A very small intro to CTMC in phylogenetics

Rosana Zenil-Ferguson

University of Hawai'i

December 2019

In a Bayesian framework

We are always interested in knowing the posterior distribution

$$P(\theta|D) \propto P(D|\theta)\,P(\theta)$$

Example

$X(t) =$ Evolution of flower color at time changes according time t

Example

$X(t) =$ Evolution of flower color at time changes according time t

M is our Model (a.k.a.the hypothesis)

Blue flowers evolve to red and viceversa

$$\theta = (q_{BR}, q_{RB})$$

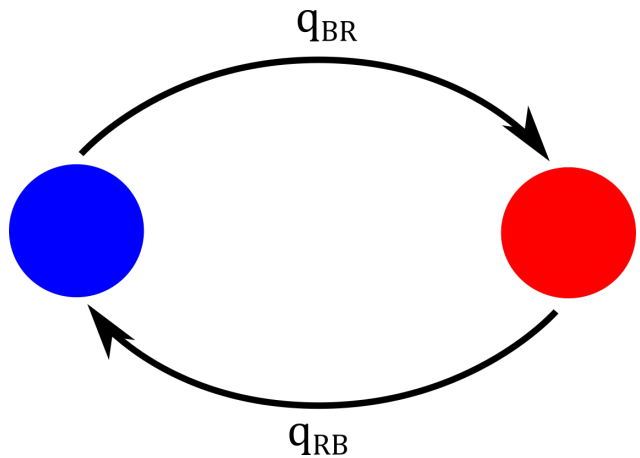
Example

$X(t) =$ Evolution of flower color at time changes according time t

M is our Model (a.k.a.the hypothesis)

Blue flowers evolve to red and viceversa

$$\theta = (q_{BR}, q_{RB})$$



The assumptions of our model in Bayesian framework

- In Bayesian framework: q_{BR}, and q_{RB} are **unknown and random variables** (they have a probability)

The assumptions of our model in Bayesian framework

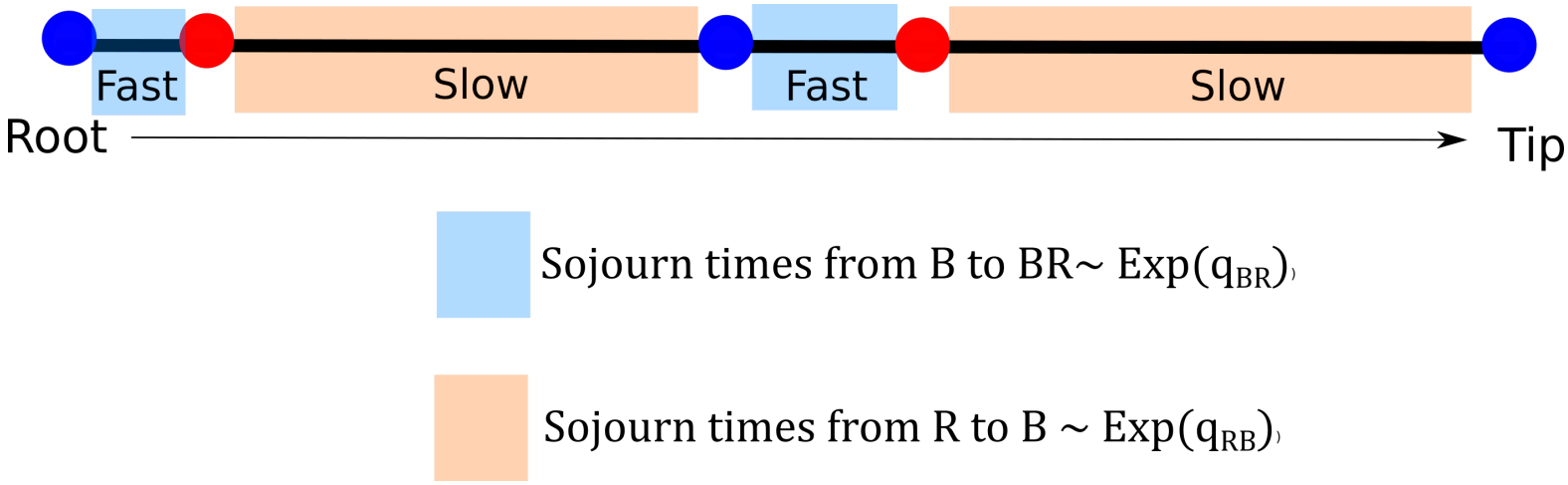
- In Bayesian framework: q_{BR}, and q_{RB} are **unknown and random variables** (they have a probability)
- q_{BR} and q_{RB} are **instantaneous rates**.

The assumptions of our model in Bayesian framework

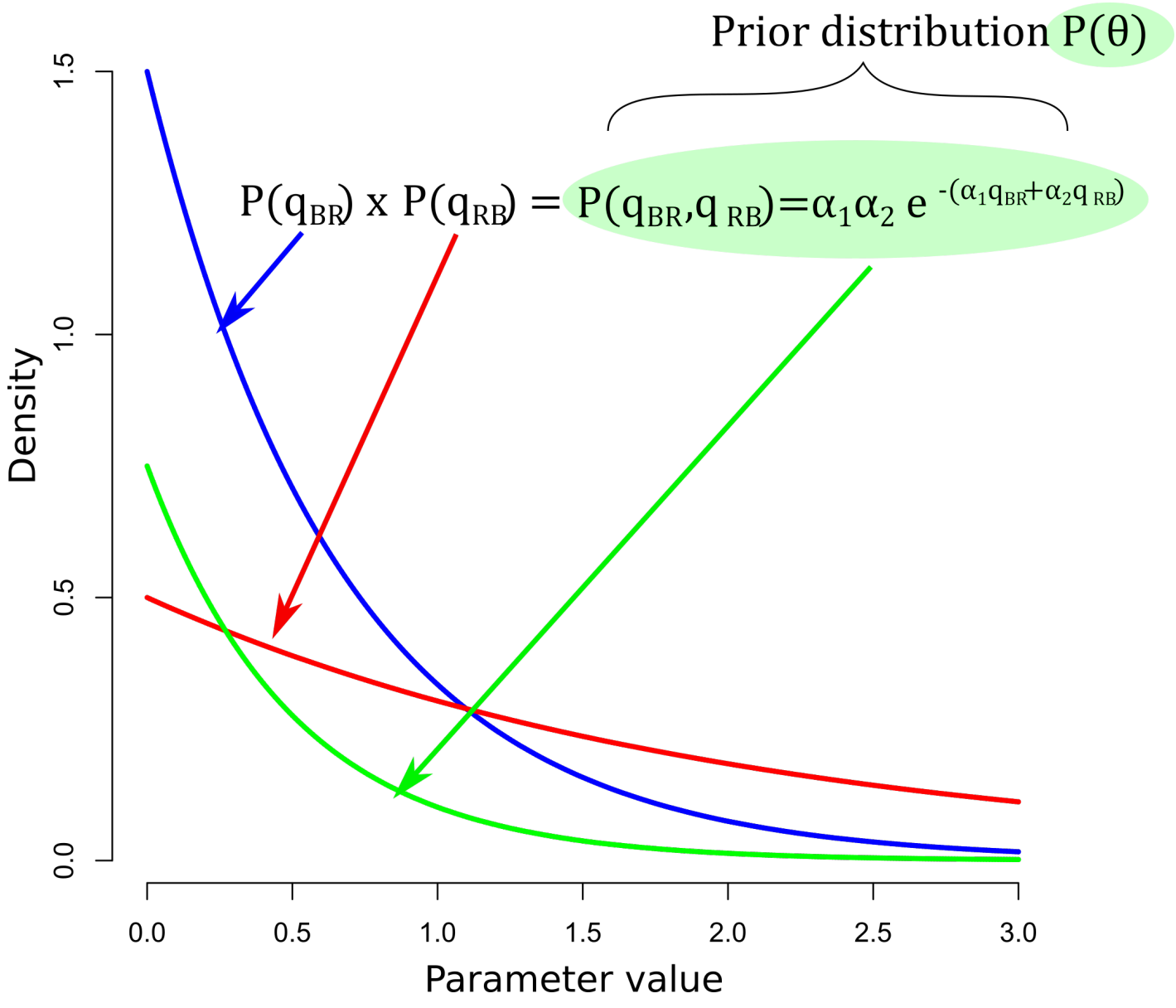
- In Bayesian framework: q_{BR}, and q_{RB} are **unknown and random variables** (they have a probability)
- q_{BR} and q_{RB} are **instantaneous rates**.

The assumptions of our model in Bayesian framework

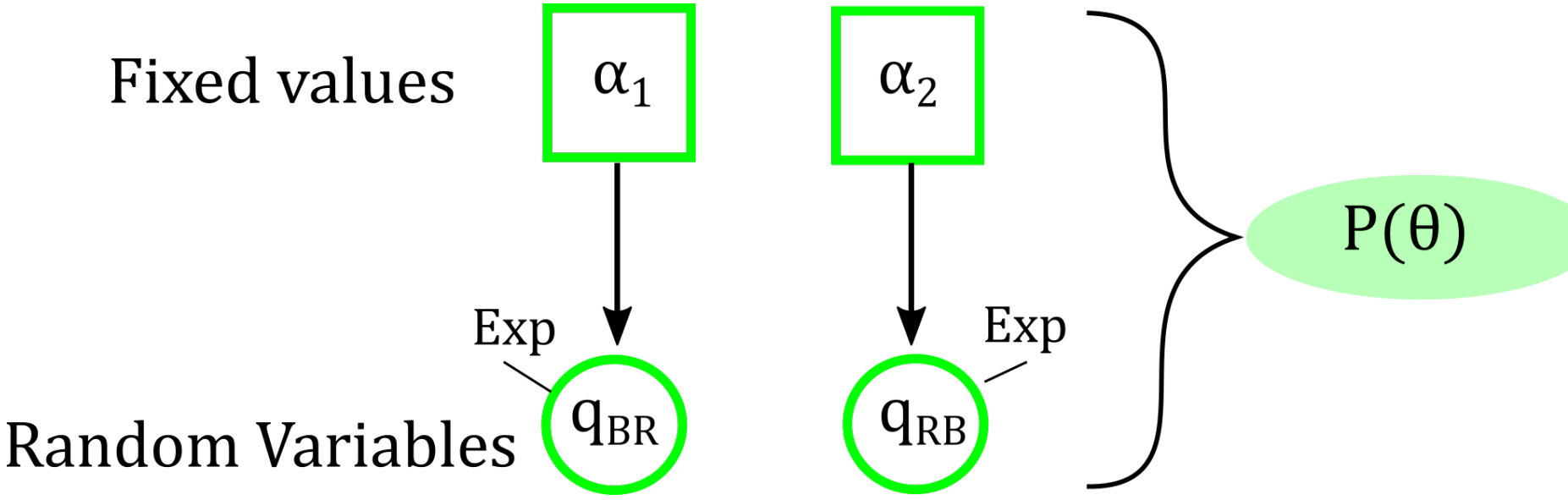
- In Bayesian framework: q_{BR}, and q_{RB} are **unknown and random variables** (they have a probability)
- q_{BR} and q_{RB} are **instantaneous rates**.



The prior distribution: $P(\theta)$



How are these assumptions represented graphically?



D is our data

We go into our favorite herbarium, field site, or green house and we collect color of multiple species

How do we integrate our model θ and our data D ?

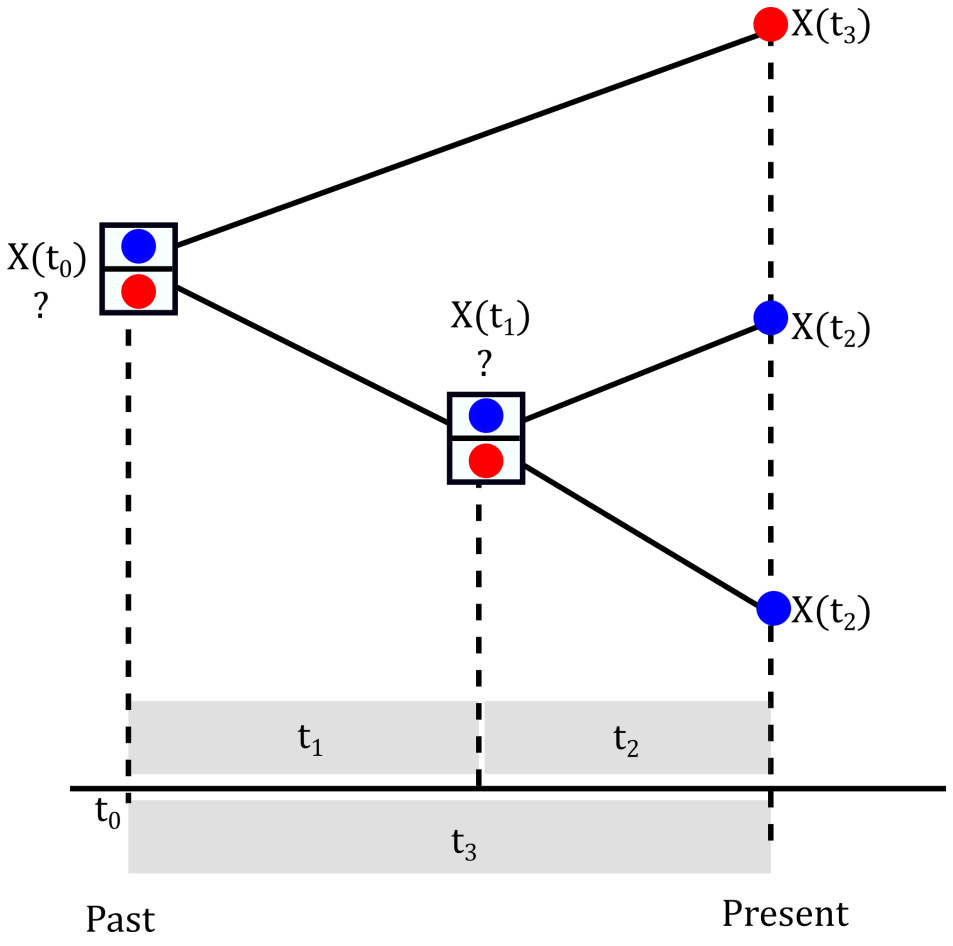
Calculating the likelihood $P(D|\theta)$

- We assume a phylogenetic tree Ψ (for this example is fixed)

Calculating the likelihood $P(D|\theta)$

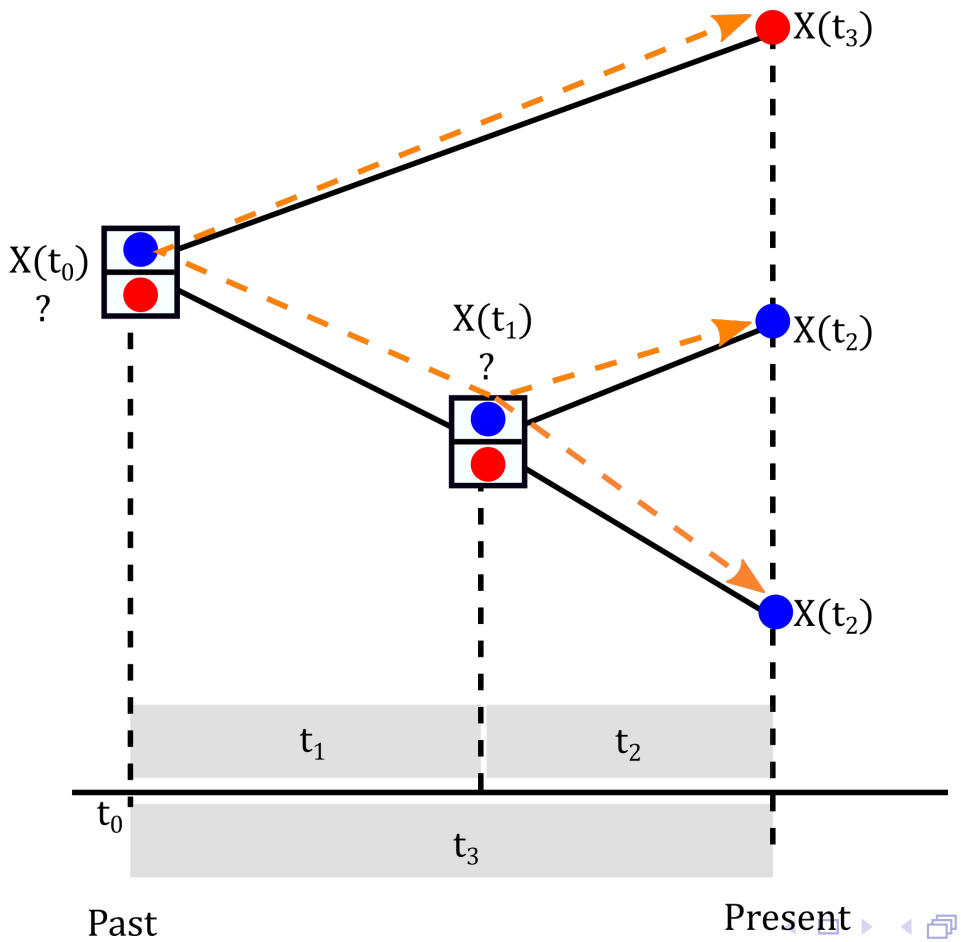
- We assume a phylogenetic tree Ψ (for this example is fixed)
- **Data**: a sample of red and blue flowers on the tips of our phylogeny tree

Likelihood function: The probability of the sample given our hypothesis θ



The probability of a single possible story in phylogenetics

$$P(X(t_2) = B|X(t_1) = B)P(X(t_2) = B|X(t_1) = B)\times$$
$$\times P(X(t_1) = B|X(t_0) = B)P(X(t_3) = R|X(t_0) = B)P(X(t_0) = B)$$



Calculating the likelihood is computationally challenging

- Felsenstein (1981)= Pruning algorithm, reduces the complexity in the calculation.

Calculating the likelihood is computationally challenging

- Felsenstein (1981)= Pruning algorithm, reduces the complexity in the calculation.
- Reminder: Optimizations to find maximum likelihood estimates and confident intervals require challenging numerical algorithms

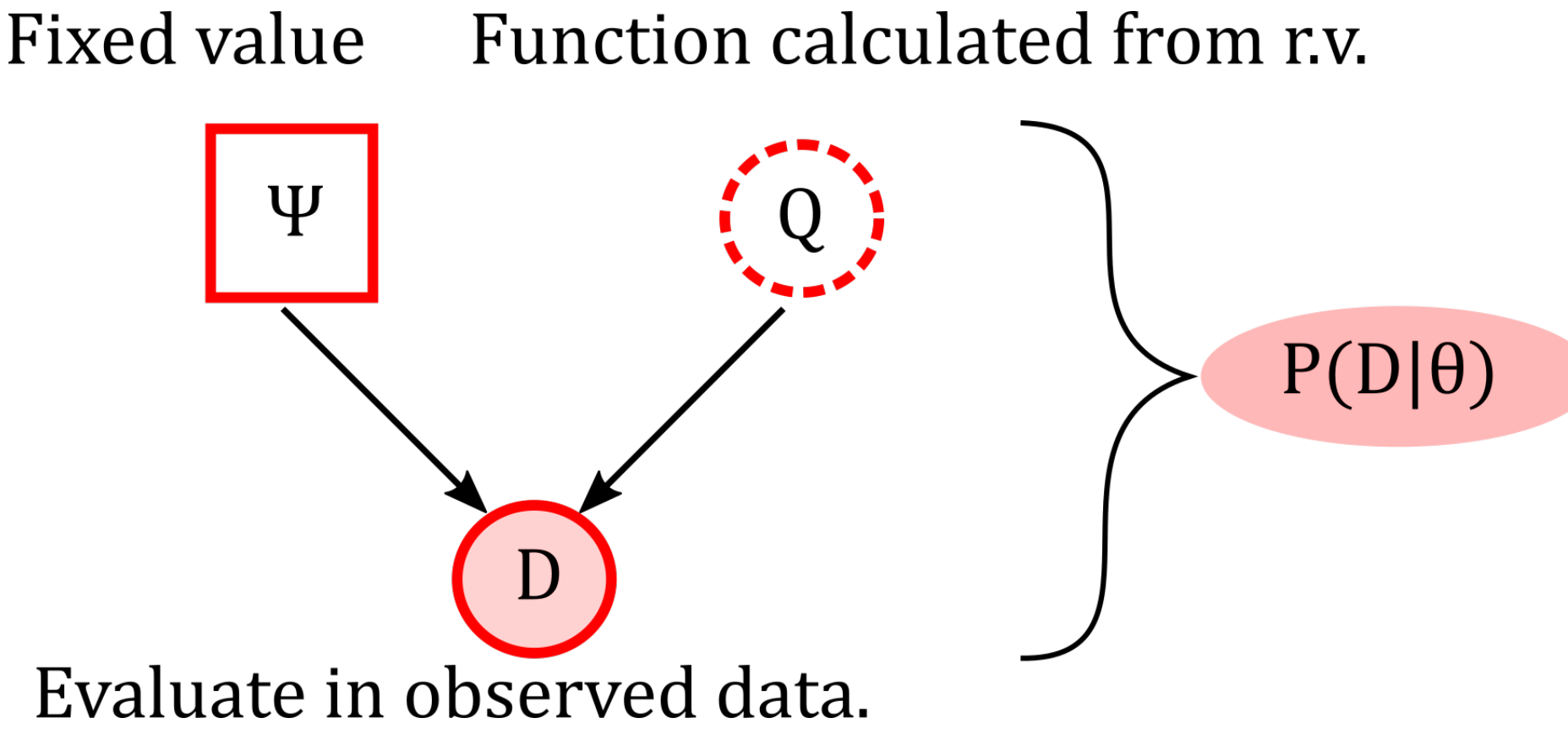
How do the rates connect with the probabilities?

Q-matrix= The infinitesimal probability matrix is the derivative of the probability

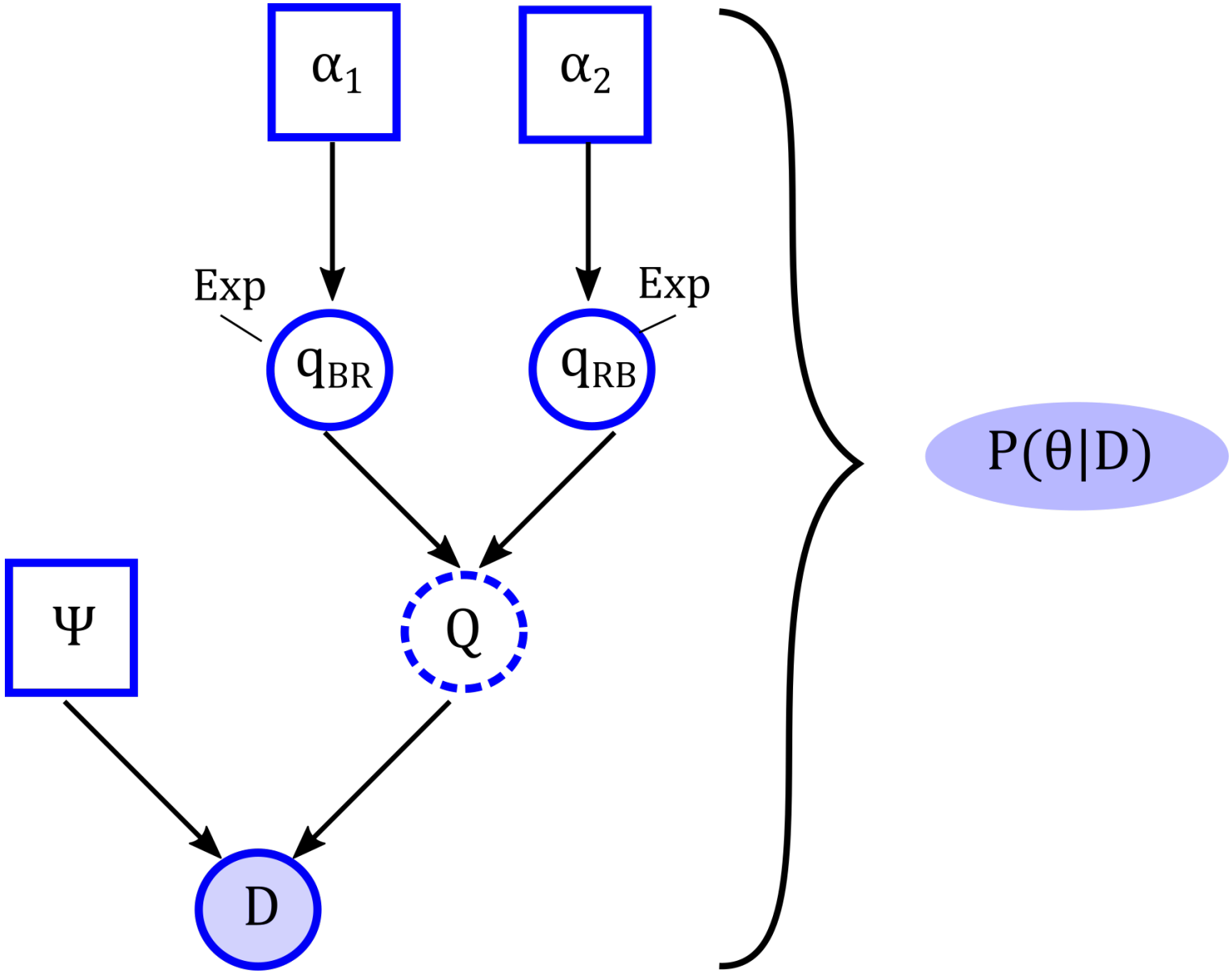
$$\frac{dP(t)}{dt} = Q$$

$$Q = \begin{pmatrix} -q_{BR} & q_{BR} \\ q_{RB} & -q_{RB} \end{pmatrix} \qquad P(t) = e^{Qt}$$

Likelihood in graphical form $P(D|\theta)$



The posterior distribution: the model conditional to the observed data



Graphical model benefits

- **Explicit notation**: In RevBayes we have notation for fixed variables, random variables, observed data, deterministic function,...

Graphical model benefits

- **Explicit notation**: In RevBayes we have notation for fixed variables, random variables, observed data, deterministic function,...
- **Modularity**: Once I have built a model I can connect other as a module (building blocks!)