BIG4: Biosystematics, informatics and genomics of the big 4 insect groups- training tomorrow's researchers and entrepreneurs

Kick-Off Meeting

14-18 September 2015

Copenhagen, Denmark

BIG4
Biosystematics, Informatics, Genomics

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642241

Statistical Phylogenetics

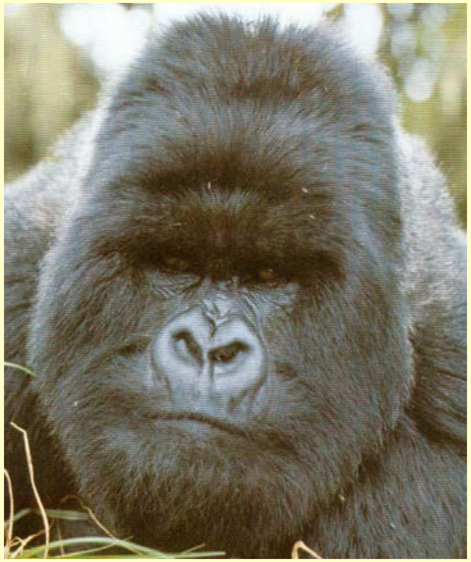
Fred(rik) Ronquist

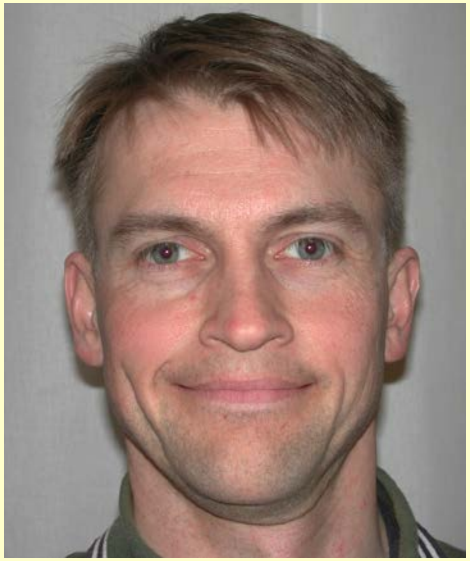
Swedish Museum of Natural History, Stockholm, Sweden

Statistical Phylogenetics

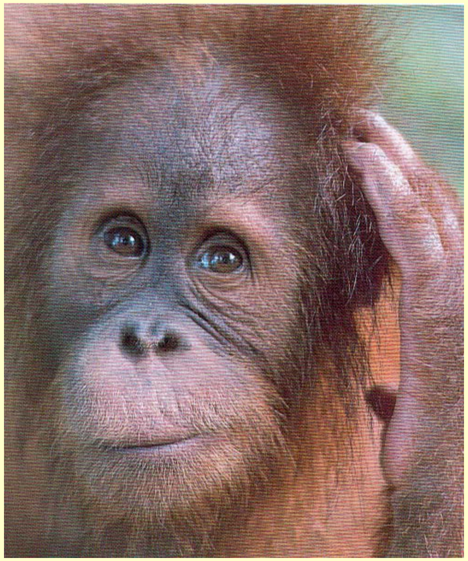
- Despite the computational complexity, statistical approaches are becoming increasingly important in phylogenetics:
  - Difficult problems requiring accurate and unbiased inference (e.g., structure of rapid radiations)
  - More aspects of molecular evolution being examined (structural dependencies, positive selection etc)
  - Combination of background knowledge and sequence information (e.g., divergence time estimation)
  - Rich evolutionary models (e.g., biogeography)

Infer relationships among three species:

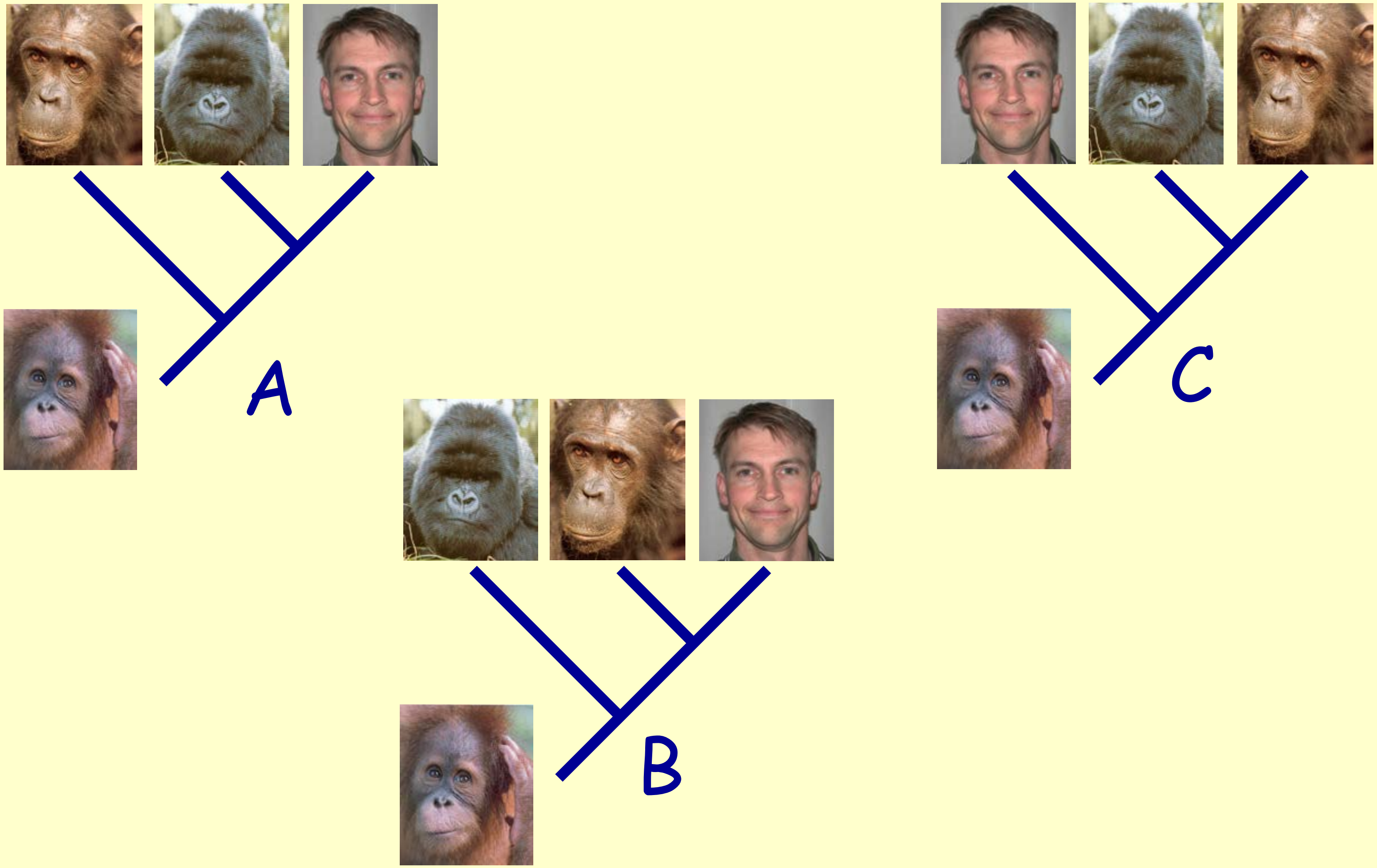




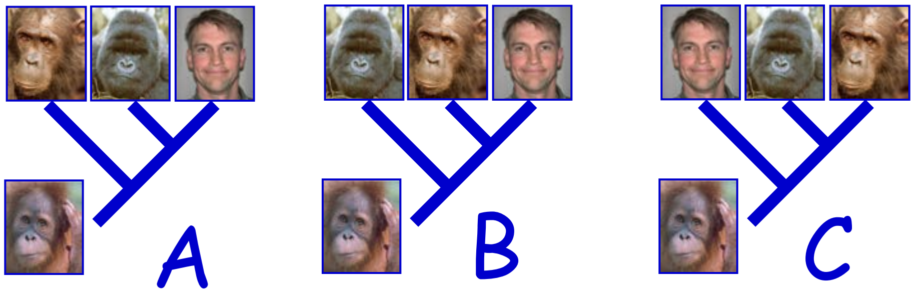
Outgroup:



Three possible trees (topologies):

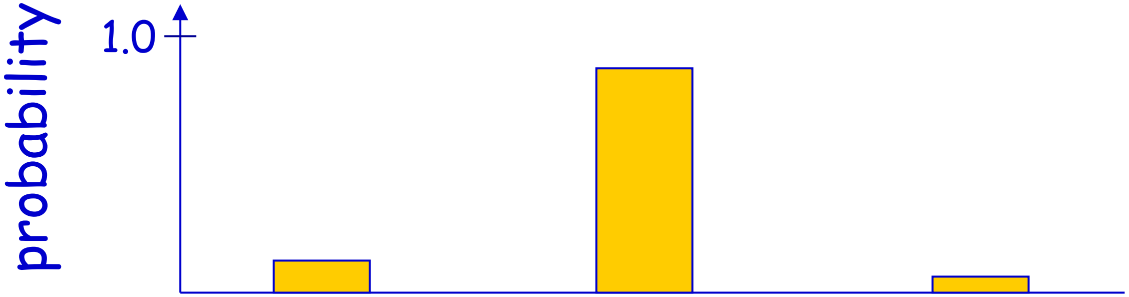


Maximum likelihood inference



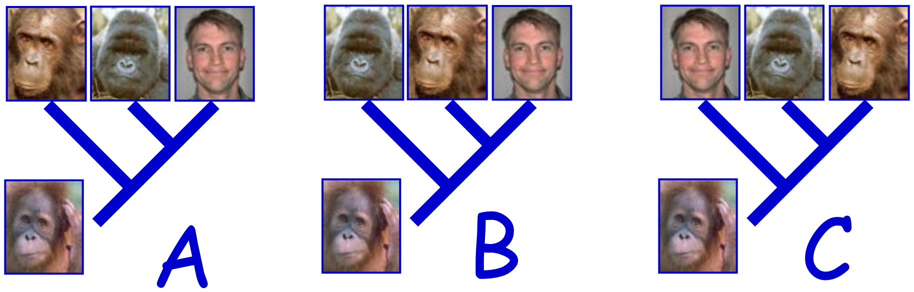
Model

Data (observations)

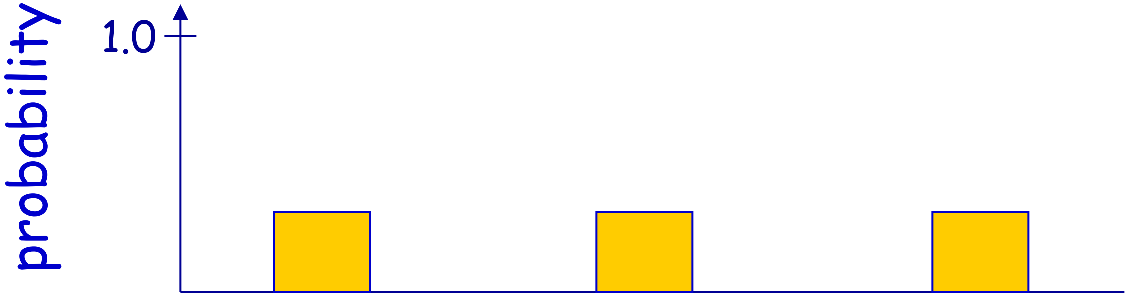


Likelihood function

Bayesian inference

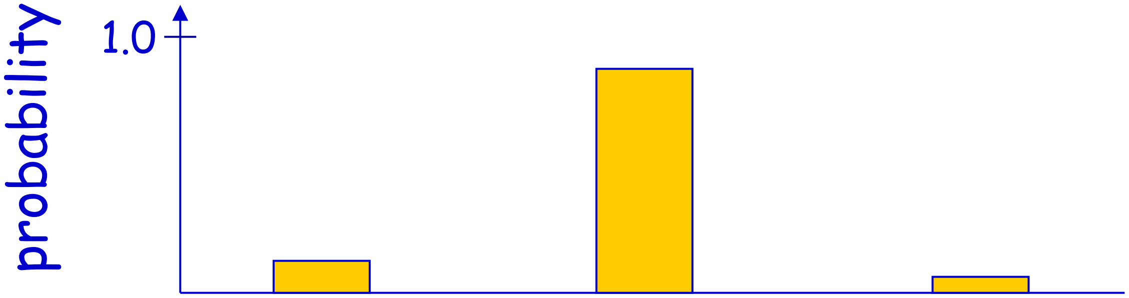


Model



Prior distribution

Data (observations)



Posterior distribution

Bayesian or ML?

- Maximum likelihood
  - Can be fast
  - Background knowledge ignored
  - No natural way of measuring uncertainty
  - Difficult to extend to complex models
  - Assessing quality of results from ML algorithms difficult
- Bayesian inference
  - May be slower
  - Possible to incorporate background knowledge
  - Natural measure of uncertainty (posterior probability distribution)
  - Standard computational machinery (Markov chain Monte Carlo), which can easily be extended to complex models
  - Convergence diagnostics for MCMC well developed

"My" Bayesian Software

- MrBayes
  - Large model space (but ~ fixed)
  - Robust and reliable
- RevBayes
  - Flexible model specification using graphical models concepts
  - Clunky but pretty competent
- Rev
  - Completely flexible graphical model specification
  - Programmable
  - Work in progress…

Software challenges

- Modeling explosion, especially in the Bayesian context
- Challenging for empiricists to communicate and correctly understand models
- Challenging for developers of inference software to cope with expanding model universe
- Addressed using switches and the like
- Can we develop more generic computational machinery?

© 1996 Randy Glasbergen.

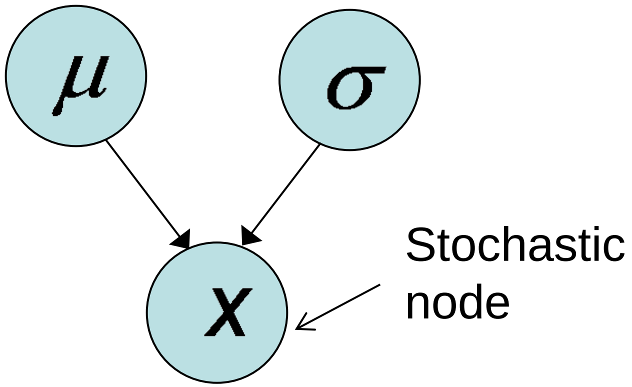
Probabilistic Graphical Models

- 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:

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

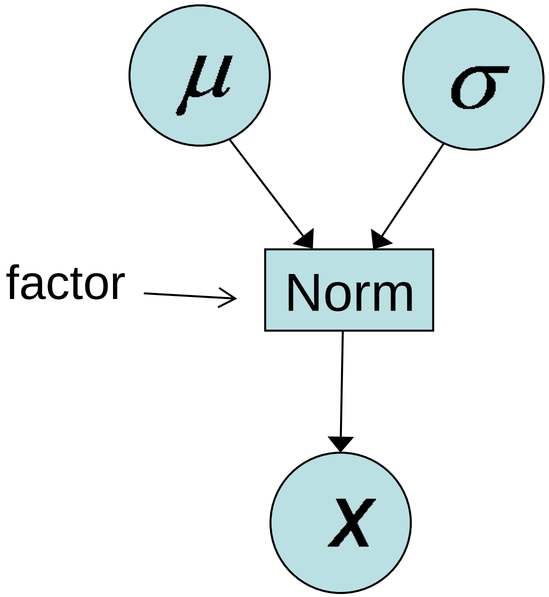
- Extensive literature on generic algorithms that apply to model graphs

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

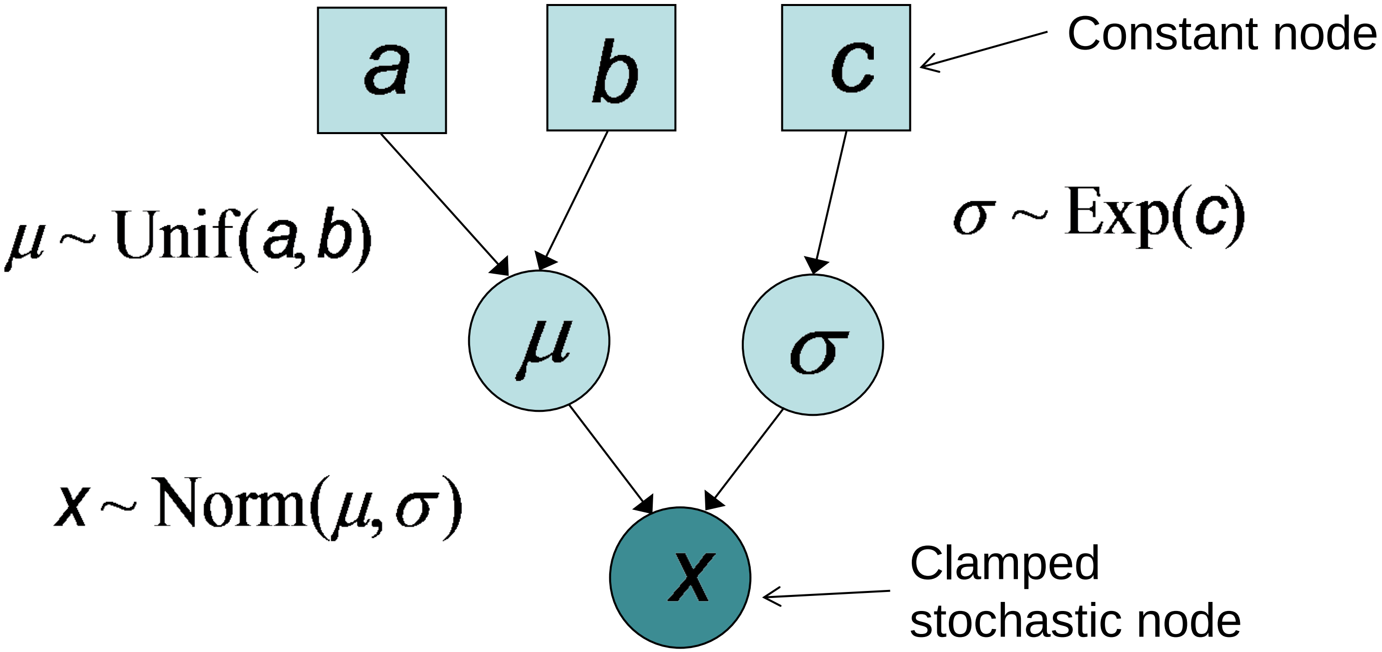


Graphical Model

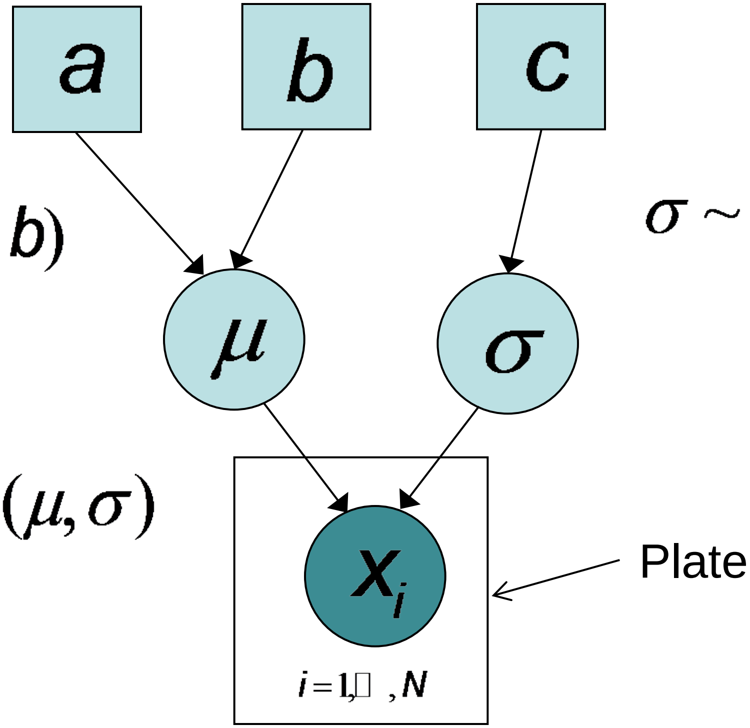
Compact Form



Factor Graph


a
b
c
Constant node
$\mu \sim \mathrm{Unif}(a,b)$
$\sigma \sim \mathrm{Exp}(c)$
μ
σ
$x \sim \mathrm{Norm}(\mu,\sigma)$
x
Clamped
stochastic node

Hierarchical Graphical Model



a

b

c

μ

σ

x_i

$i = 1, \ldots, N$

Plate

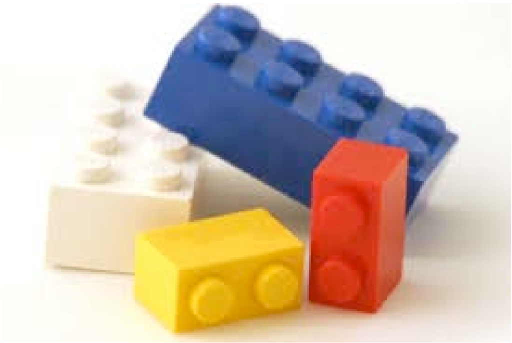
$$\mu \sim \mathrm{Unif}(a, b)$$

$$\sigma \sim \mathrm{Exp}(c)$$

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

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



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 a new phylogenetic model

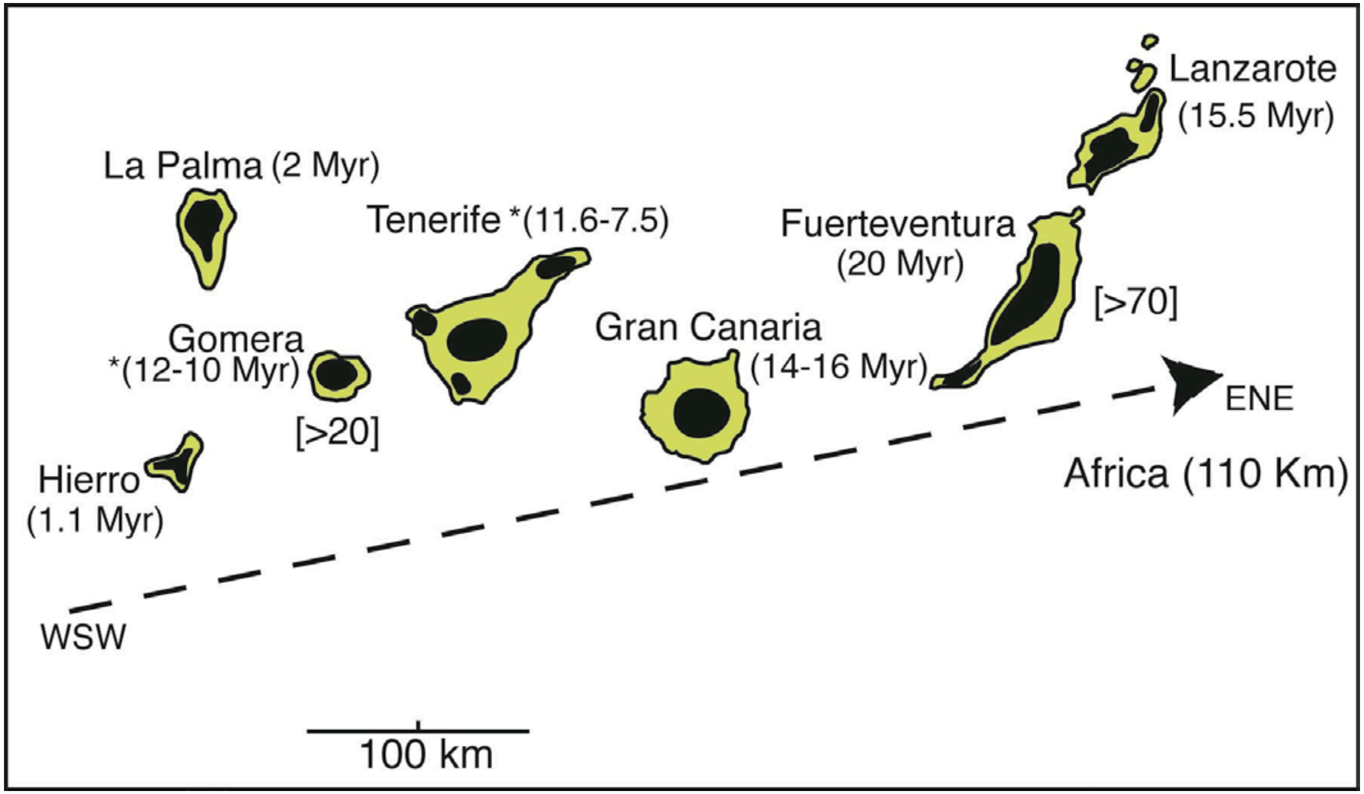
```
# definition of the myGTR function (“Ziheng's favorite”)
function model myGTR (CharacterMatrix data) {

   # 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 );
}
```

Appr. 20 lines

Complexity hidden from normal user

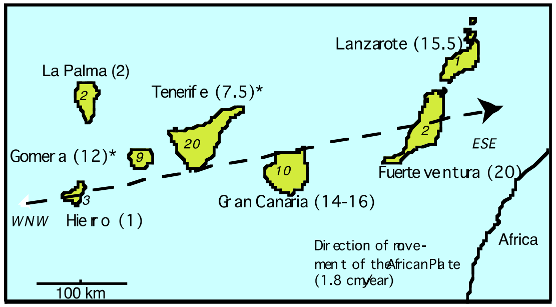
```
# Read in data
myData <- read( “data.nex” )

# Apply model
myModel = zihengGTR( myData )

# Construct mcmc
myMCMC = mcmc( myModel )

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

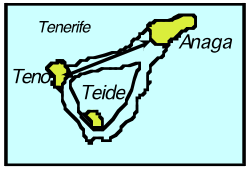
The Canary Islands



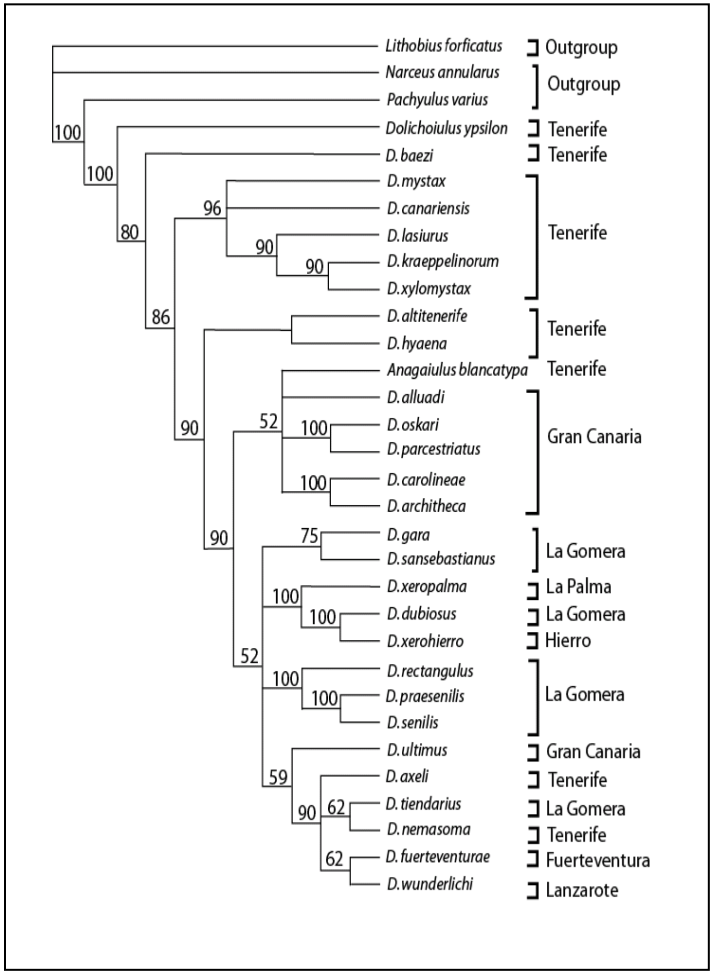
Dolichoiulus (Diplopoda)

***Dolichoiulus* (Diplopoda, Julida, Julidae, Pachyulinae)**

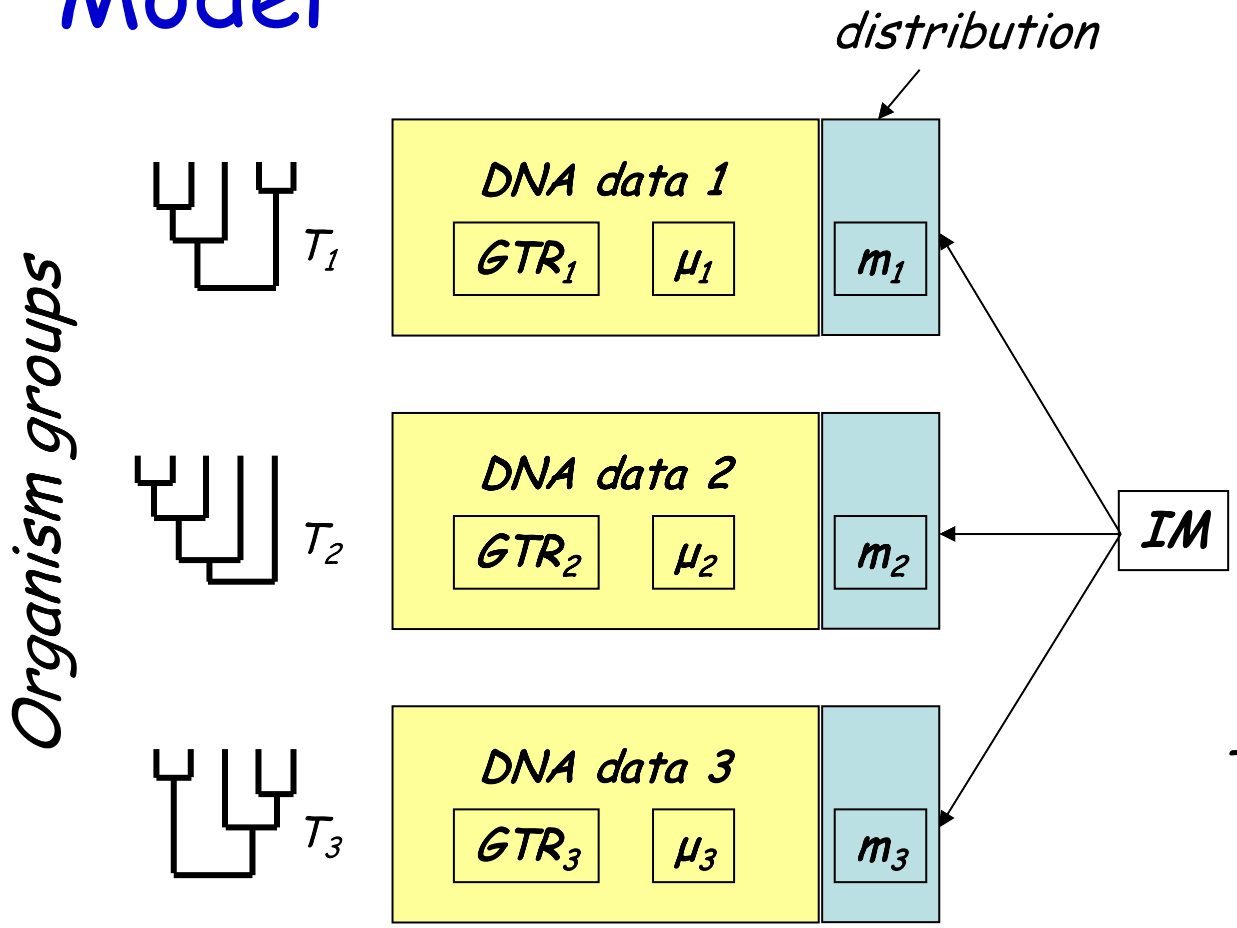




46 endemic species



Model

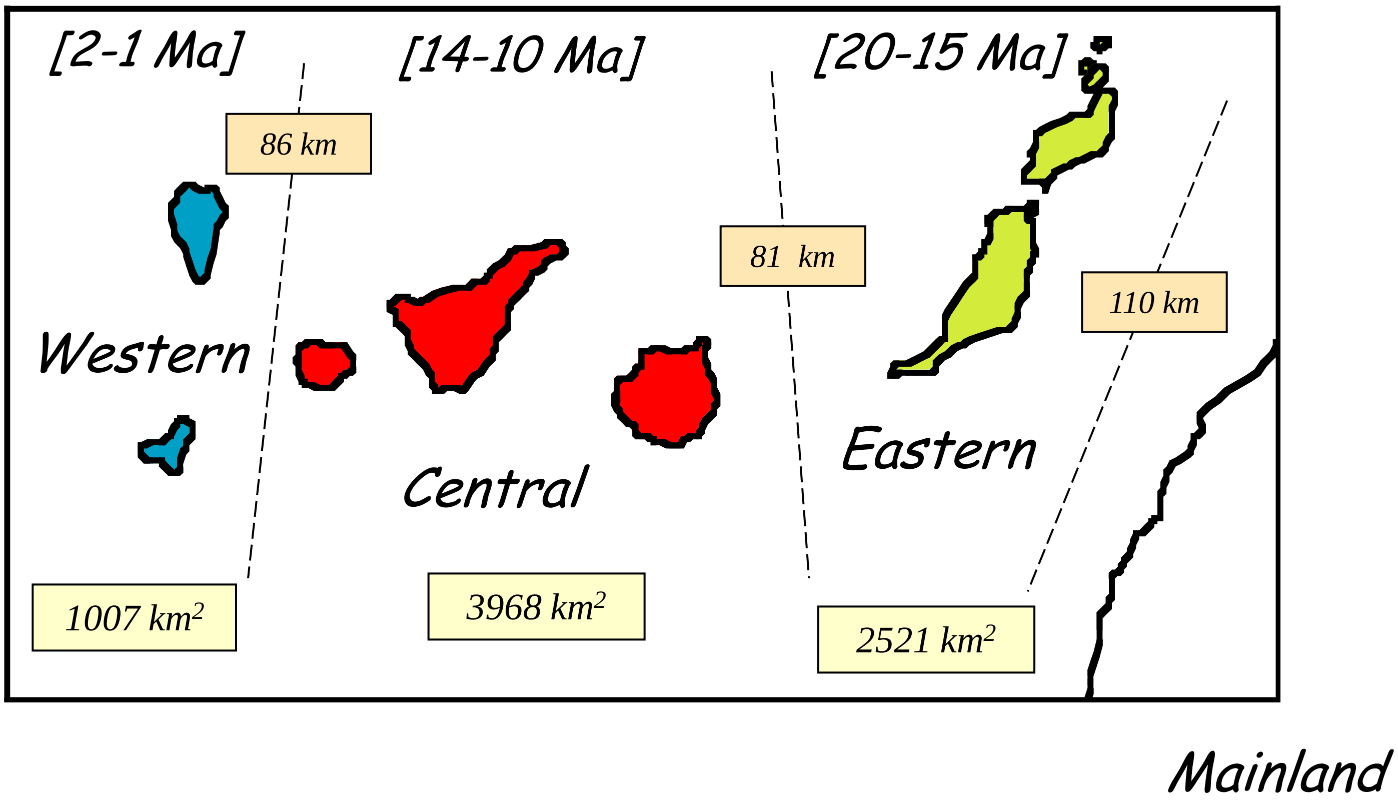


IM – island model
μ_i – mutation rate
m_i - dispersal rate

Inference

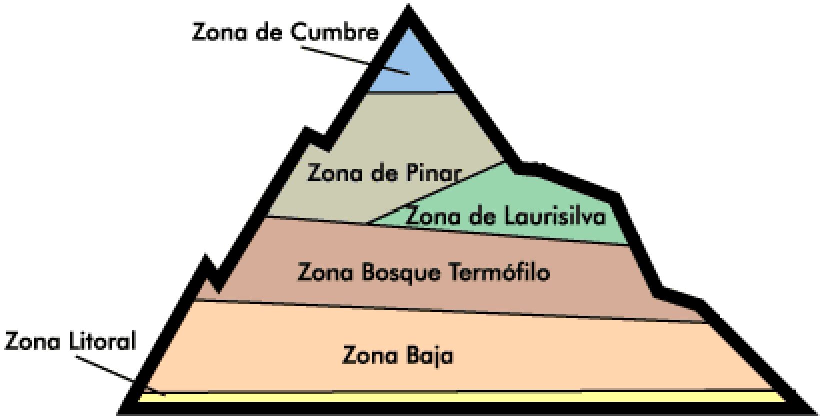
Bayesian inference using MCMC sampling, accommodating uncertainty in all model parameters

Canary Islands: 3-island model



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

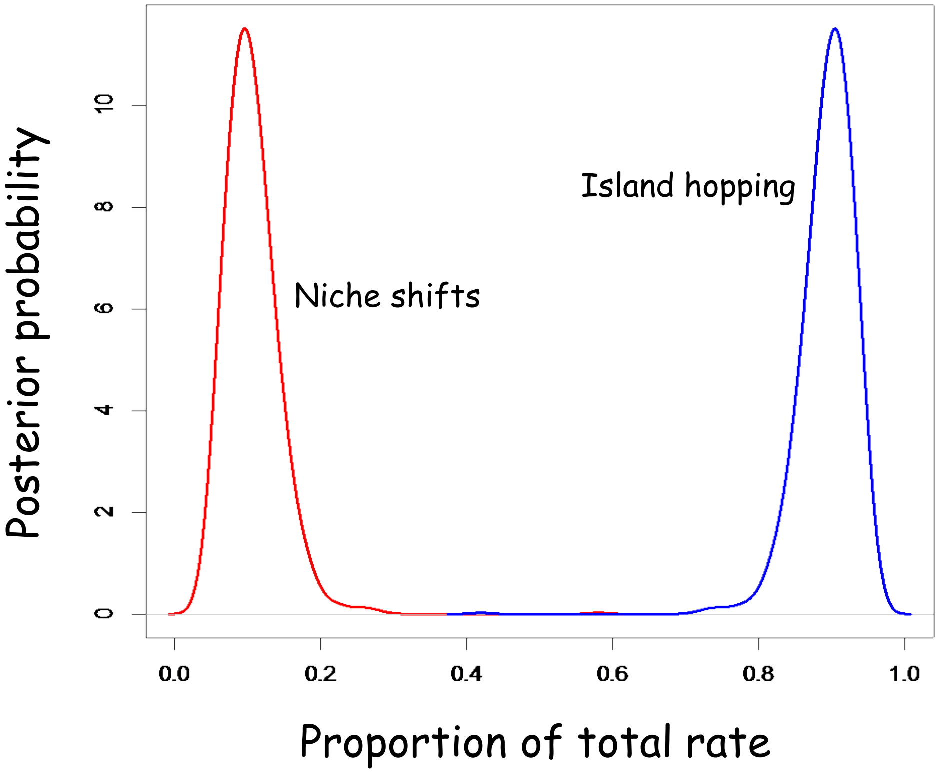
$$
r \left\{ \begin{array}{ll} r_i & \text{Shift between islands} \\ r_e & \text{Shift between niches} \\ r_i r_e & \text{Shift between islands and niches} \end{array} \right.
$$

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