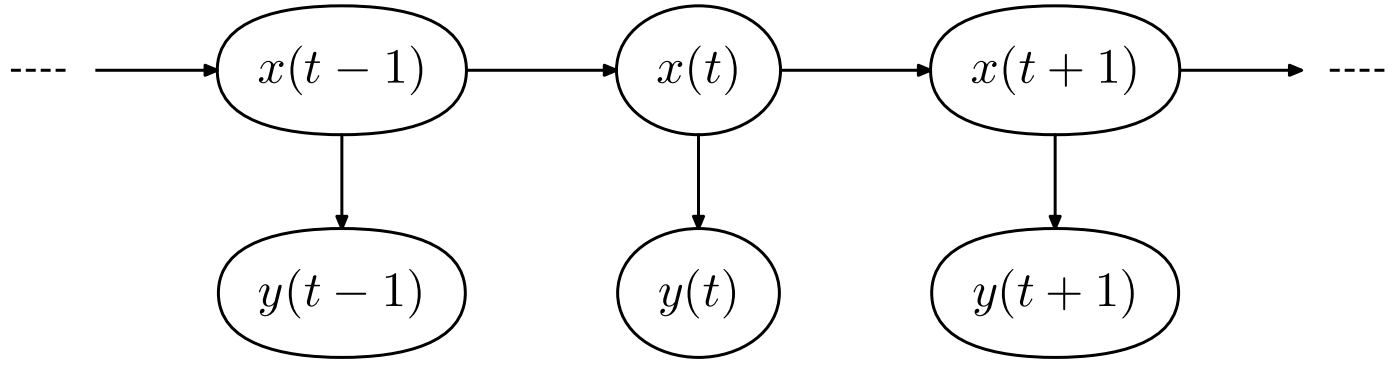


Specifying Graphical Models in `RevBayes`

Will Freyman

Department of Integrative Biology
University of California, Berkeley

freyman@berkeley.edu
http://willfreyman.org

UC Berkeley, February 26-27 2018

Workshop goals:

- not simply demonstrate "standard" phylogenetic analyses in `RevBayes`
- instead we'll explore the flexibility of a graphical modeling framework
- use graphical models to see how the models underlying tree inference and downstream tree use (comparative methods) are linked
- enable participants to specify custom and unique phylogenetic analyses in `RevBayes`

What are we doing in phylogenetics?

- "inference"? (i.e. statistical inference?)
- "learning"? (i.e. machine learning?)
- "prediction"?

In a probabilistic framework (maximum likelihood or Bayesian) inference and learning are the same

When we "train" a machine learning algorithm we are doing parameter estimation

The field of machine learning includes camps that use principled probabilistic approaches and camps that use heuristic ad hoc methods (like in phylogenetics!)

In phylogenetics we should be doing more prediction!
i.e. model adequacy/posterior predictive tests

What is a model?

- in statistics?
- in machine learning?
- in biology?

maybe:

- a way to relate data to hypotheses?
  - what about heuristic or ad hoc approaches?
  - is parsimony in phylogenetics a model, an algorithm, or a philosophy?
- a set of assumptions about the data-generating process?
- "a formal representation of a theory"?
- a set of mathematical equations that relate one or more random variables?

The distinction between models and algorithms:

model	algorithm
phylogenetic models	pruning algorithm
Bayesian graphical model	belief propagation
neural networks	backpropagation
Hidden Markov model	forward-backward
k-means clustering	Lloyd's algorithm
linear regression	least-squares

Learning algorithms typically either optimize $\hat{\theta}$ or integrate to infer $p(\theta|D)$

They are often *very similar* and can be used with other models

Any well defined model can be treated in a *probabilistic* framework and then we can use Bayesian or maximum likelihood approaches

Probabilistic models:

Instead of a hodgepodge of different heuristic methods these models use the principles of probability theory

Why use them?

- Quantify uncertainty: they know when they don't know
  - what is the best prediction/decision/inference given data?
  - what is the best model/hypothesis given the data?
  - do I need more/different data?
- natural complexity control
  - preventing overfitting / regularization
- modularity
  - models as "lego kits"
  - different inferential algorithms can use the same model
  - different models can use the same inferential algorithm

Discriminative vs Generative Models

Discriminative (or conditional) models:

1. models a response variable conditioned on a predictor variable
2. models the conditional distribution $p(y|x)$
3. makes fewer assumptions about the data: $p(x)$ not necessary

Phylogenetic examples:

- estimating divergence times over a fixed topology
- estimating ancestral states on a fixed tree
- estimating shifts in diversification rates over a fixed tree

Discriminative vs Generative Models

Generative models:

1. models the entire process used to generate the data
2. models the joint distribution $p(x, y)$
3. makes more assumptions about the data: need to define $p(x)$
4. richer representation of the relations between variables
5. more powerful: allows us to compute $p(y|x)$ or $p(x|y)$
6. more powerful: can simulate both x and y

Phylogenetic examples:

- jointly estimating divergence times and the tree topology
- jointly estimating ancestral states and the tree
- jointly estimating shifting diversification rates and the tree

What is a graphical model?

Also called:

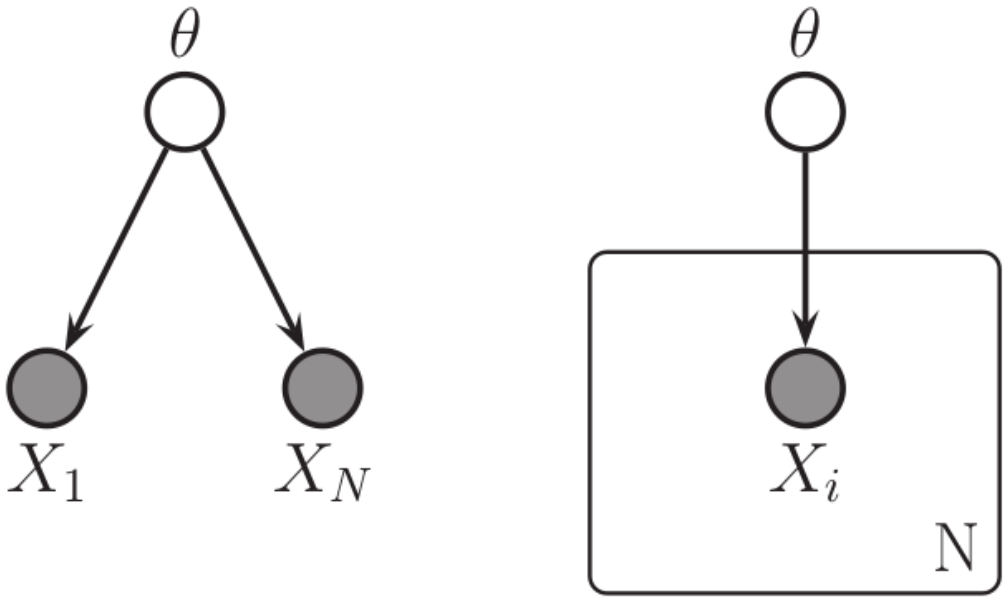
1. Bayesian networks
2. belief networks
3. causal networks

A useful way to represent a probabilistic model: a joint distribution of random variables.

We can specify both generative and discriminative models as graphical models.

What is a graphical model?

Nodes represent variables and edges represent conditional dependencies:



$$p(\theta, \mathcal{D}) = p(\theta)\Big[\sum_{i=1}^{N} p(x_i|\theta)\Big]$$

Image from Murphy (2012)

What is a graphical model?

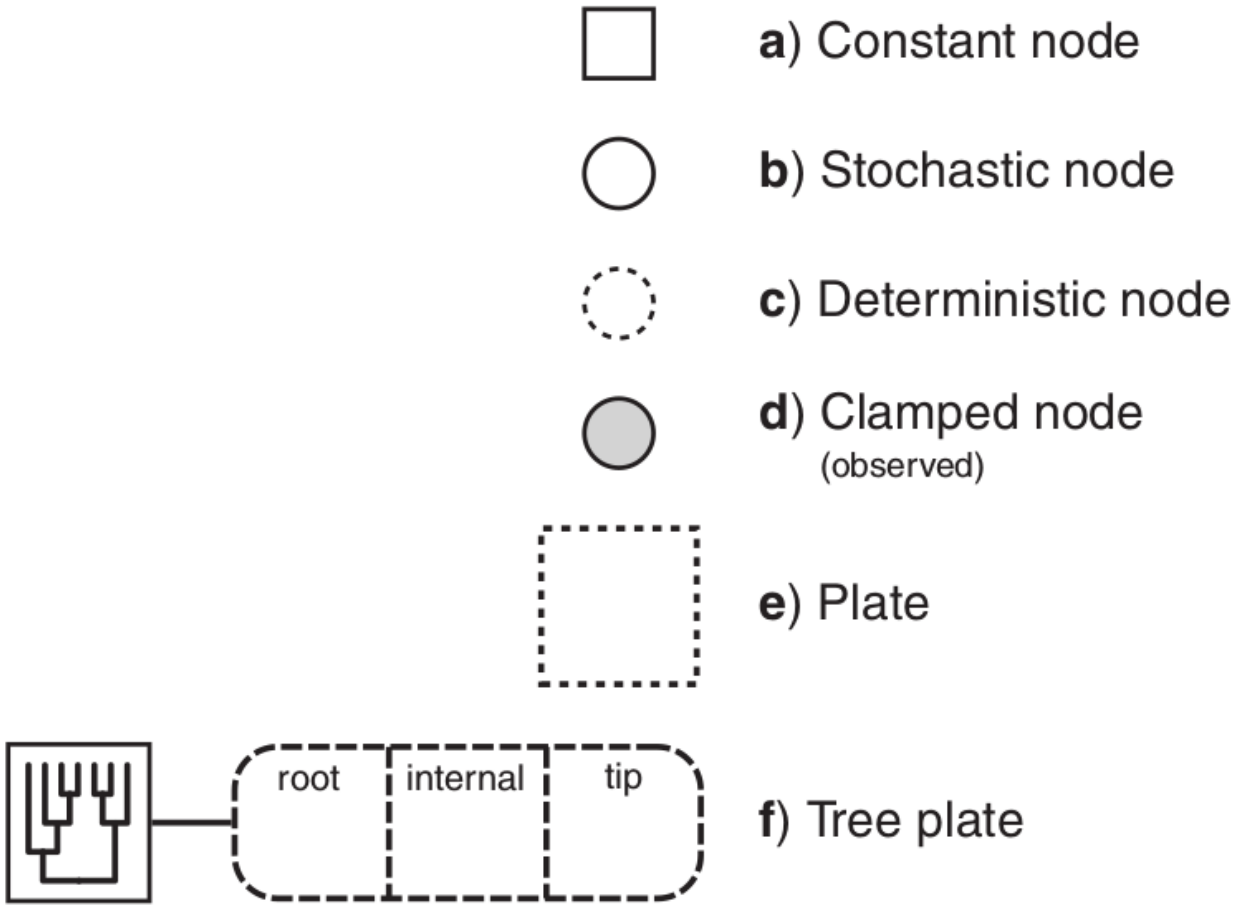


Image from Höehna et al. (2014)

phylogenetic graphical model

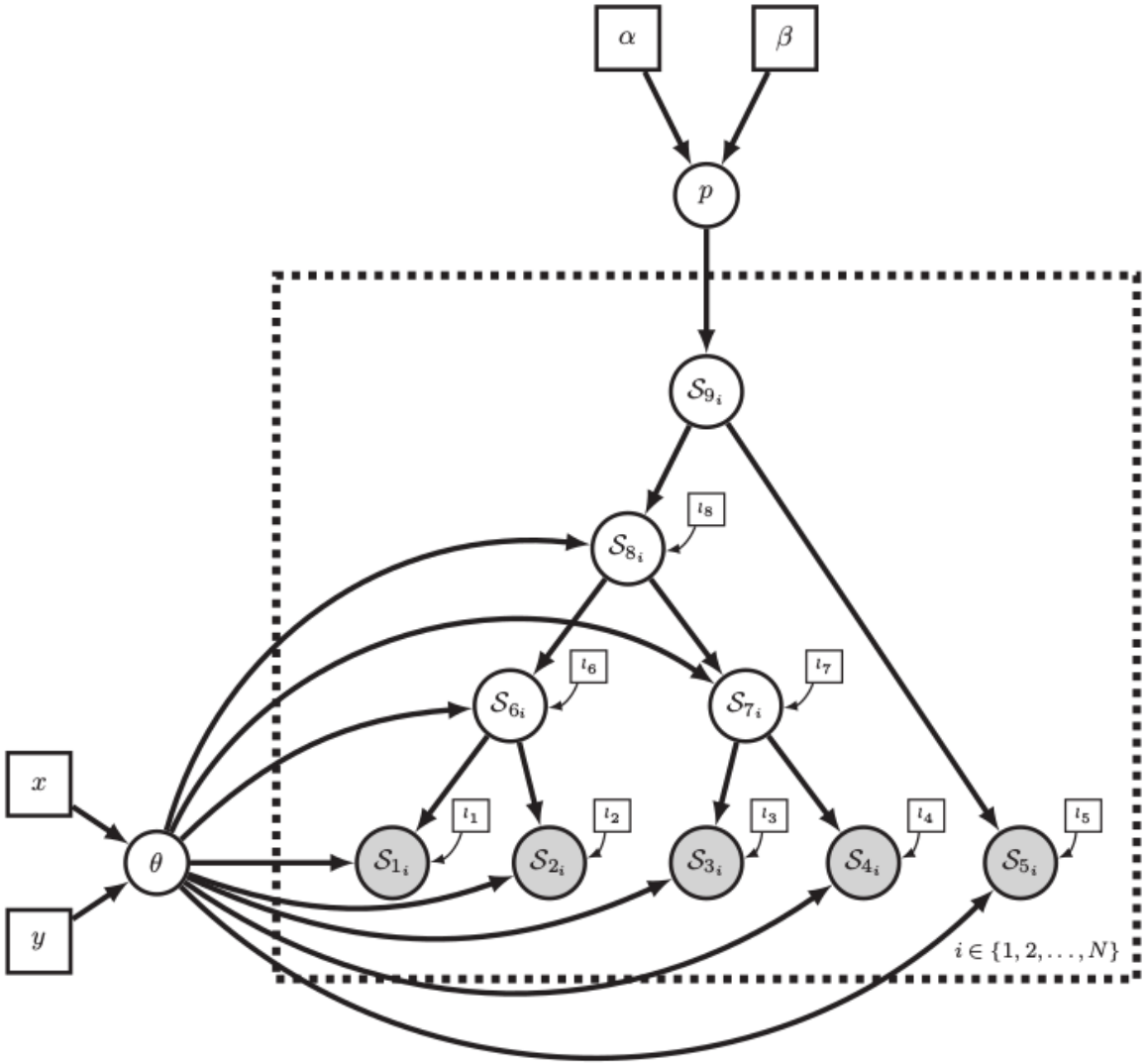


FIGURE 4. A phylogenetic graphical model of N independently evolving binary characters. When sampling N different binary characters for each extant species, we assume that these characters are independent and identically distributed. Thus the model for each character is the same as in Figure 3b. Yet, the state for each character $1,\ldots,N$ can be different. We use the *plate* notation to represent repetition over a vector of elements. In this figure, the dashed box and the iterator i indicate the replicated variables. Thus, the plate represents separate variables of binary character evolution for i in characters $1,2,3,\ldots,N$.

Image from Höehna et al. (2014)

phylogenetic graphical models as modules

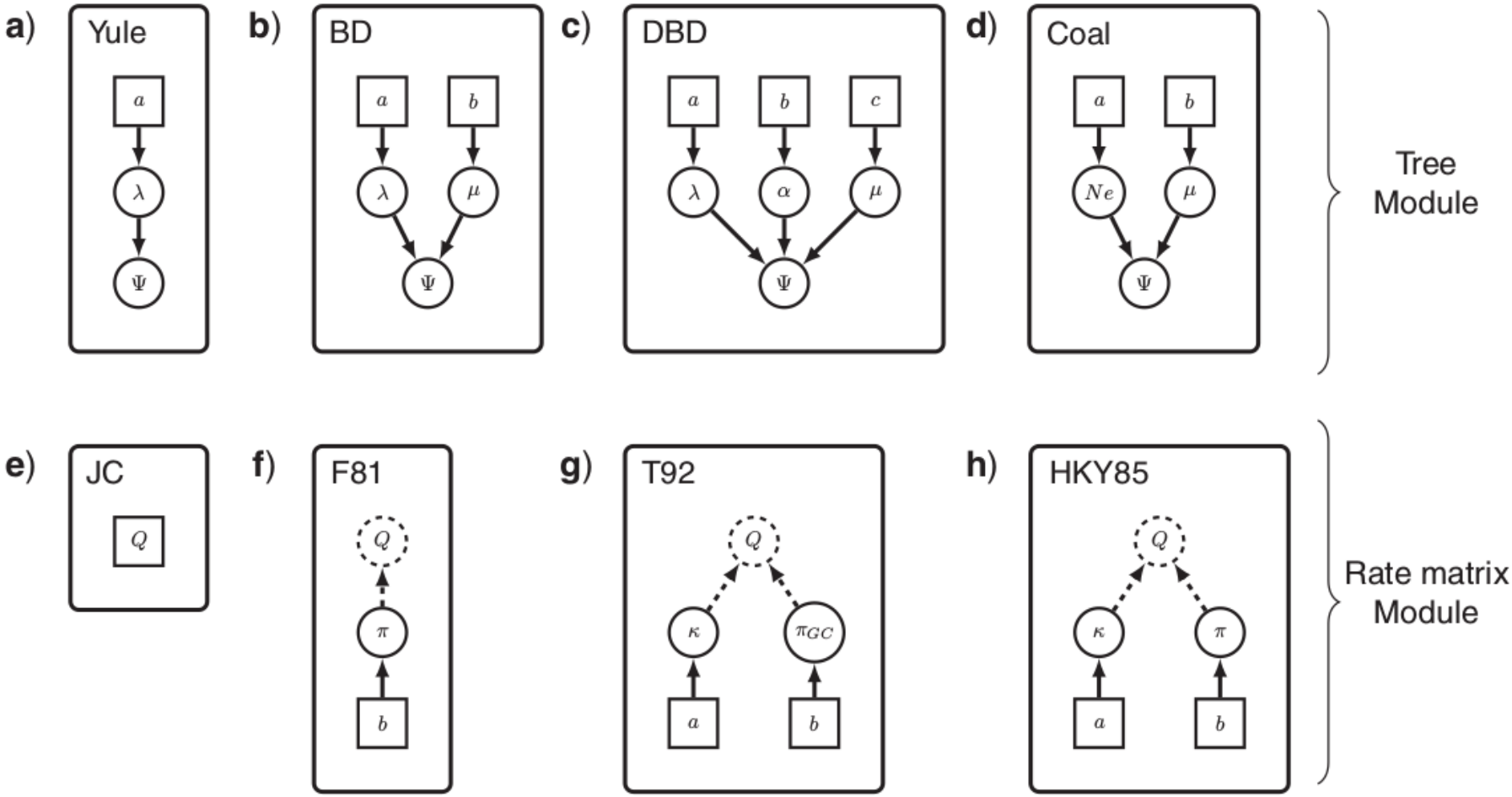


Image from Höehna et al. (2014)

assembling phylogenetic models like lego kits

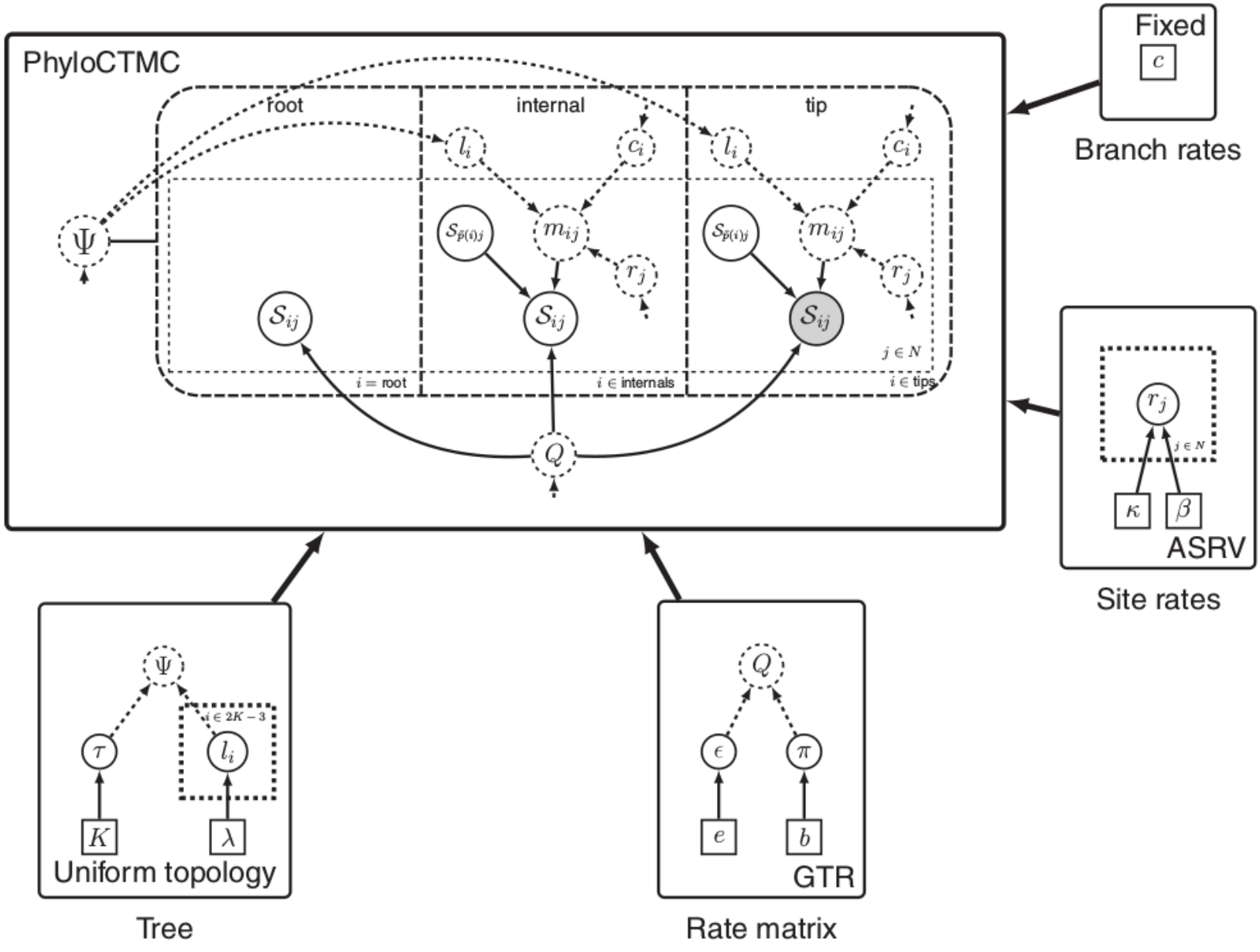


Image from Höehna et al. (2014)

Is the graphical model paradigm really helpful?

Disadvantages:

1. steep learning curve...
   - constant, stochastic, deterministic nodes
   - clamping
   - MCMC proposals

Advantages:

1. transparency: all modeling assumptions are specified
2. power and flexibility: build custom models that test your specific hypotheses
3. efficiency: customize inference algorithm to efficiently perform inference
4. applicability: the same concepts are widely used in many probabilistic programming languages like Stan, BUGS, Edward, PyMC3... and `Rev`!

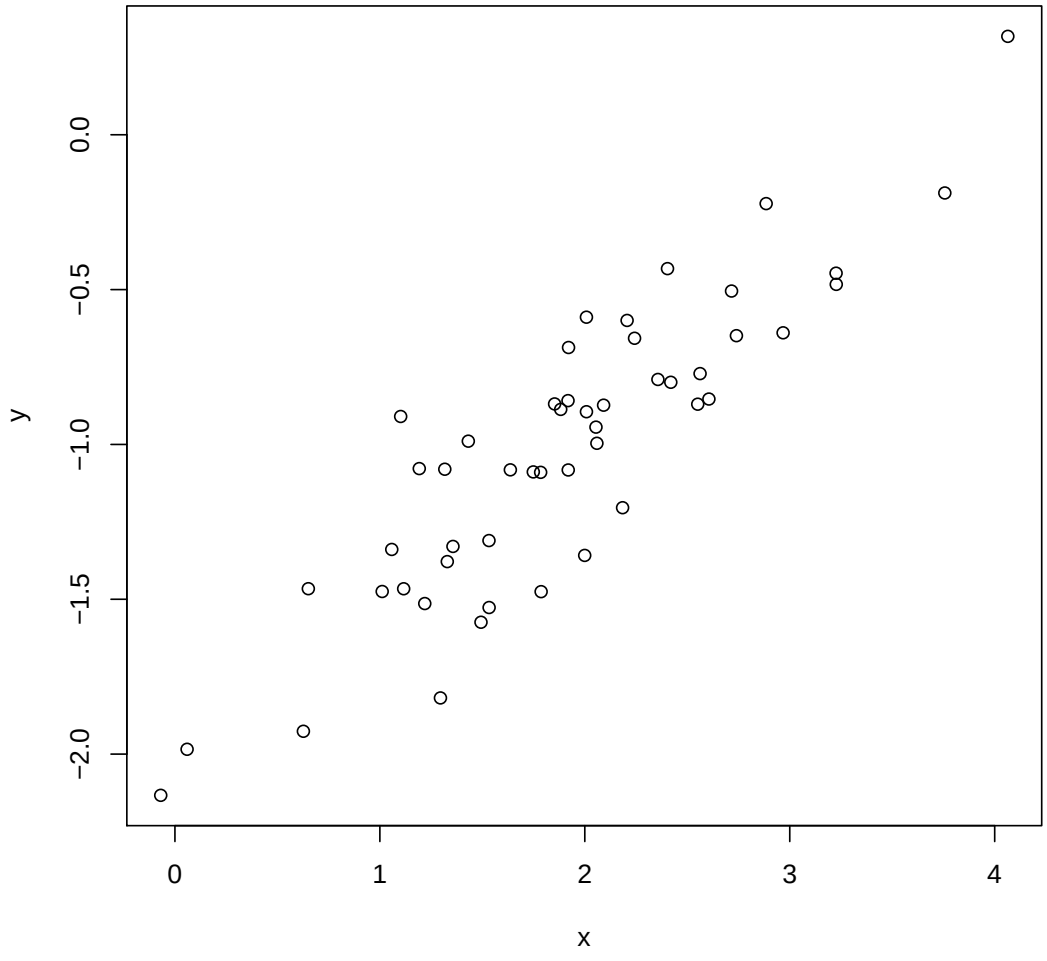
The `Rev` probabilistic programming language:

Most `Rev` scripts have two important aspects woven together:

1. specify the graphical model
2. define the inference algorithm (MCMC moves, etc.)

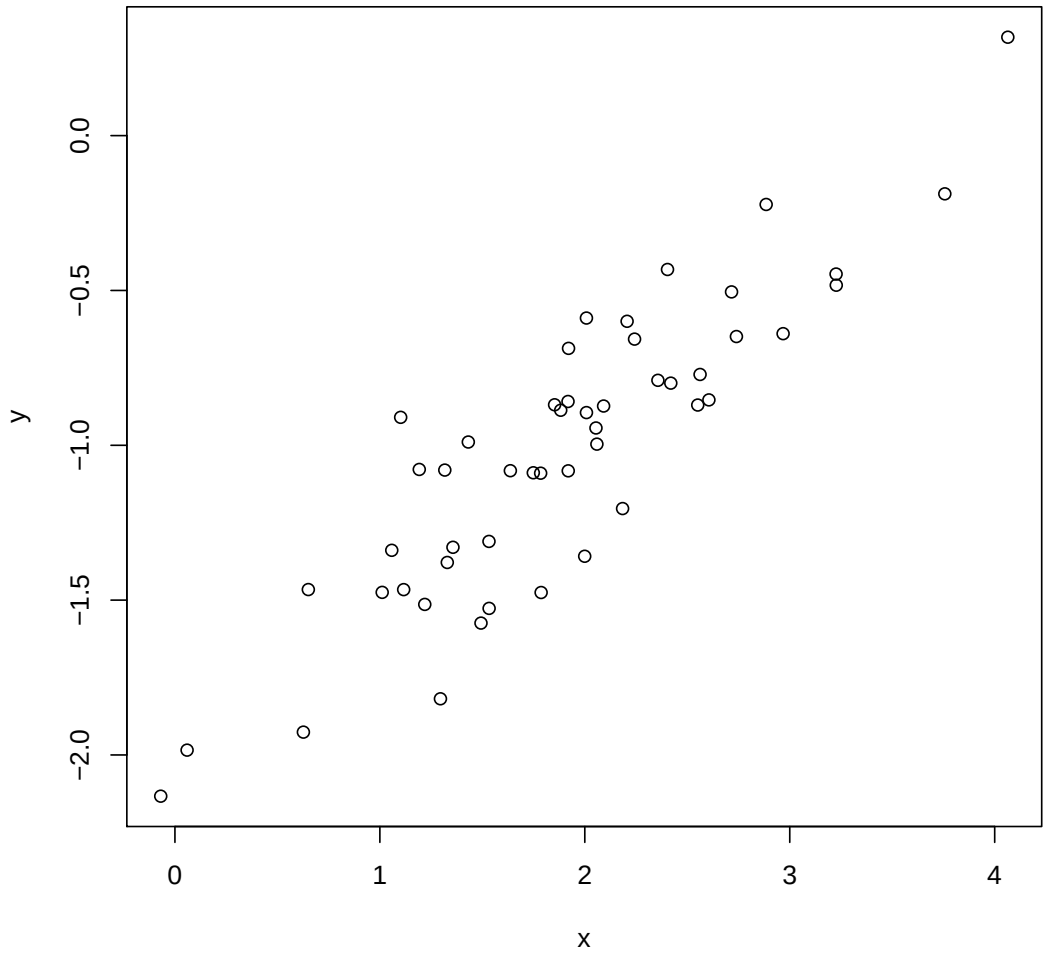
Since our goal is think abstractly in terms of graphical models we're going to learn these two aspects separately.

linear regression as a graphical model



$$y = \beta x + \alpha + \epsilon$$

linear regression as a graphical model



$$\mu_y = \beta x + \alpha$$
$$y \sim \text{Normal}(\mu_y, \sigma_\epsilon)$$

Bayesian linear regression

$$\mu_y = \beta x + \alpha$$
$$y \sim \text{Normal}(\mu_y, \sigma_\epsilon)$$

We need some priors!

$$\beta \sim \text{Normal}(\mu = 0, \sigma^2 = 1)$$
$$\alpha \sim \text{Normal}(\mu = 0, \sigma^2 = 1)$$
$$\sigma_\epsilon \sim \text{Exponential}(\lambda = 1)$$