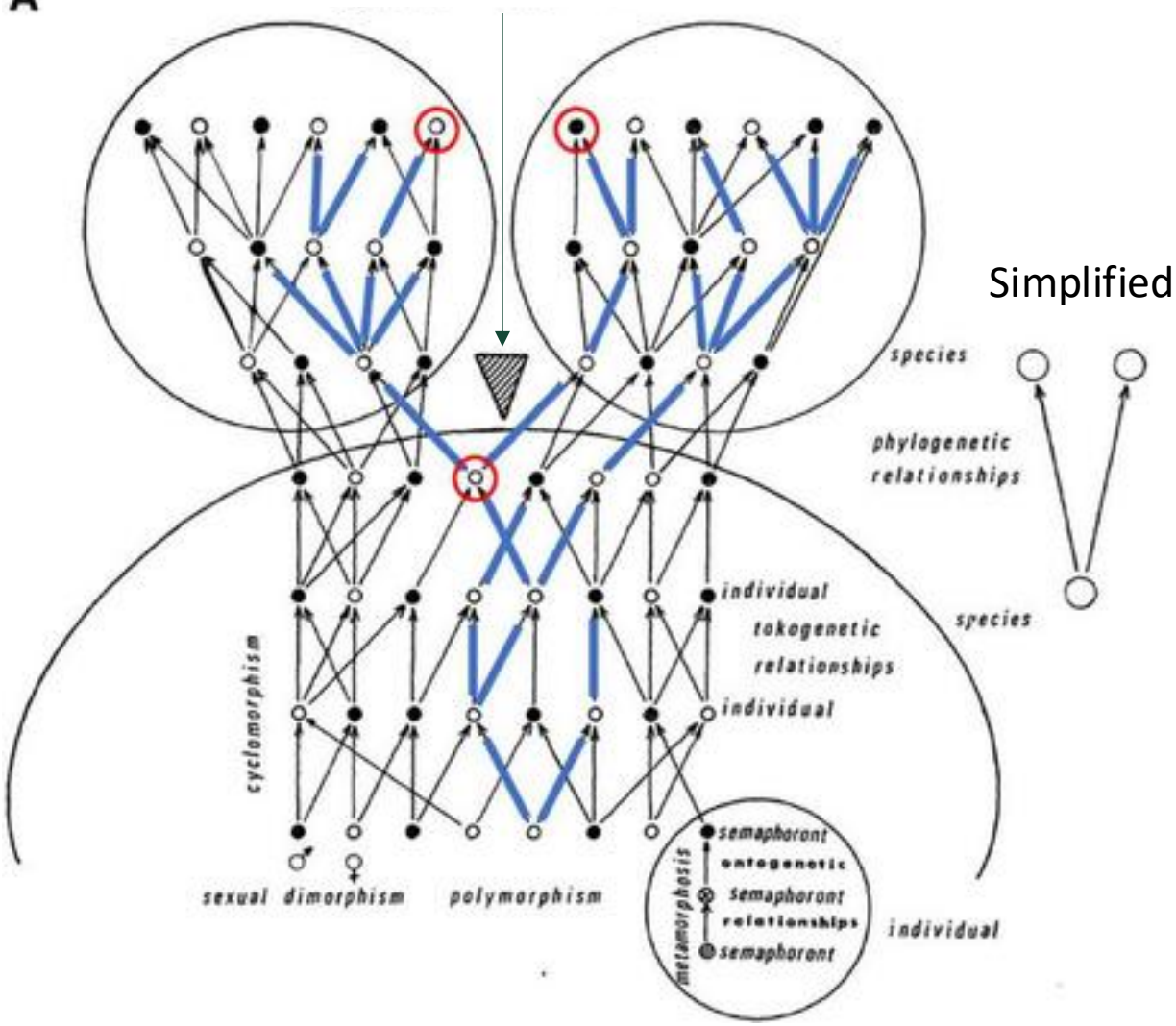
population/lineage/species

individual



Sexually reproducing species

Spatial or temporal barrier



Simplified phylogeny

Applications of Statistical Phylogenetics

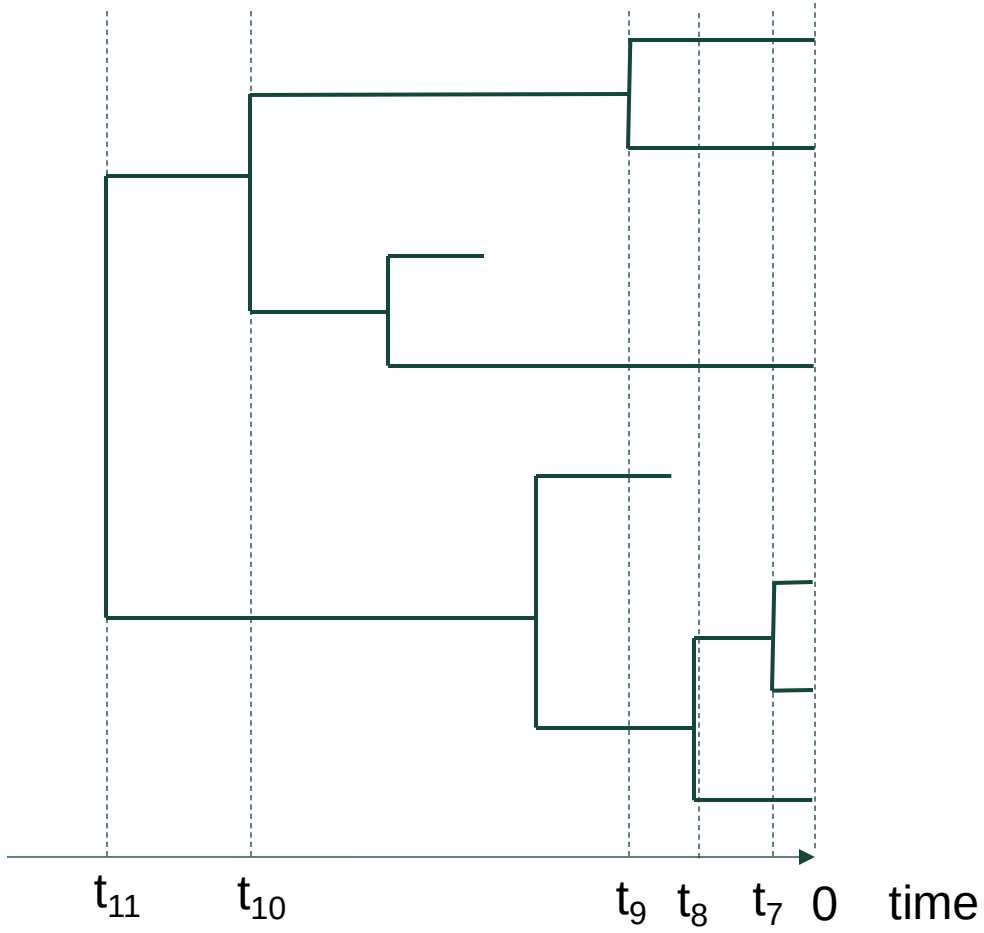
- Virus transmission pathways (HIV, SARS-CoV-2,…)
- Epidemiology
- Predicting next year’s influenza outbreak
- Identification of pathogens
- Relationships among organisms
- Divergence time estimation
- Molecular evolution processes
- Selection analysis (which genes cause X)
- Patterns of diversification and extinction
- Biogeography (where does an organism come from)
- …

NATURHISTORISKA RIKSMUSEET

”Nothing in Biology Makes Sense Except in the Light of Evolution”

Theodosius Dobzhansky (1900-1975)

Tree model



Birth-death model:

Birth rate λ

Death rate μ

Root time $t_{\mathrm{mrca}} = t_{11}$

The birth-death model induces a probability distribution on

Topology τ

Speciation times t

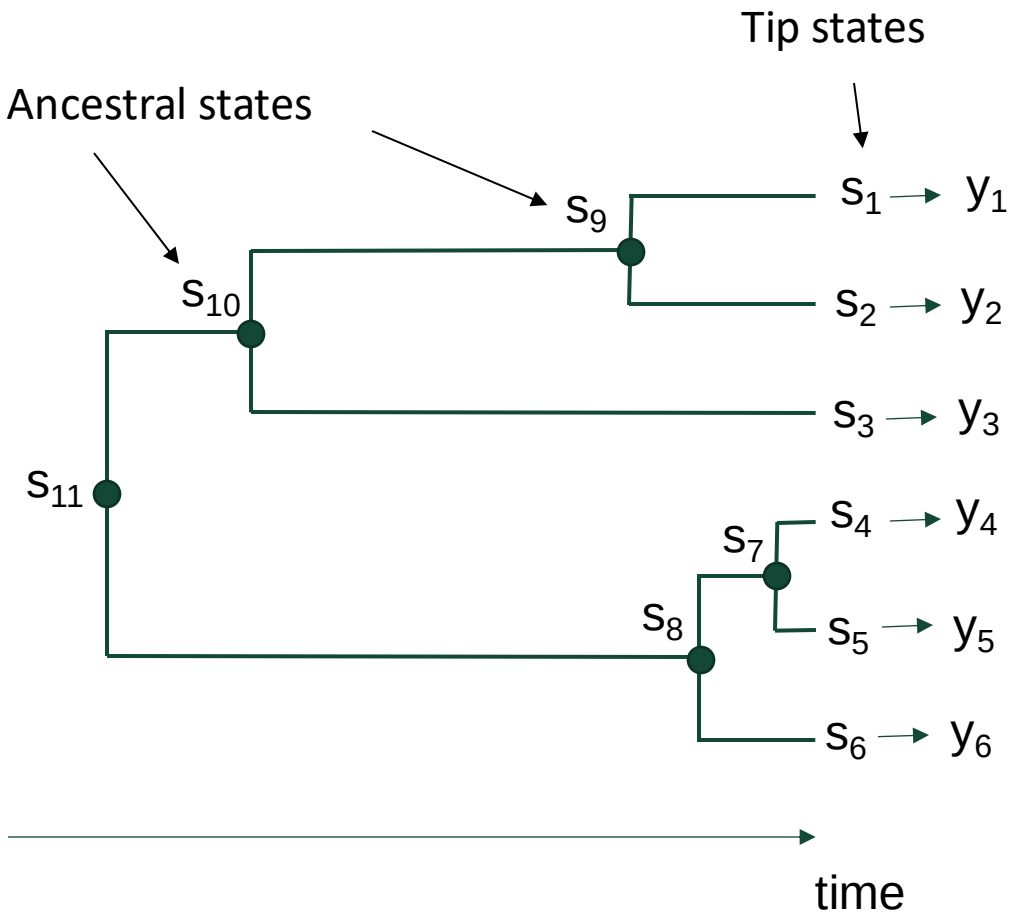
Given a substitution rate r, branch lengths b are given by $b_i = r(t_{a_i} - t_i)$

Substitution model

DNA sequence

GATAAATAATATAAGATTTTGAC...

site (*s*) assumed iid



Observation error usually ignored, that is, it is assumed that $y_i = s_i$ for all leaves i.

DNA sequences are drawn iid from a discrete-state continuous-time Markov chain over four nucleotides: A, C, G, T

$$Q = \begin{pmatrix} - & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & - & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & - & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & - \end{pmatrix}$$

Instantaneous rate matrix for the General Time Reversible (GTR) substitution model

π Stationary state frequencies

r Exchangeability rates

Standard approach in statistical phylogenetics

- Specify model: script or command-line settings
- Hard-coded Bayesian MCMC (or maximum likelihood) inference machinery

NATURHISTORISKA RIKSMUSEET

Home
» Download «
Manual
Bug Report
Authors
Links

MrBayes: Bayesian Inference of Phylogeny

Download MrBayes

MrBayes may be downloaded as a pre-compiled executable or in source form (recommended).

Current release

The most recent release version of MrBayes is 3.2.7a, released March 6, 2019.

The 3.2.7a source code is available for compilation on Unix machines.

Pre-compiled (provisional) executables are available for Windows (MrBayes-3.2.7-WIN.zip). These are, however, serial versions compiled without the Beagle library. The serial version works well for smaller analyses but if you plan to run large analyses using many parallel chains, you should use the MPI version instead. Refer to the User Manual, and the INSTALL document on GitHub for help with installation of the program.

MrBayes may also be installed through the Homebrew package manager on macOS, Linux, and Windows Subsystem for Linux (WSL). Please see the INSTALL document for instructions.

Older releases

You can get access to older releases (from release 3.2.0 onwards), by browsing the releases directory on github.

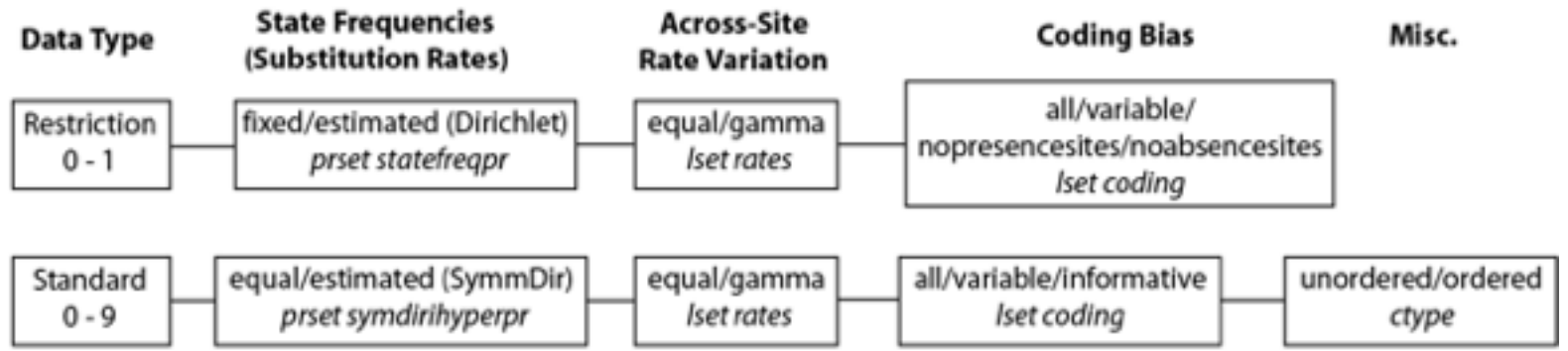
Developer version

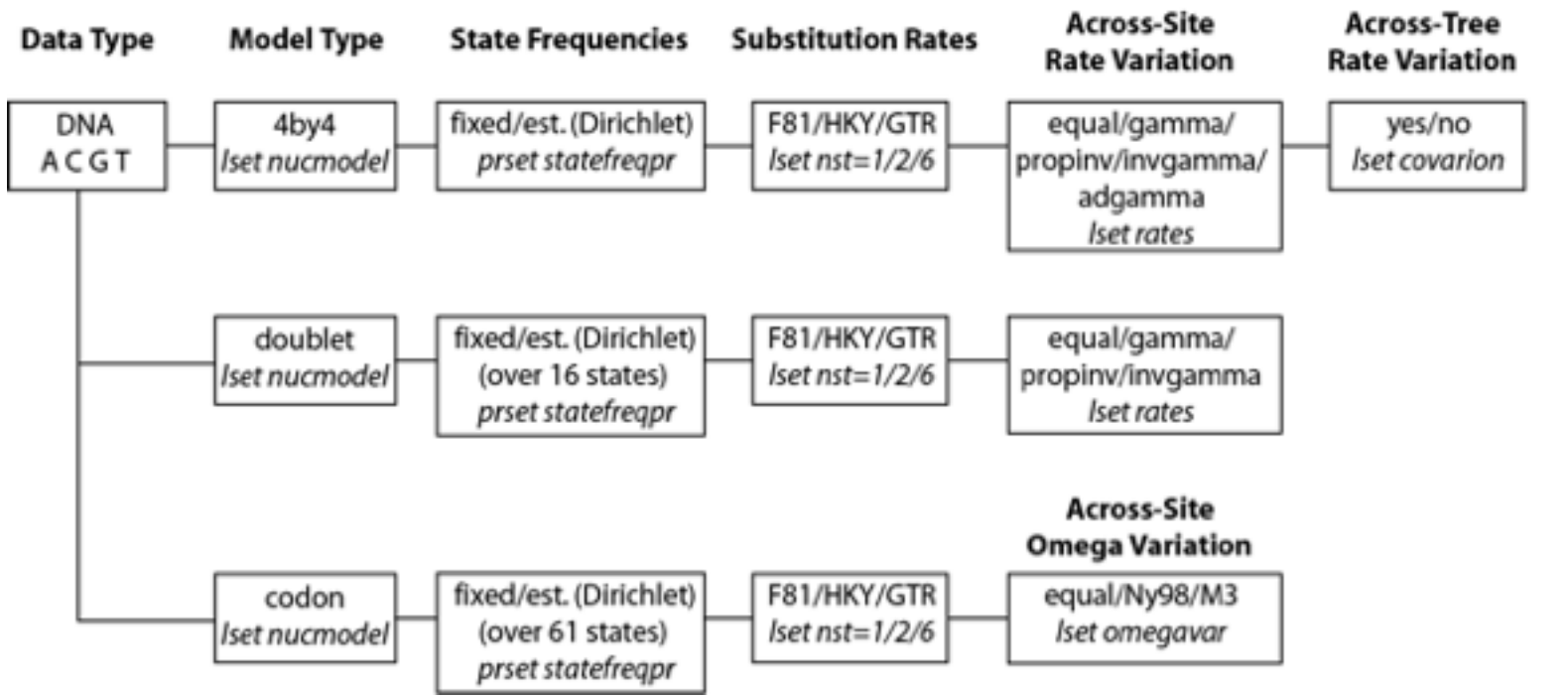
You may also download cutting edge developer version of MrBayes from the Git repository hosted at GitHub. Note that you have to compile the code yourself. Read instructions in the INSTALL file in the source code repository for further instructions.

https://nbisweden.github.io/MrBayes/download.html

Huelsenbeck & Ronquist (Bioinformatics, 2001)

Models supported by MrBayes 3 (simplified)

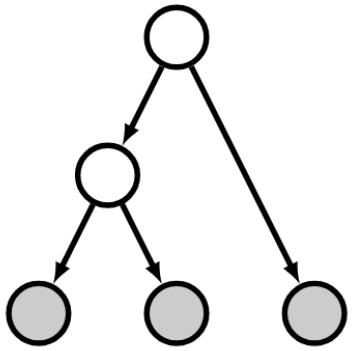




MrBayes script

```
#NEXUS

begin mrbayes;

    execute data.nex;

    outgroup Ibalia;
    charset morphology = 1-166;
    charset molecules = 167-3246;
    charset COI = 167-1244;
    charset EF1a = 1245-1611;
    charset LWRh = 1612-2092;
    charset 28S = 2093-3246;
    partition favored= 5: morphology, COI, EF1a, LWRh, 28S;

    set partition=favored;
    lset applyto=(1) rates=gamma;
    lset applyto=(2,3,4,5) rates=invgamma nst=mixed;
    unlink revmat=(all) pinvar=(all) shape=(all) statefreq=(all);
    prset ratepr=variable;

end;
```

Download Tutorials Documentation Interfaces Workshops Jobs Developer



RevBayes

Bayesian phylogenetic inference using probabilistic graphical models and an interpreted language

About

RevBayes provides an interactive environment for statistical computation in phylogenetics. It is primarily intended for modeling, simulation, and Bayesian inference in evolutionary biology, particularly phylogenetics. However, the environment is quite general and can be useful for many complex modeling tasks.

RevBayes uses its own language, Rev, which is a probabilistic programming language like JAGS, STAN, Edward, PyMC3, and related software. However, phylogenetic models require inference machinery and distributions that are unavailable in these other tools.

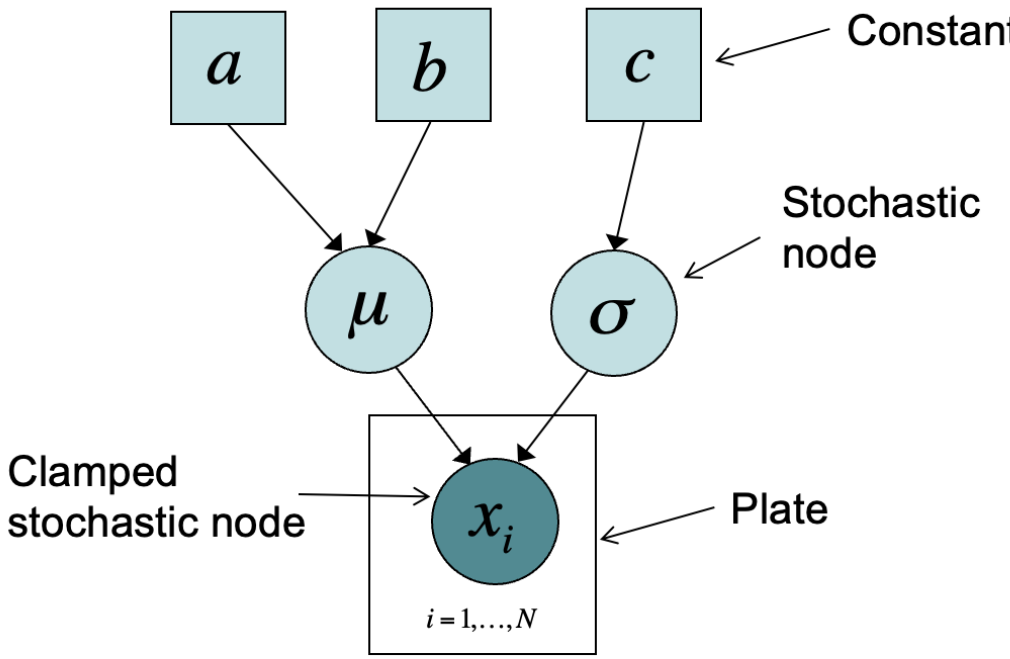
The Rev language is similar to the language used in R. Like the R language, Rev is designed to support interactive analysis. It supports both functional and procedural programming models, and makes a clear distinction between the two. Rev is also more strongly typed than R.

RevBayes is a collaboratively developed software project.

GitHub | License | Citation | Users Forum

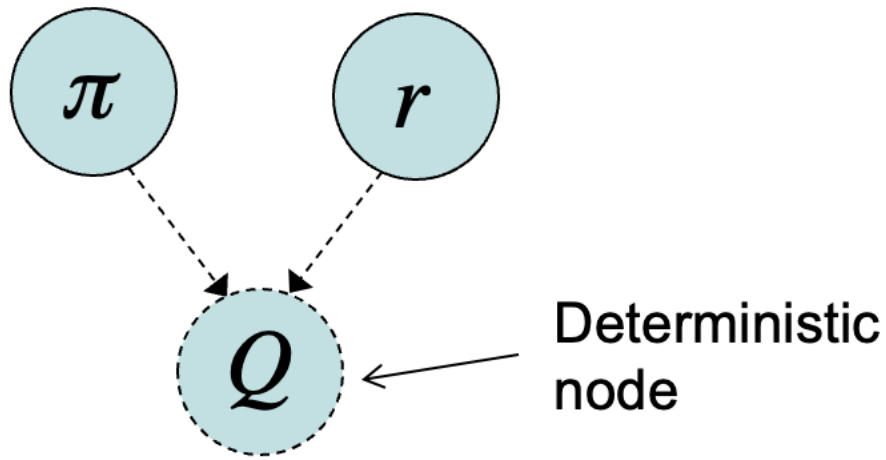
http://www.revbayes.com

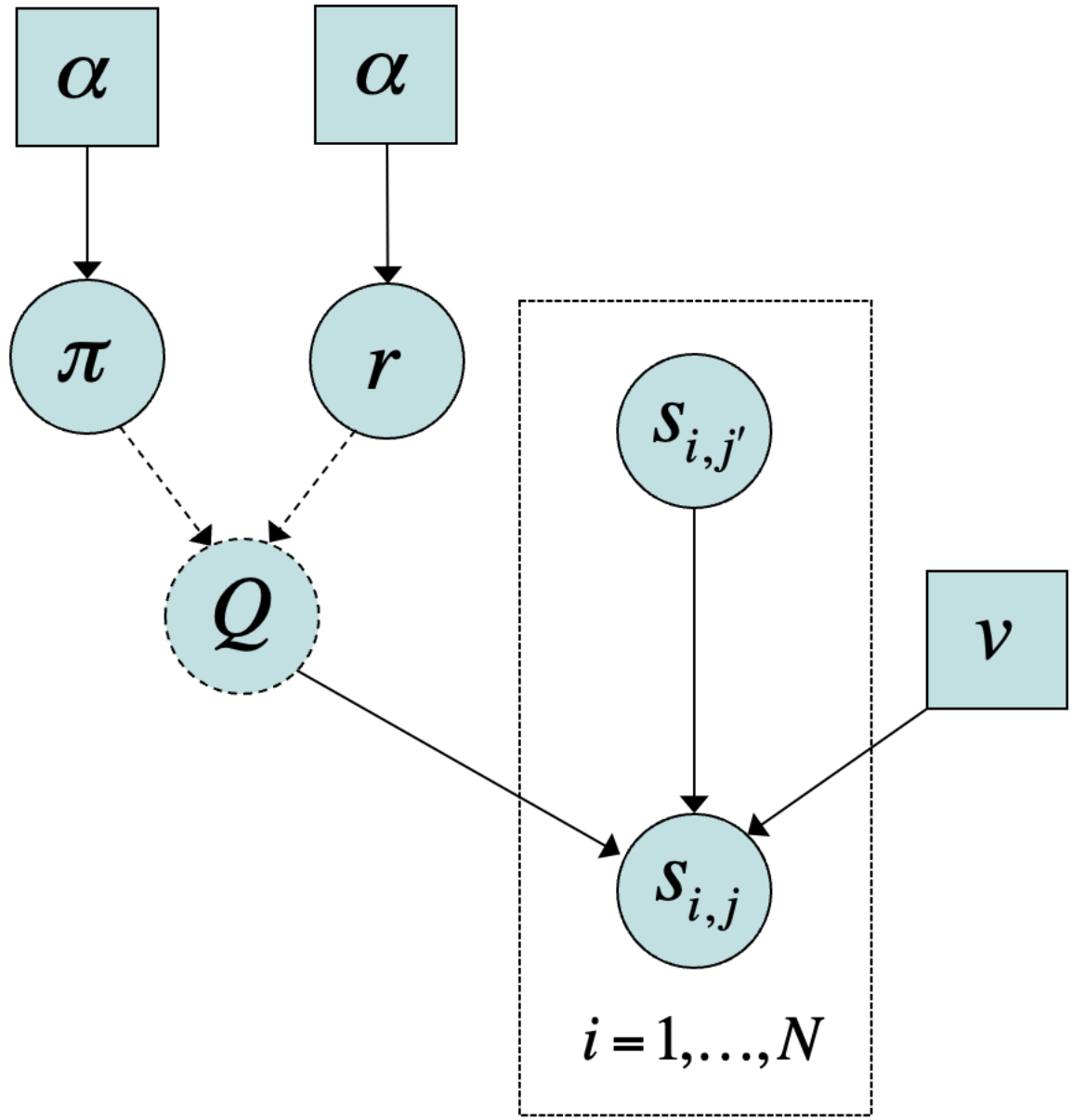
Höhna et al. (Systematic Biology, 2014)

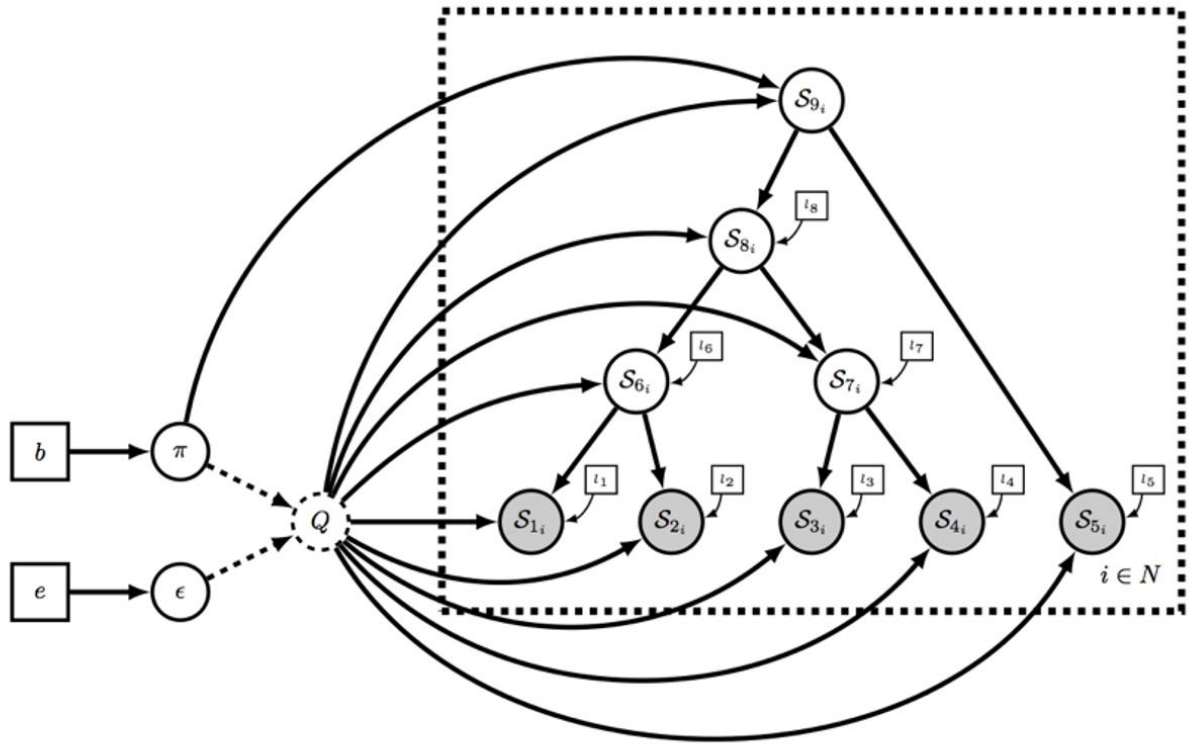


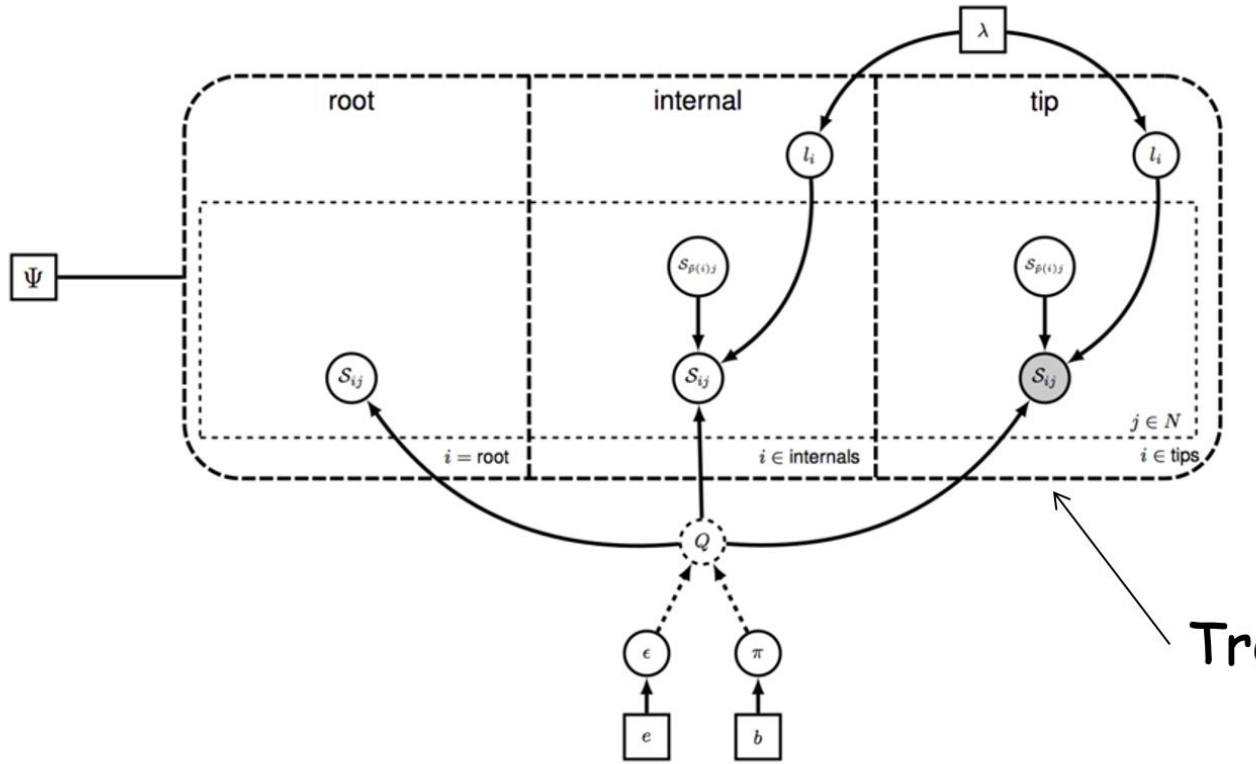
$$\mu \sim \text{Unif}(a, b)$$
$$\sigma \sim \text{Exp}(c)$$
$$x \sim \text{iid Norm}(\mu, \sigma)$$

Hierarchical Normal Model

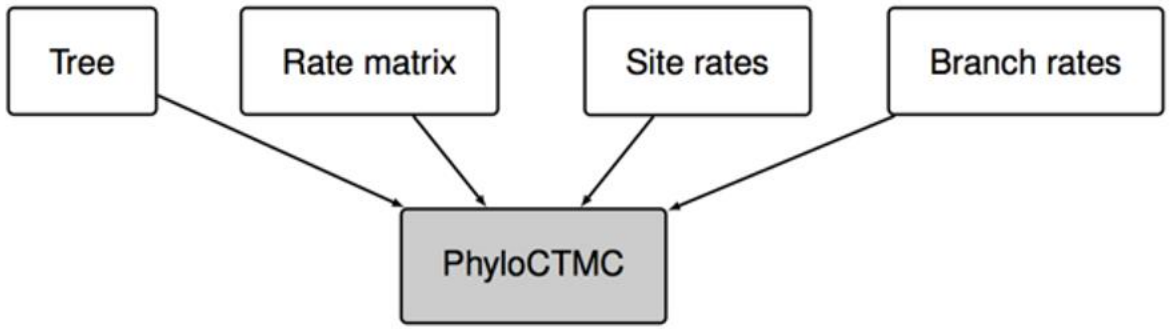

π
r
Q
Deterministic
node


α
α
π
r
$s_{i,j'}$
Q
v
$s_{i,j}$
$i = 1,\ldots,N$

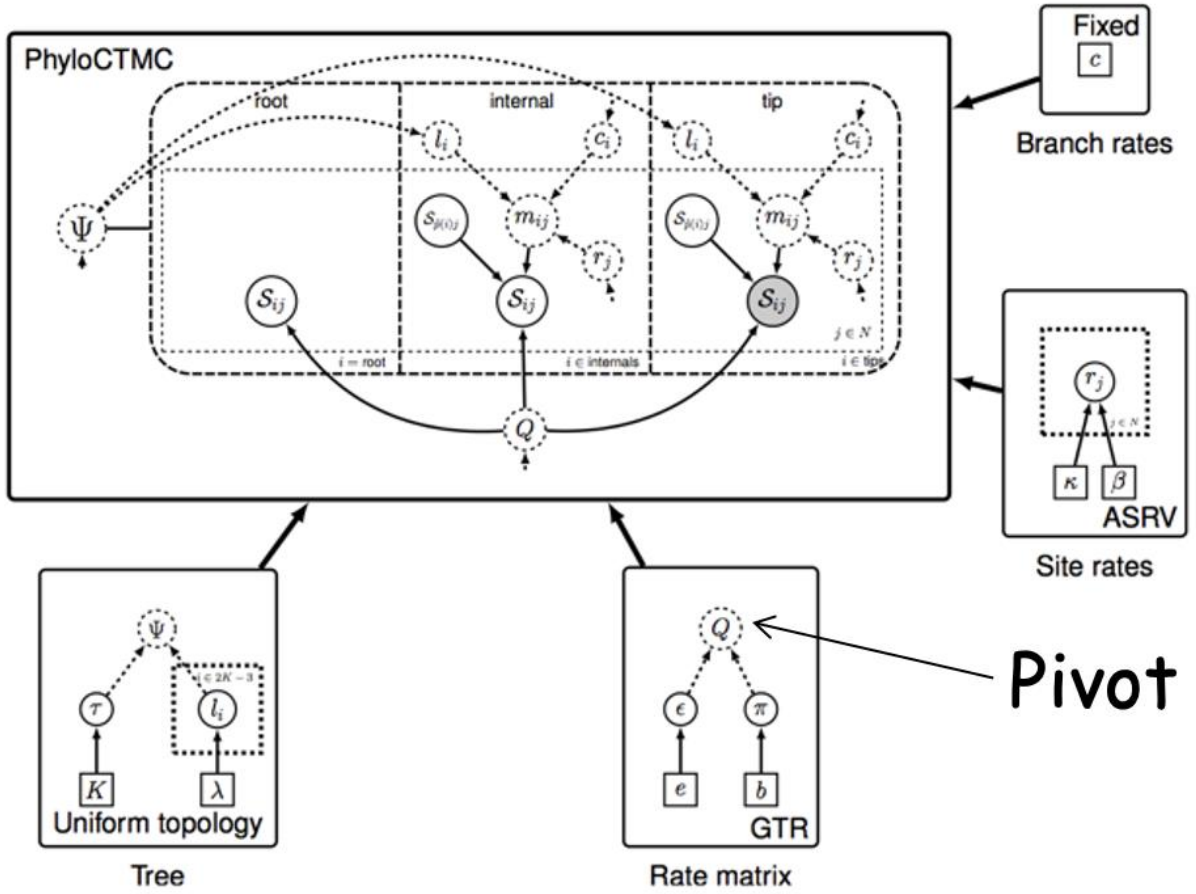

b
π
e
ϵ
Q
$\mathcal{S}_{9_i}$
$\mathcal{S}_{8_i}$
$\mathcal{S}_{6_i}$
$\mathcal{S}_{7_i}$
$\mathcal{S}_{1_i}$
$\mathcal{S}_{2_i}$
$\mathcal{S}_{3_i}$
$\mathcal{S}_{4_i}$
$\mathcal{S}_{5_i}$
l_1
l_2
l_3
l_4
l_5
l_6
l_7
l_8
$i \in N$


root
internal
tip
λ
l_i
l_i
Ψ
$s_{\hat{p}(i)j}$
$s_{\hat{p}(i)j}$
s_{ij}
s_{ij}
s_{ij}
$j \in N$
$i = \text{root}$
$i \in \text{internals}$
$i \in \text{tips}$
Q
ϵ
π
e
b

Tree plate



Modular representation



Pivot variable

Challenges for a universal PPL for phylogenetics

- Ease of use in describing models and inference algorithms
- Computational efficiency [problem sizes are large, convergence is difficult]
  - A range of inference strategies
  - Fast computation of likelihoods/weights
  - Informative diagnostics (are we there yet?)
- Belief propagation
  - Makes a huge difference when it can be applied
  - Standard model (GTR+Gamma) requires belief propagation in two dimensions (rates across sites and ancestral nucleotide states)
  - Automated?
- Alignment
  - Required when number of random variables varies (birth-death models, sampling of change histories)
  - Automated?
- Proposal/guide distributions
  - Using quick-and-dirty approximations essential for efficient inference
  - Generic `propose` – `update` mechanism?

NATURHISTORISKA RIKSMUSEET