Phylogenetic Graphical Models and RevBayes: Introduction

Fred(rik) Ronquist

Swedish Museum of Natural History, Stockholm, Sweden

Statistical Phylogenetics

- Statistical approaches increasingly important:
  - Difficult problems requiring accurate and unbiased inference (e.g., structure of rapid radiations)
  - More aspects of molecular evolution being examined (structural dependencies, etc)
  - Combination of background knowledge and sequence information (e.g., divergence time estimation)
- Modeling explosion, especially in the Bayesian context
- Challenging for empiricists to communicate and correctly understand models
- Challenging for developers of inference software

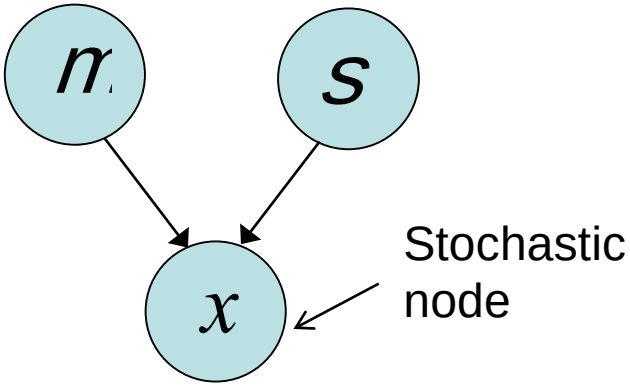
Probabilistic Graphical Models

- Theoretical framework for specifying dependencies in complex statistical models
- Allows a complex model to be broken down into conditionally independent distributions
- Closely related to standard statistical model formulae:

$$x \sim \text{Norm}(\mu, \sigma)$$

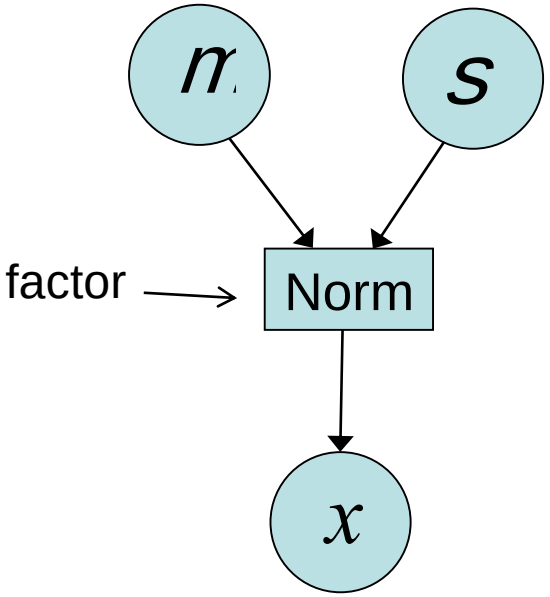
- Extensive literature on generic algorithms that apply to model graphs

$$x \sim \text{Norm}(m, s)$$

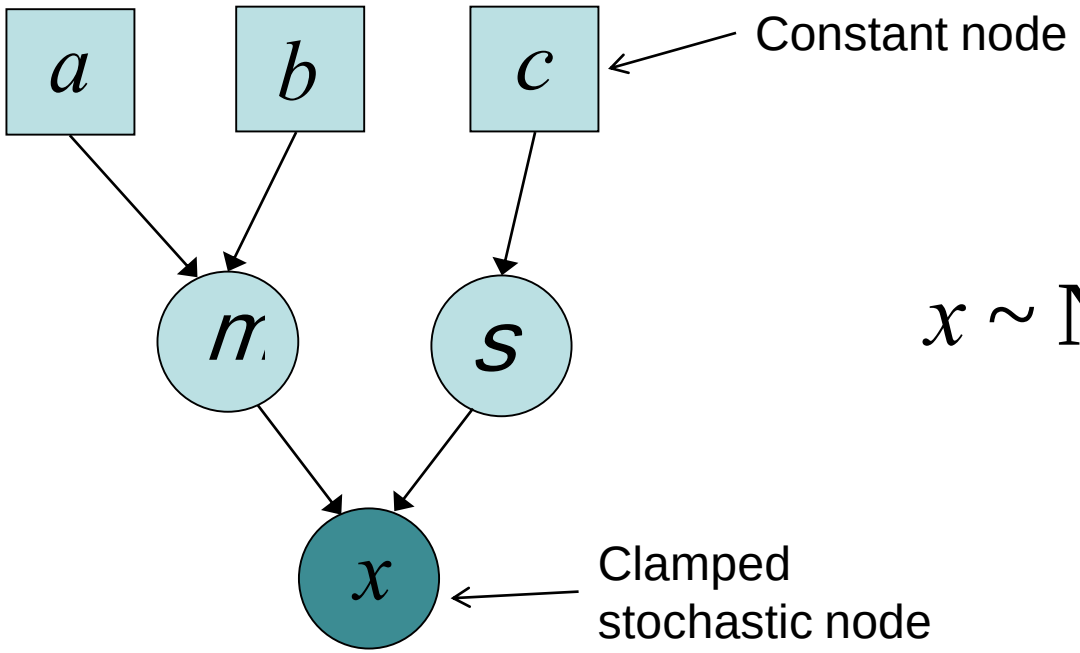


Graphical Model

Compact Form

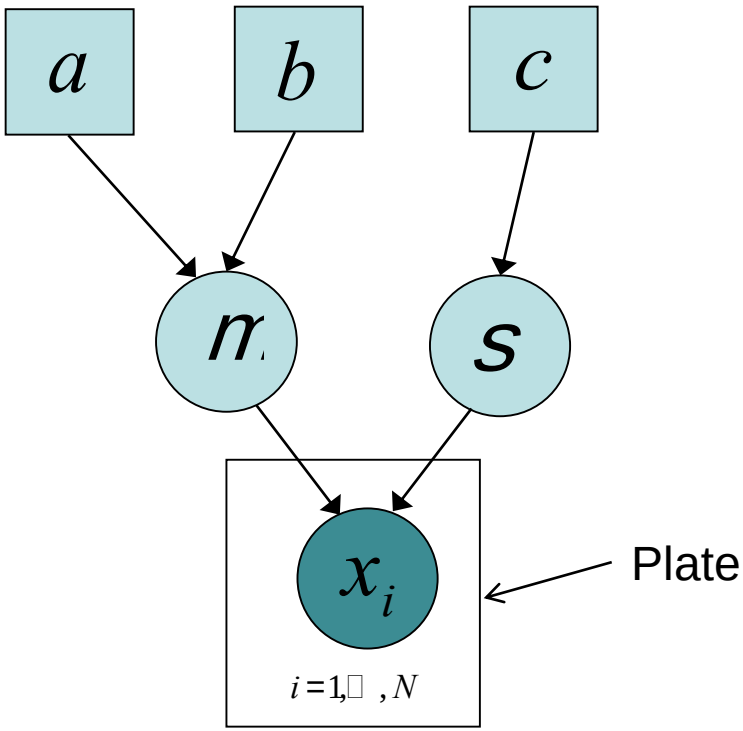


Factor Graph



$$x \sim \text{Norm}(m, s)$$

Hierarchical Graphical Model



$$x \sim \text{Norm}(m, s)$$

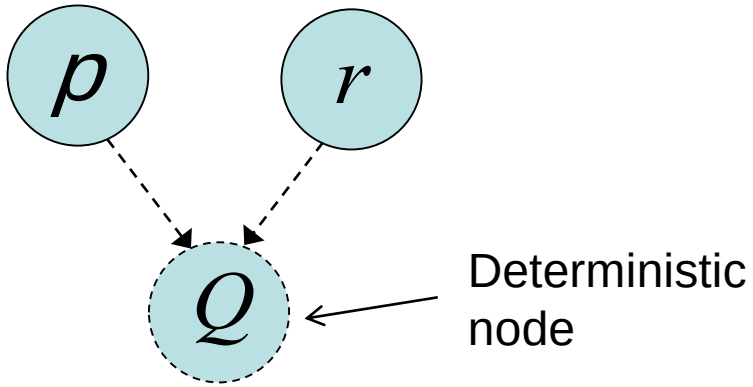
Hierarchical Graphical Model

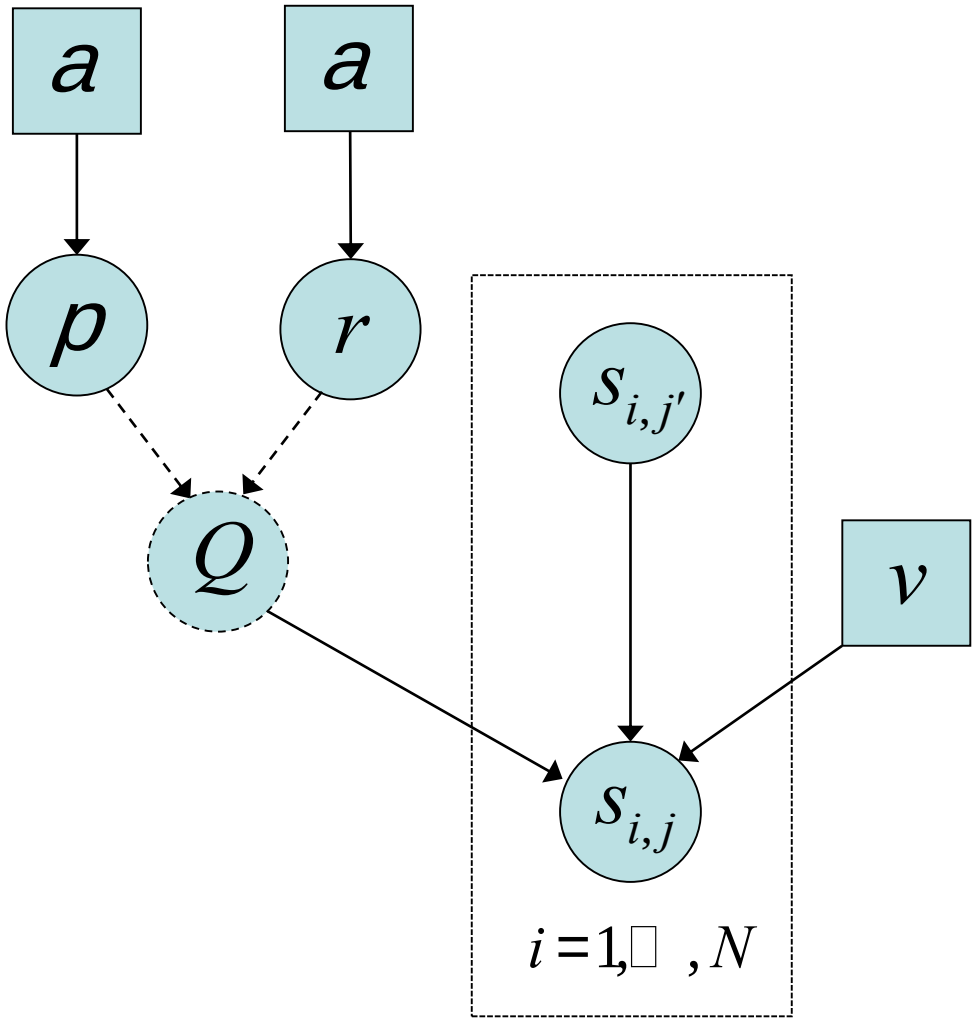
$$Q = \begin{pmatrix} - & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & - & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & - & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & - \end{pmatrix}$$

General Time Reversible (GTR) substitution model

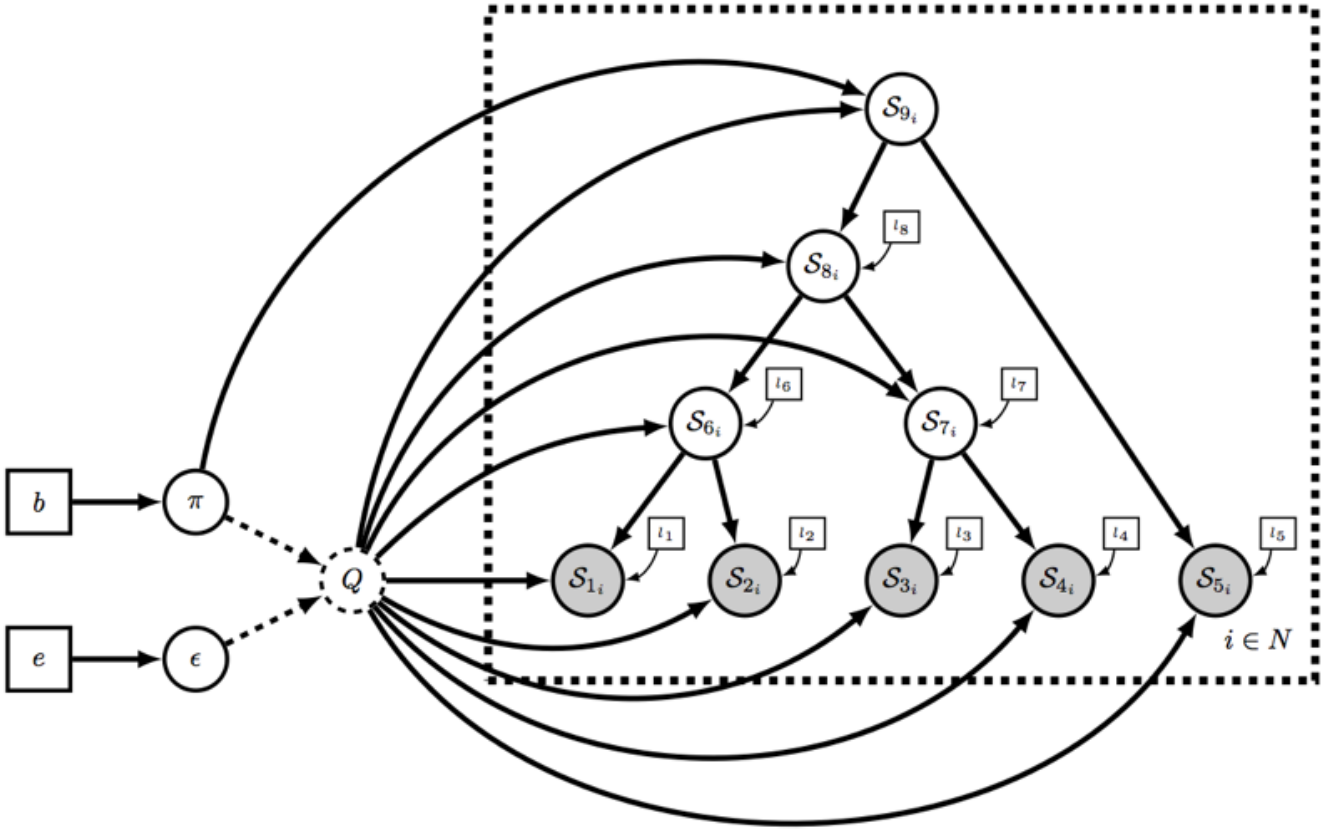
π Stationary state frequencies

r Exchangeability rates

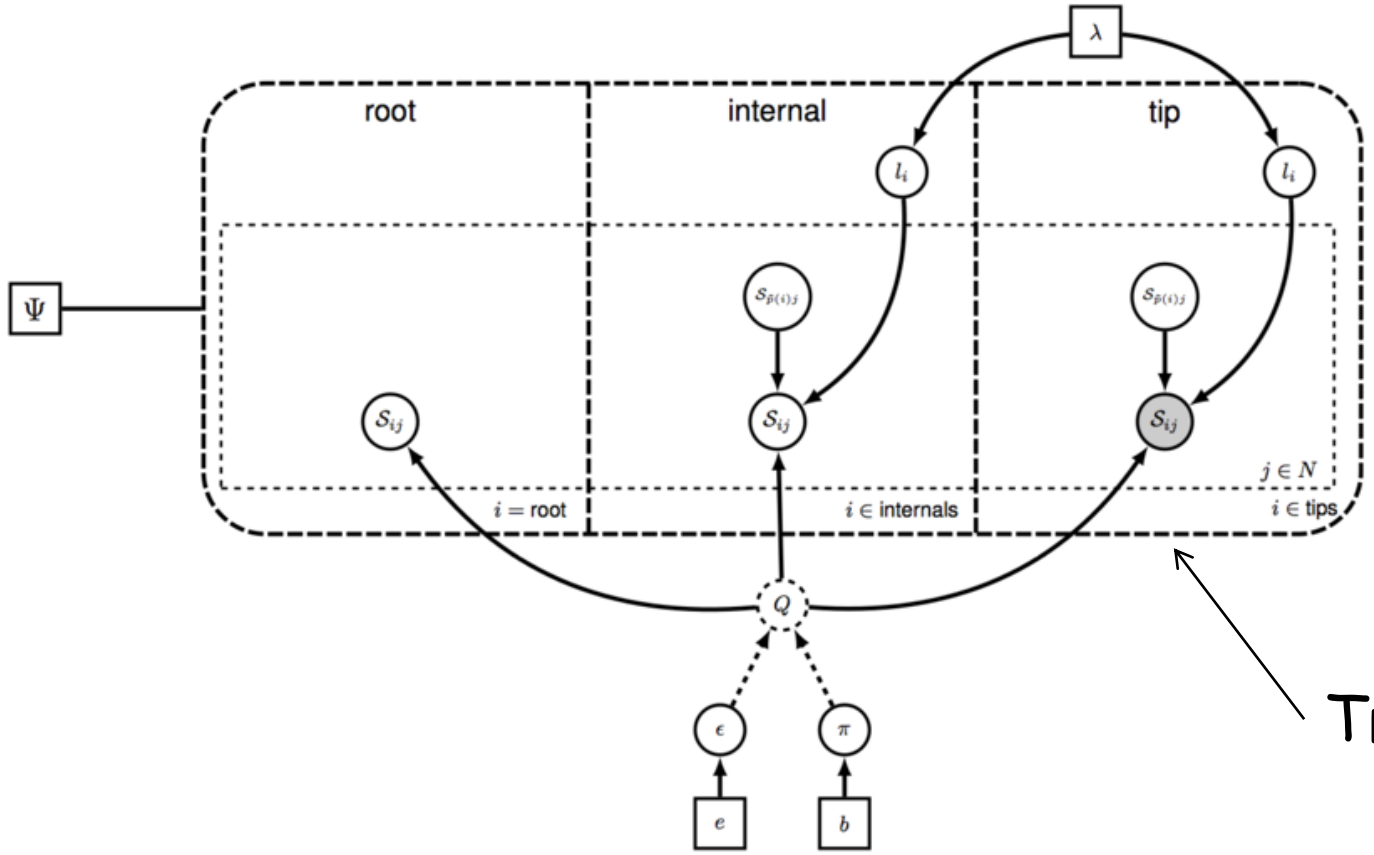

p
r
Q
Deterministic
node


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

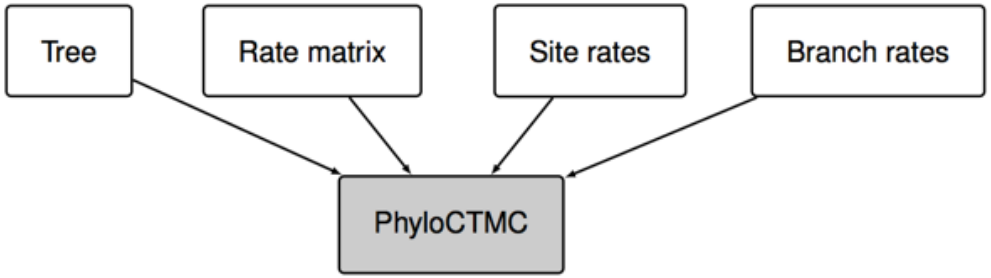
GTR Phylogeny Model



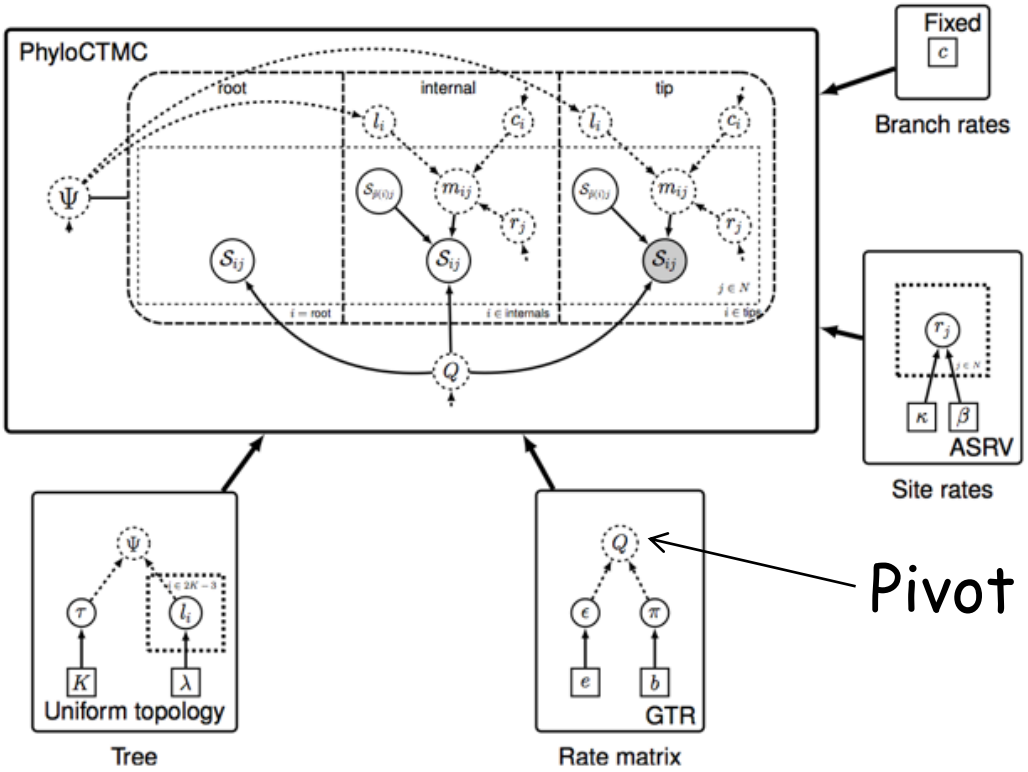
Tree Plate Representation



Tree plate



Modular Representation



Pivot variable

RevBayes Project

- Interactive computing environment intended primarily for Bayesian phylogenetic inference
- Uses a special language, Rev, for constructing probabilistic phylogenetic and evolutionary graphical models interactively, step by step
- Rev is similar to R and the BUGS modeling language
- RevBayes provides generic computing machinery for simulation, inference and model testing

Basic properties of the Rev language

There are three kinds of statements in the language

1. Arrow assignment (value assignment, create constant nodes)

```
> a <- 4              # Give a the value 4
> b <- sqrt(a)        # Giva b the value of sqrt(a), that is, 2
> b                   # Print the value of b
 2
```

2. Equation assignment (create deterministic nodes)

```
> c := sqrt(a)        # Make c a dynamic function node evaluating sqrt(a)
> c
 2
> a <- 9              # Give a the value 9
> b                   # Print the value of b
 2
> c                   # Print the value of c
 3
```

Basic properties of the Rev language

3. Tilde assignment (create stochastic variables (nodes))

> a ~ dnExp(rate = x) # a is drawn from exp dist with rate = x

Basic properties of the Rev language

```
# -------------------------------
# Declaring and defining functions
# -------------------------------

> function foo ( x ) { x * x }
> foo( 2 )
 4

# If you wish, you can specify types as well

> function PosReal foo ( Real x ) { x * x }

# Without explicit types, RevObject is the assumed type


# -------------------------------
# Declaring and defining new types
# -------------------------------

> class myclass : Move {
+     Real myTuningParam;
+     procedure Real move( Real x ) { myTuningParam * x }
+ }

# Inheritance, function overriding and overloading
```

A complete MCMC analysis in Rev

```
a <- -1.0
b <- 1.0

mu    ~ dnUnif(a, b)
sigma ~ dnExp(1.0)

for (i in 1:10) {
   x[i] ~ dnNorm(mu, sigma)
   x[i].clamp(0.5)
}

mymodel = model(mu)   # Any stochastic node in the model works

mymcmc = mcmc(mymodel)

mymcmc.run(1000)
```

```
# definition of the myGTR function (“Ziheng's favorite”)
function model myGTR (CharacterMatrix data) {

   # describe Q matrix
   pi ~ dflatdir(4);
   r  ~ dflatdir(6);
   Q := gtr(pi, r);

   # describe tree
   tau ~ dtopuni(data.taxa(), rooted=false);

   # gamma shape
   alpha ~ dunif(0.0, 50.0);

   # discrete gamma mixture
   for (i in 1:4)
        catRate[i] := qgamma(i*0.25-0.125, alpha, alpha);
   for (i in 1:data.size())
        ratecat[i] ~ dcat(simplex(0.25,0.25,0.25,0.25));

   # associate distributions with tree parts
   for (i in 1:data.size()) {
        for (n in 1:tau.numNodes()) {
             if (tau.isTerminal(n)) {
                  tau.length[n] ~ exp(1.0);
                  tau.state[n] ~ ctmc(Q, e.length*catRate[ratecat[i]],
                       tau.state[tau.parent(n));
                  tau.state[n] <- data[i][tau.tipIndex(n)];
             }
             else {
                  tau.length[n] ~ exp(10.0);
                  tau.state[n] ~ ctmc(Q, e.length*catRate[ratecat[i]],
                       tau.state[n]);
             }
        }
   }

   # return model
   return model( Q );
}
```

Definition of a new phylogenetic model

Appr. 20 lines

Complexity hidden from normal user

```
# Read in data
myData <- read( “data.nex” )

# Apply model
myModel = zihengGTR( myData )

# Construct mcmc
myMCMC = mcmc( myModel )

# Run mcmc
myMCMC.run(10000)
```

RevBayes

- RevBayes is still experimental software
- Help is still incomplete
- There may be various bugs and other problems, for instance related to type conversion
- A number of practical features are still missing
  - Post-hoc analysis very limited
  - No support for convergence diagnostics on the fly

RevBayes Limitations

- Coarse-grained representation of tree plates -> some models are impossible to specify
- No ability to manipulate modules
- No default moves or monitors
- Some programming features are missing, notably specification of user-defined types
- Most additions need to be made in the back end using heavily templated C++ code

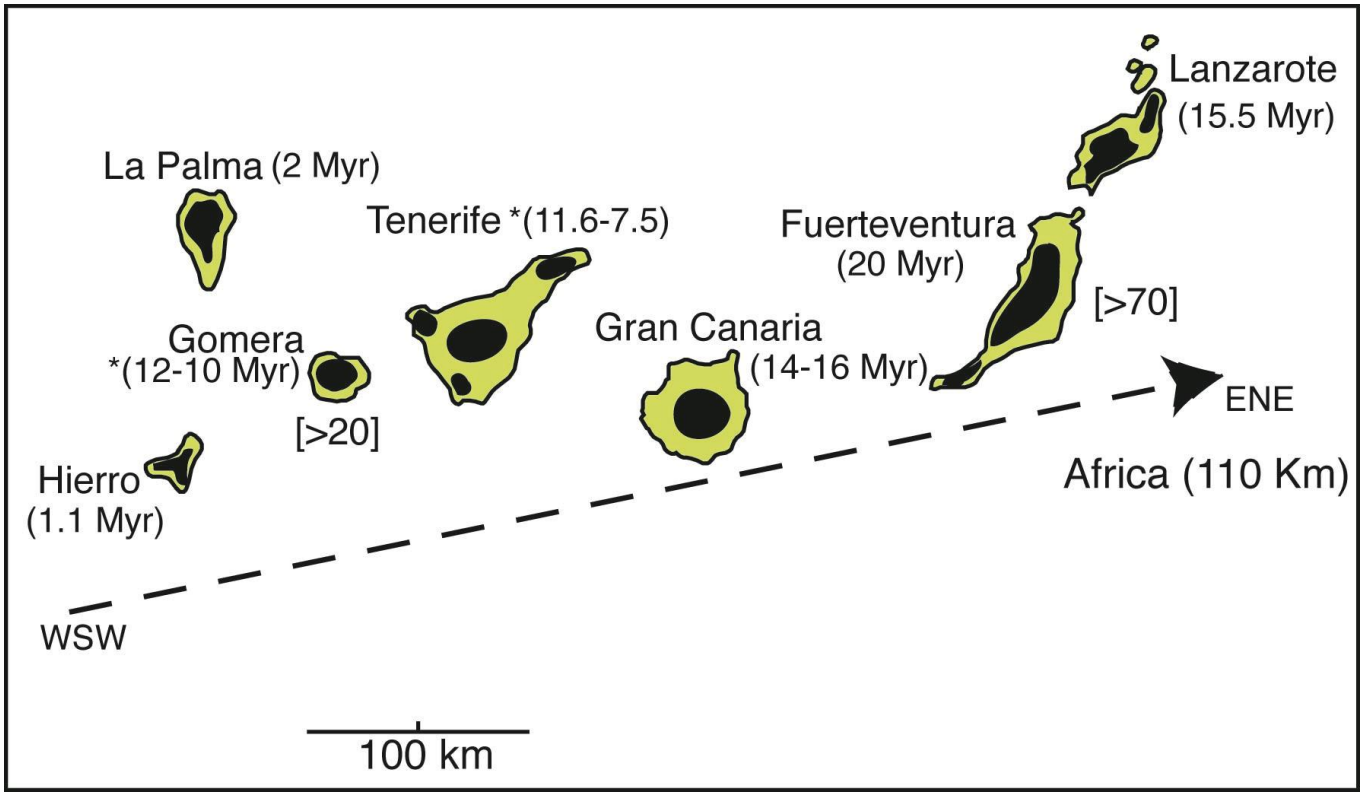
RevLang Project

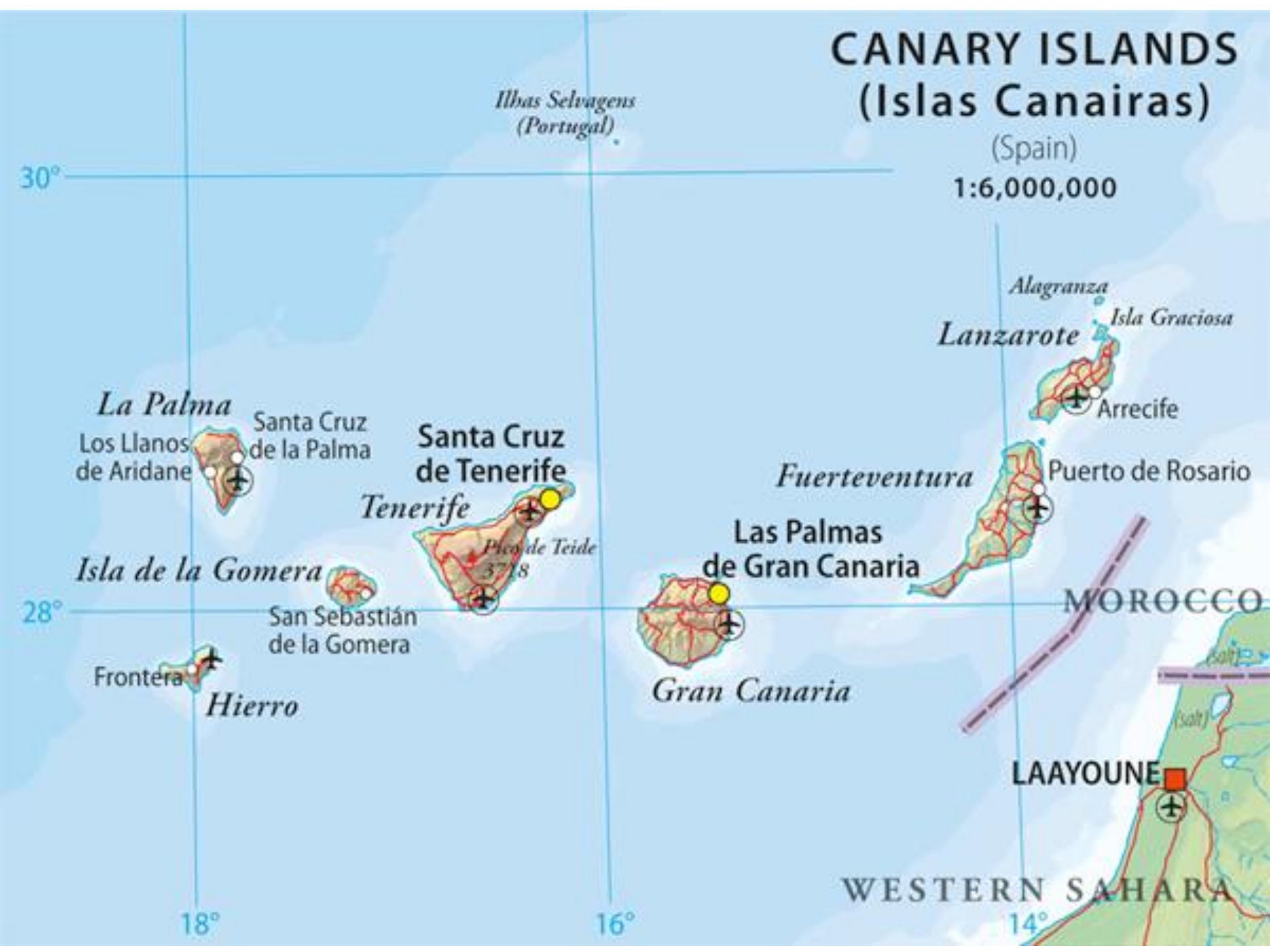
- Rev -> probabilistic programming language
- Moves, distributions etc programmed in Rev itself
- Professional JIT compiler
- Fully supported interactive environment, like R
- Modular design makes it easy to extend the environment

RevBayes Exercises

- Download and install RevBayes according to instructions at the RevBayes web site (http://revbayes.com).
- Download material and follow tutorials of interest
- RevBayes code at github: https://github.com/revbayes/revbayes

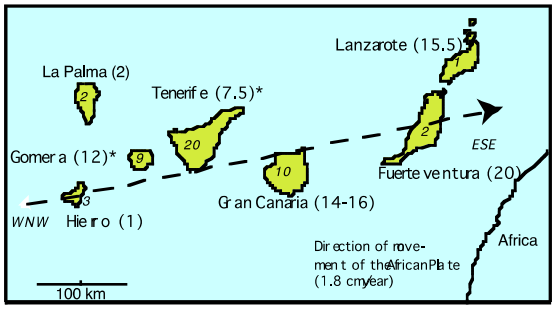
The Canary Islands

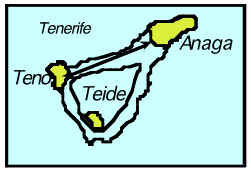



CANARY ISLANDS
(Islas Canairas)
(Spain)
1:6,000,000
Ilhas Selvagens
(Portugal)
30°
La Palma
Los Llanos
de Aridane
Santa Cruz
de la Palma
Santa Cruz
de Tenerife
Tenerife
Pico de Teide
3718
Isla de la Gomera
San Sebastián
de la Gomera
28°
Frontera
Hierro
Las Palmas
de Gran Canaria
Gran Canaria
Alagranza
Isla Graciosa
Lanzarote
Arrecife
Fuerteventura
Puerto de Rosario
MOROCCO
LAAYOUNE
WESTERN SAHARA
18°
16°
14°

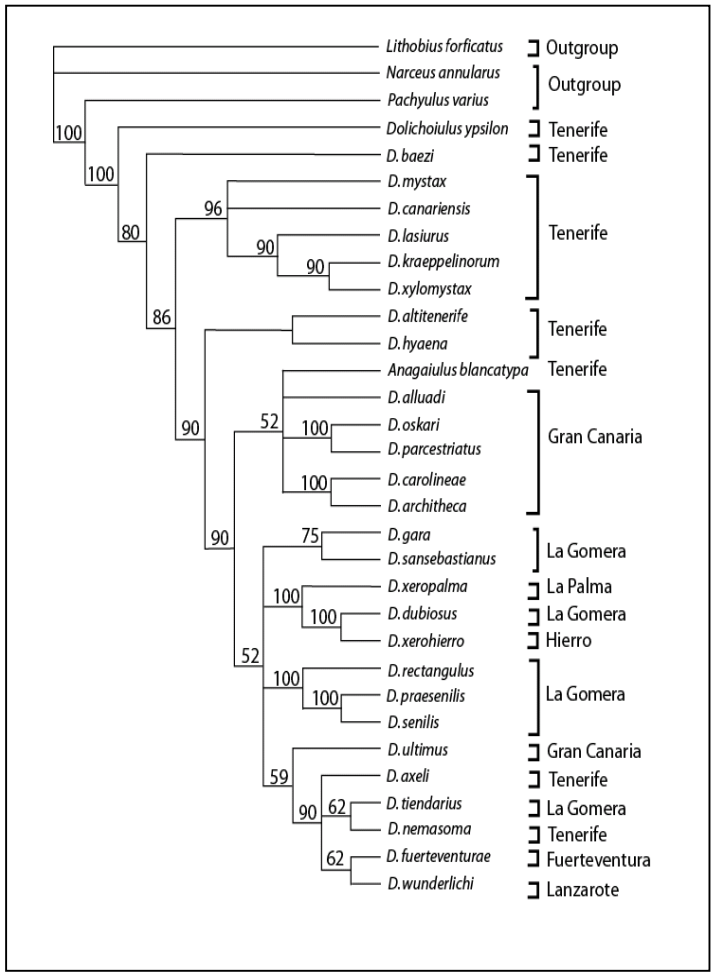
Dolichoiulus (Diplopoda)

Dolichoiulus (Diplopoda, Julida, Julidae, Pachyulinae)

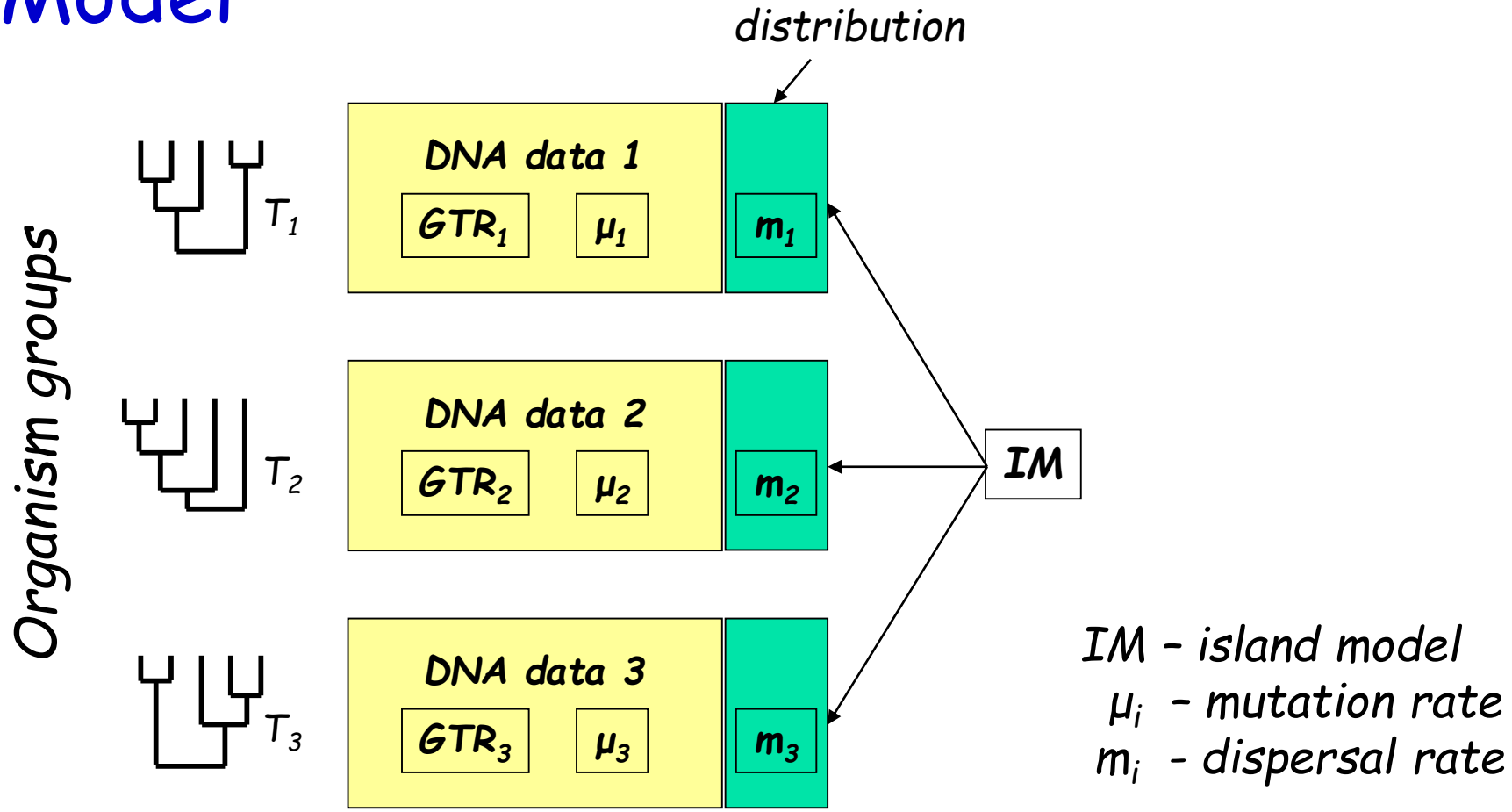




46 endemic species



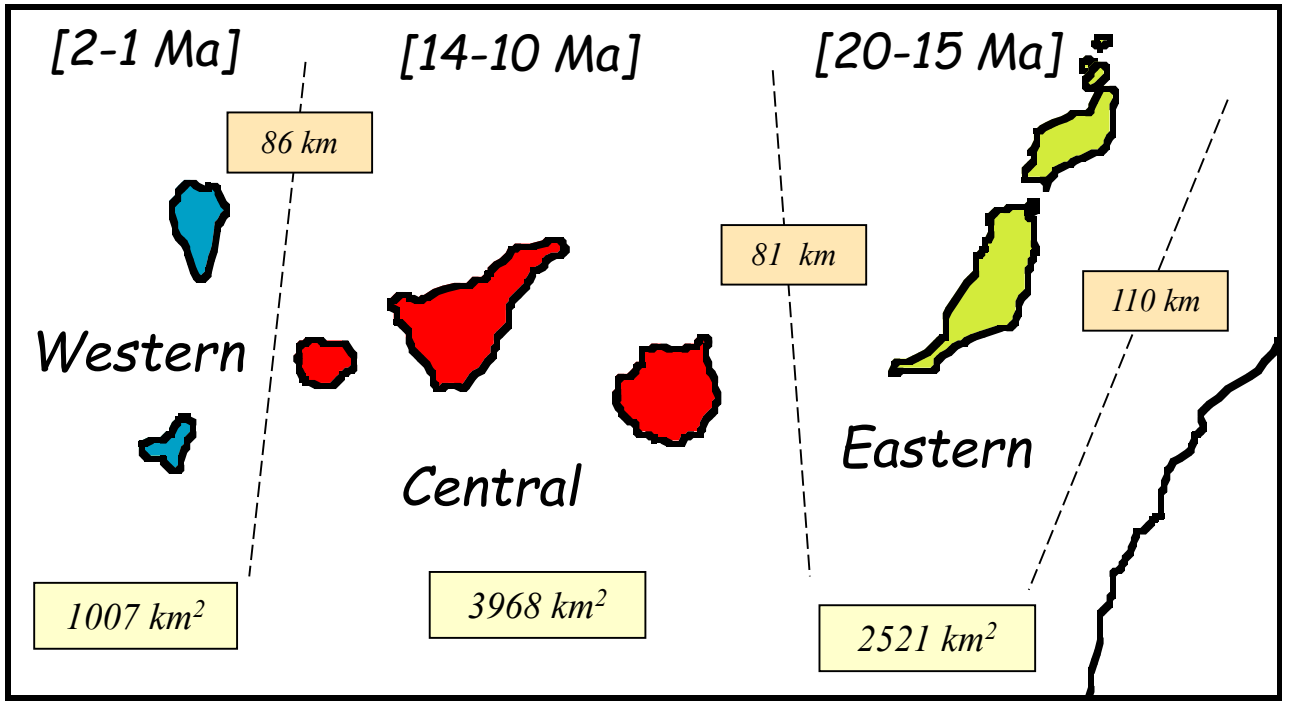
Model



Inference

Bayesian inference using MCMC sampling, accommodating uncertainty in all model parameters

Canary Islands: 3-island model



Instantaneous rate matrix

to

from

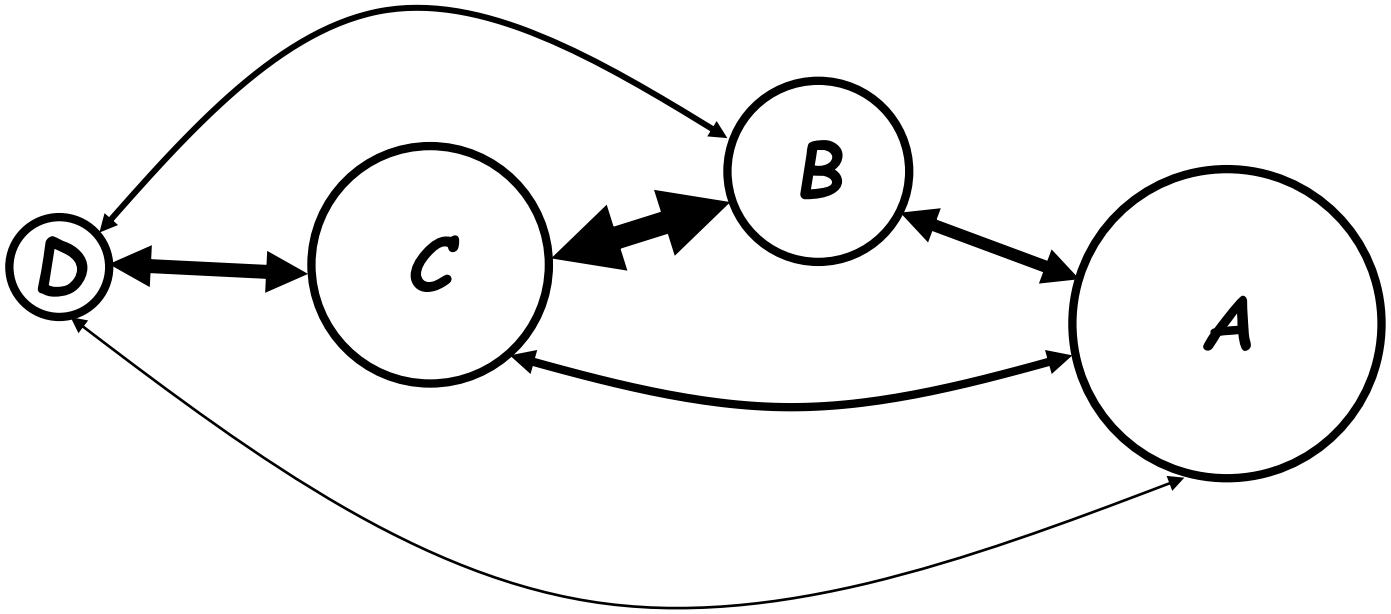
$$Q = \begin{array}{c} \\ [\mathrm{A}] \\ [\mathrm{B}] \\ [\mathrm{C}] \\ [\mathrm{D}] \end{array} \begin{pmatrix} [\mathrm{A}] & [\mathrm{B}] & [\mathrm{C}] & [\mathrm{D}] \\ - & \pi_{\mathrm{B}} r_{\mathrm{AB}} & \pi_{\mathrm{C}} r_{\mathrm{AC}} & \pi_{\mathrm{D}} r_{\mathrm{AD}} \\ \pi_{\mathrm{A}} r_{\mathrm{AB}} & - & \pi_{\mathrm{C}} r_{\mathrm{BC}} & \pi_{\mathrm{D}} r_{\mathrm{BD}} \\ \pi_{\mathrm{A}} r_{\mathrm{AC}} & \pi_{\mathrm{B}} r_{\mathrm{BC}} & - & \pi_{\mathrm{D}} r_{\mathrm{CD}} \\ \pi_{\mathrm{A}} r_{\mathrm{AD}} & \pi_{\mathrm{B}} r_{\mathrm{BD}} & \pi_{\mathrm{C}} r_{\mathrm{CD}} & - \end{pmatrix}$$

π_i *Relative carrying capacity of island i*

r_{ij} *Relative dispersal rate between islands i and j*

Island "GTR" model

Time reversible continuous time Markov chain



Differing relative carrying capacities of islands
Differing intensities of biotic exchange between islands







Laura Martinez

Canary Island flora

15 groups, 567 lineages

Javier Fuertes







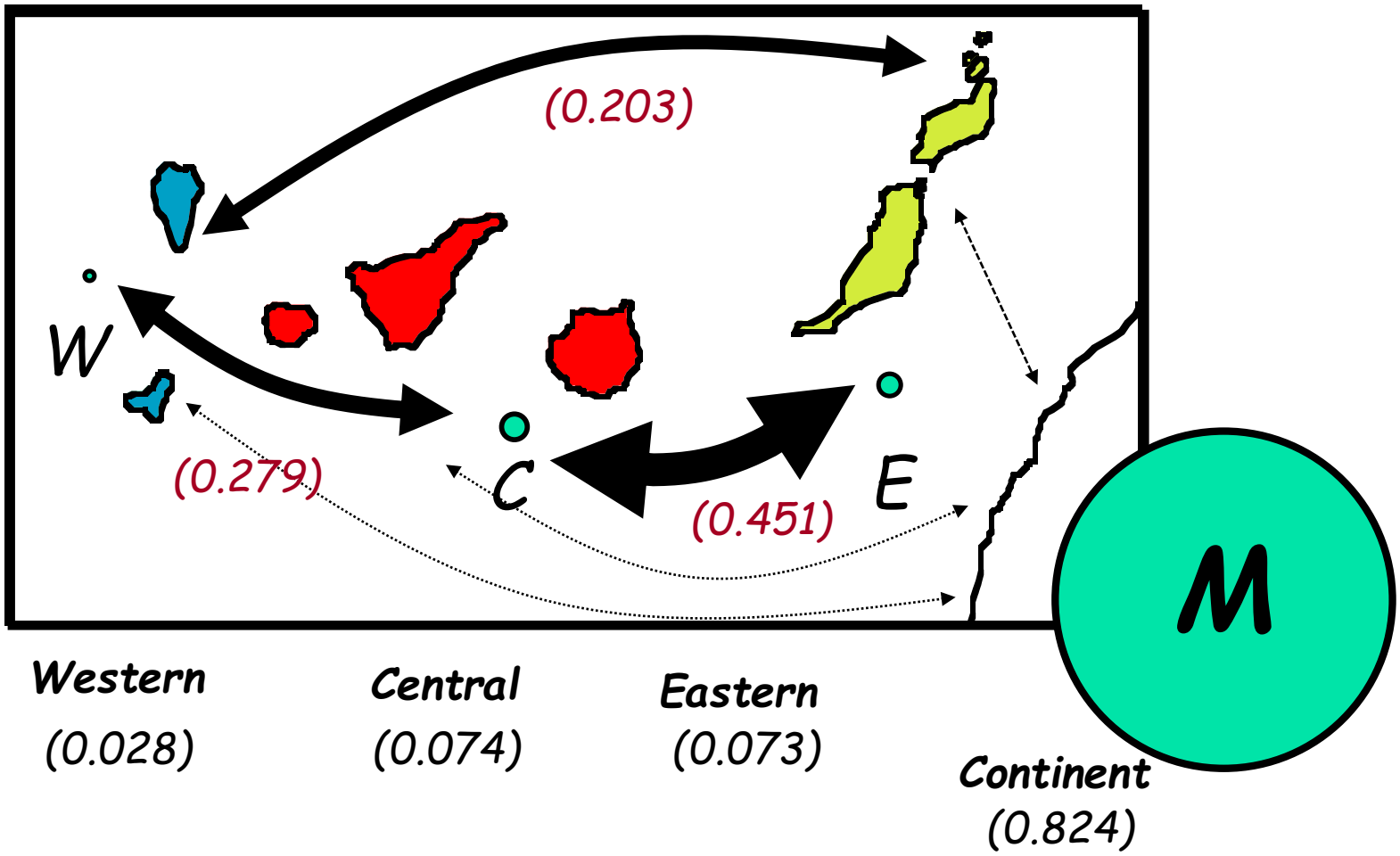
Raquel Martin

Canary Island fauna

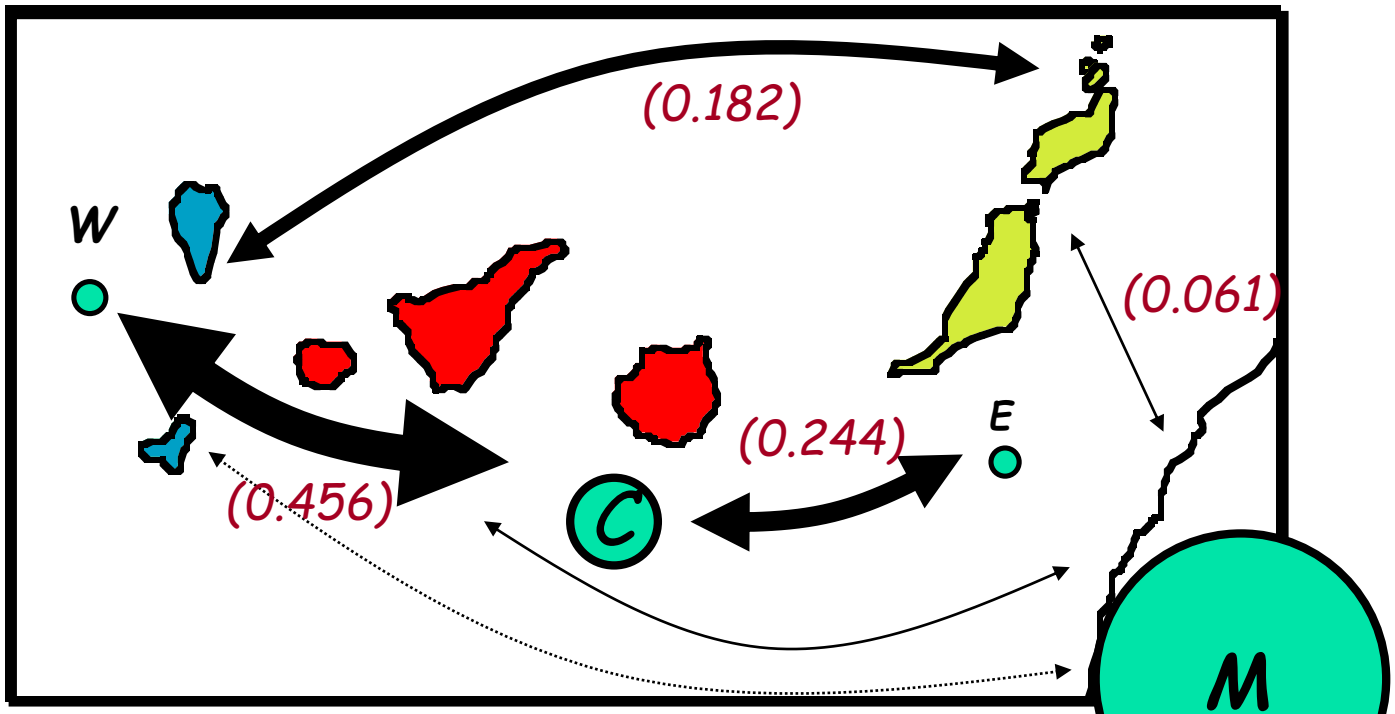
19 groups, 578 lineages

Javier Fuertes

Colonization patterns (plants)

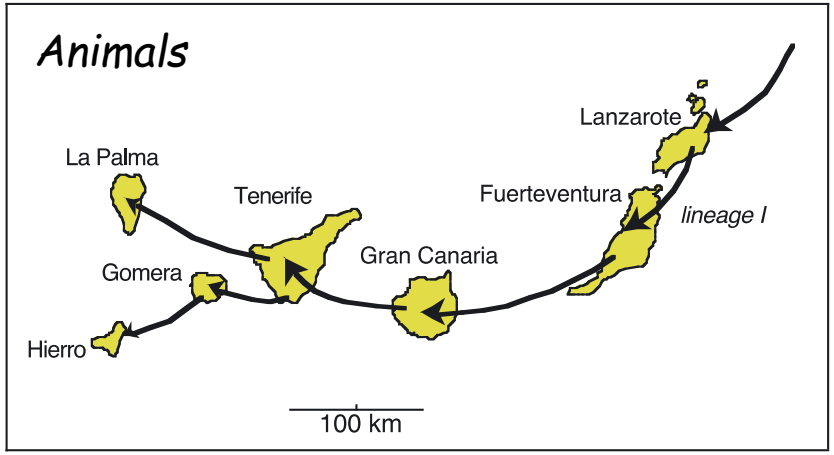


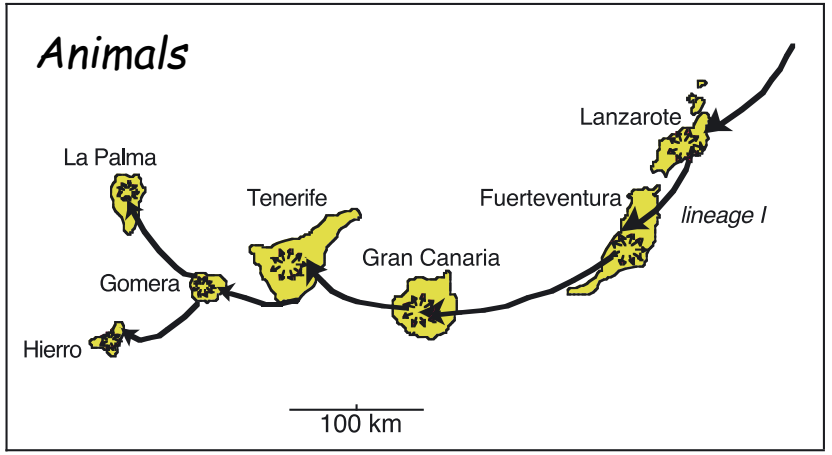
Colonization patterns (animals)

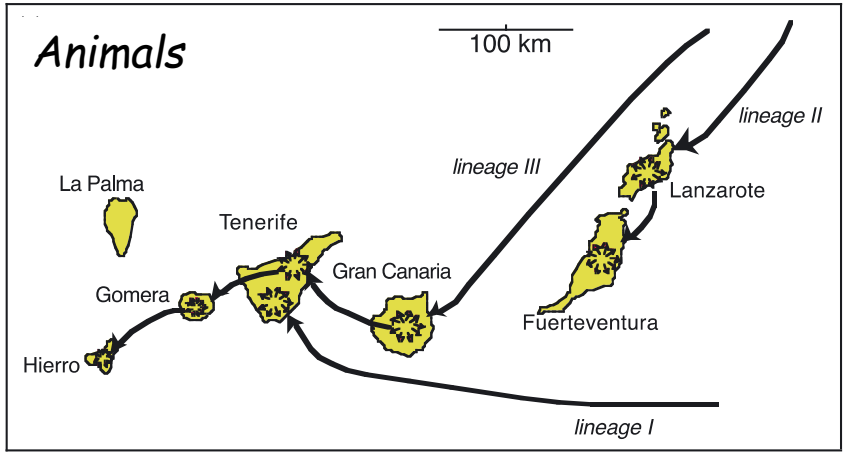


Western	Central	Eastern	Continent
(0.083)	(0.205)	(0.059)	(0.653)

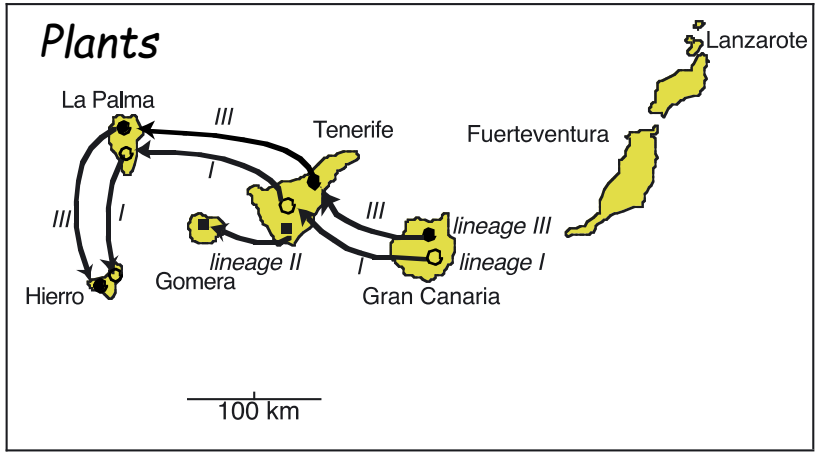
Colonization models



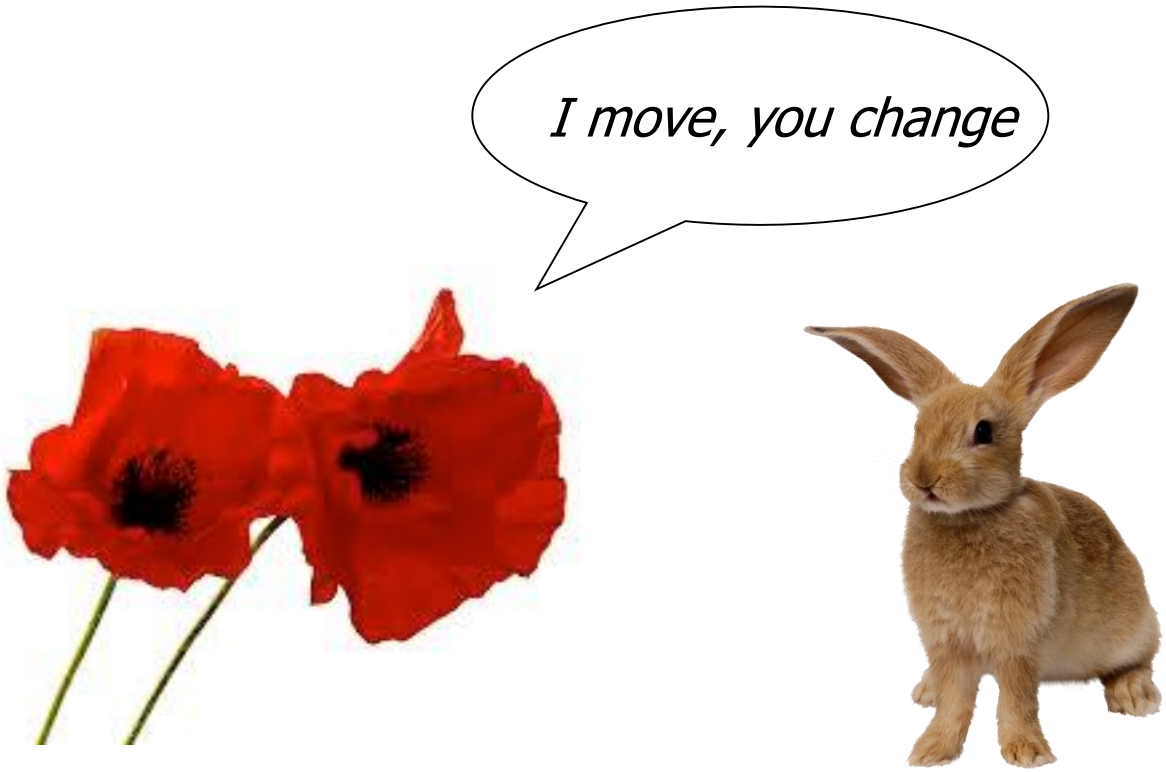




Within-island diversification
Niche evolution

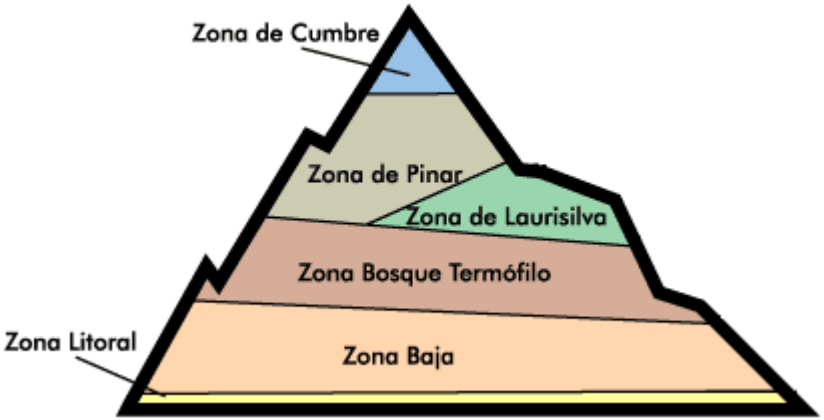


Inter-island colonization of similar biomes (niche conservatism)


I move, you change

Ecological zones

- Coastal belt
- Open habitat
- Thermophilous forest
- Laurisilva
- Pine forest
- Sub-alpine



Ten island-habitat types

M1	*Other Mainland*
E2	*Eastern-Open*
C2	*Central-Open*
W2	*Western Open*
C3	*Central-laurel forest*
W3	*Western-laurel forest*
C4	*Central-pine forest*
W4	*Western-pine forest*
C5	*Central-alpine vegetation*
W5	*Western-alpine vegetation*

L. Martinez 2010

Separating island-hopping and niche-shift rates

r { r_i Shift between islands; r_e Shift between niches; $r_i r_e$ Shift between islands and niches

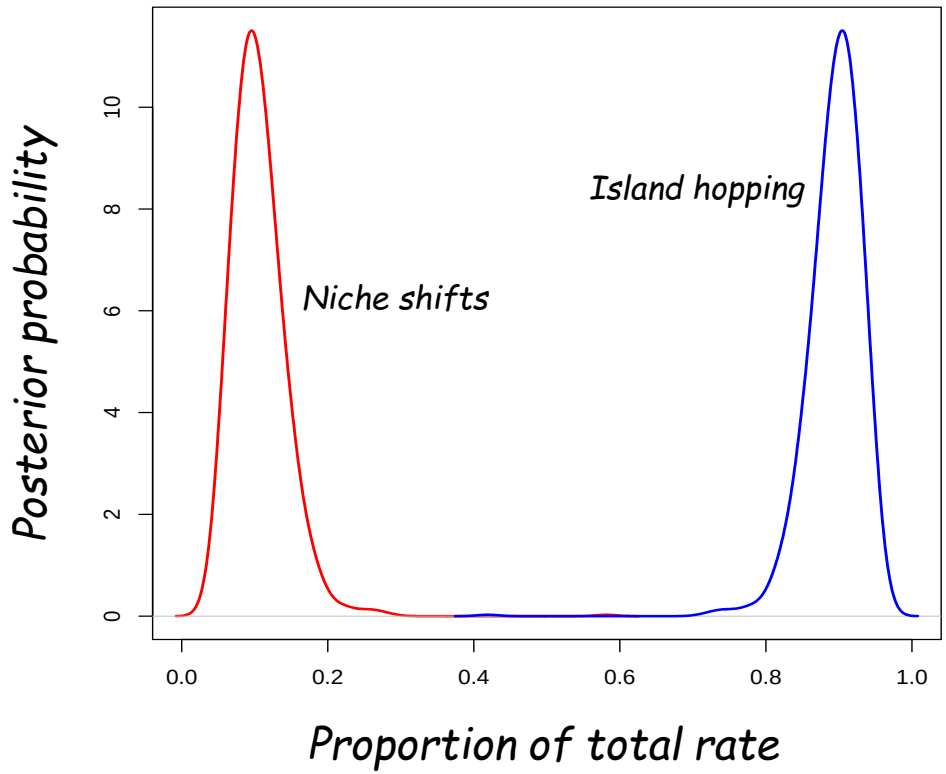
Standard biogeography model:

r ~ dirichlet(1, 1, 1, ...)

Islands-ecology model:

mu ~ dirichlet(1, 1)

*r := simplex(mu[1], mu[2], mu[1] * mu[2], ...)*


Posterior probability
0
2
4
6
8
10
Niche shifts
Island hopping
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of total rate

Separating area and ecology contributions to carrying capacity

a, e	Area and ecology components
$\beta a + (1-\beta)e$	Linear mix
$a^{\beta} e^{(1-\beta)}$	Power mix

$S = c A^{z}$ $z = 0.25$ (0.22, 0.28)

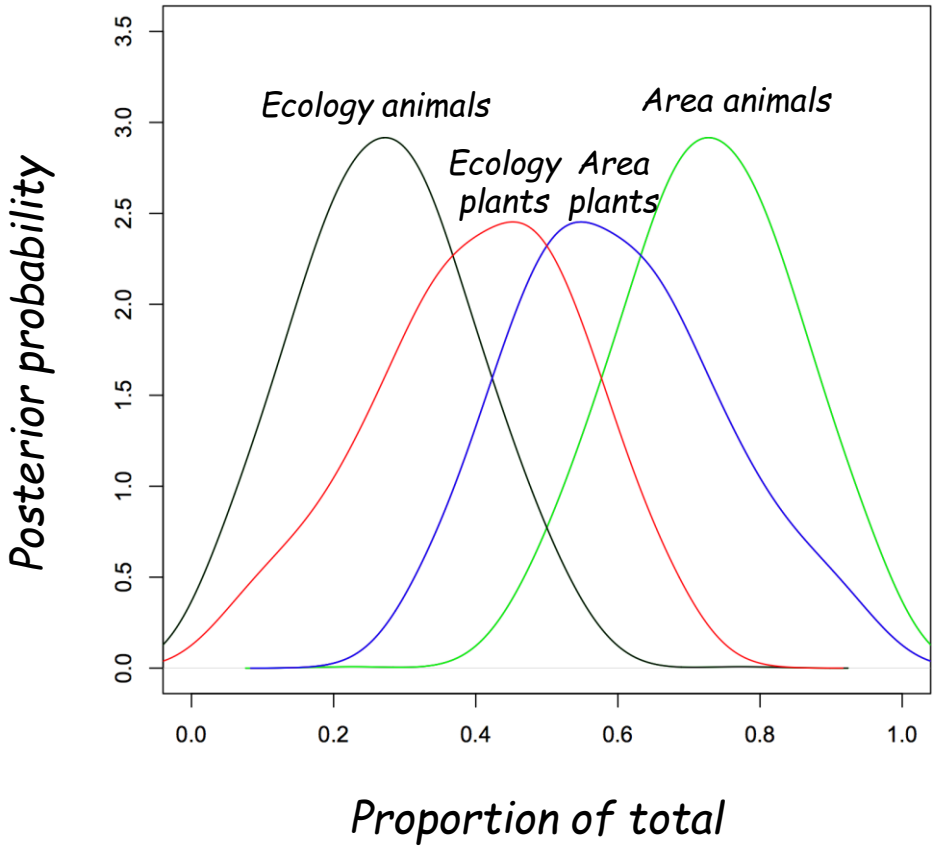
Standard model:

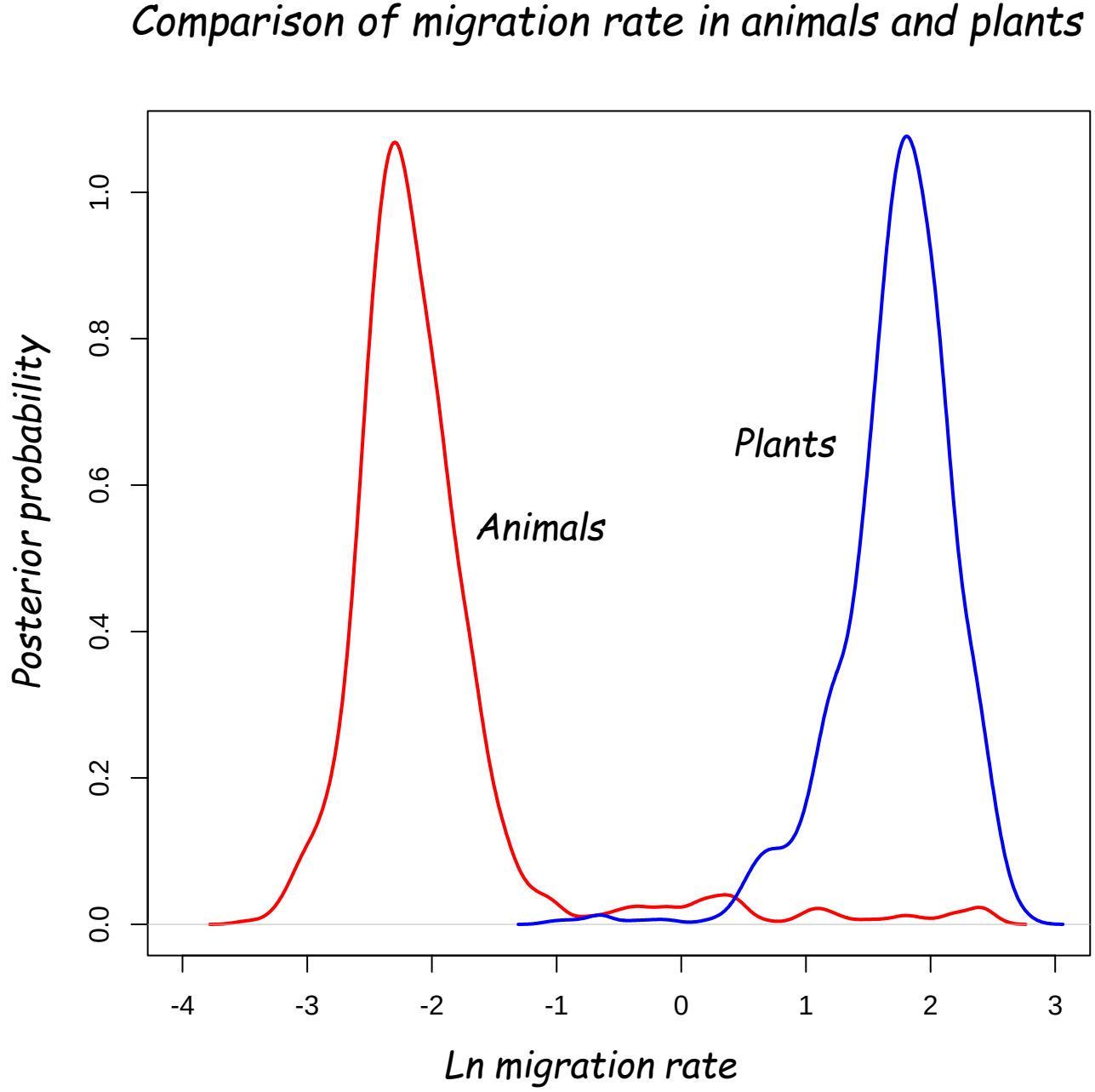
```
pi ~ dirichlet( 1, 1, 1, ...)
```

Power-mix model:

```
areaK <- simplex( A_1 ^ z, A_2 ^z, ...)
ecoK ~ dirichlet( 1, 1, ...)
prop ~ dirichlet( 1, 1 )
pi := powermix( areaK, ecoK, prop )
```

Carrying capacity factored into area and other factors (ecology etc)




Comparison of migration rate in animals and plants
Posterior probability
0.0
0.2
0.4
0.6
0.8
1.0
Animals
Plants
-4
-3
-2
-1
0
1
2
3
Ln migration rate

Summary

- Carrying capacities and relative magnitudes of biotic exchange between islands are similar in animals and plants
- Nevertheless, animal dispersal between islands is an order of magnitude slower than plant dispersal
- Area effects are more important, relatively speaking, in determining island carrying capacities in animals
- Plants shift between islands much more readily than they adapt to new niches
- Phylogenetic graphical models are good tools in modeling and analyzing these phenomena…