From Individual-based Population Models to Lineage-based Models of Phylogenies

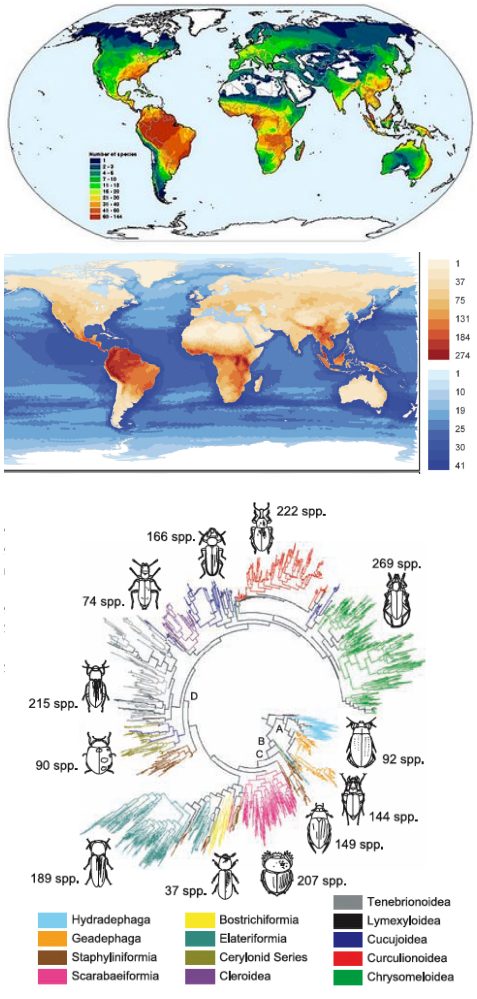
Amaury Lambert
(joint works with G. Achaz, H.K. Alexander, R.S. Etienne, N. Lartillot, H. Morlon, T.L. Parsons, T. Stadler)

New Directions in Probabilistic Models of Evolution
U.C. Berkeley, Simons Institute, April 30th 2014

SMILE : an interdisciplinary group in Paris

- SMILE = Stochastic Models for the Inference of Life Evolution
- CIRB = Center for Interdisciplinary Research in Biology (Collège de France)

Modeling, and inferring from, phylogenies



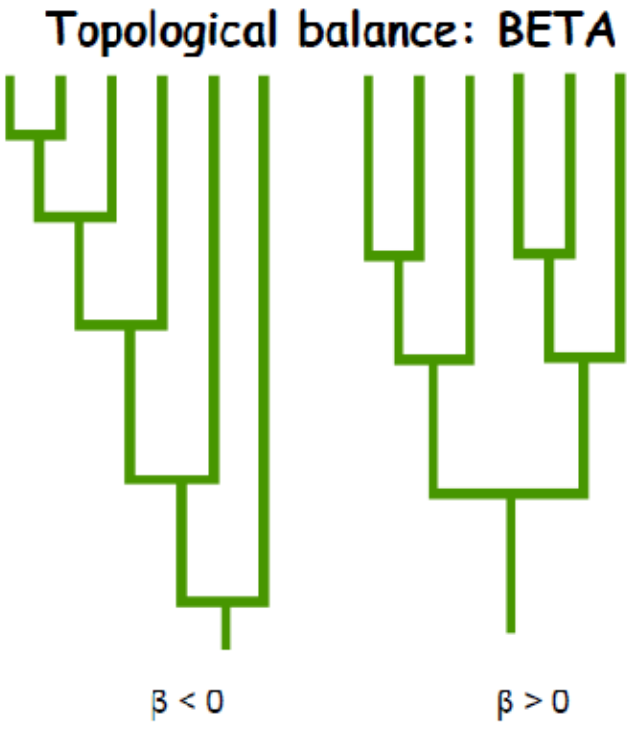
In our group, we try to

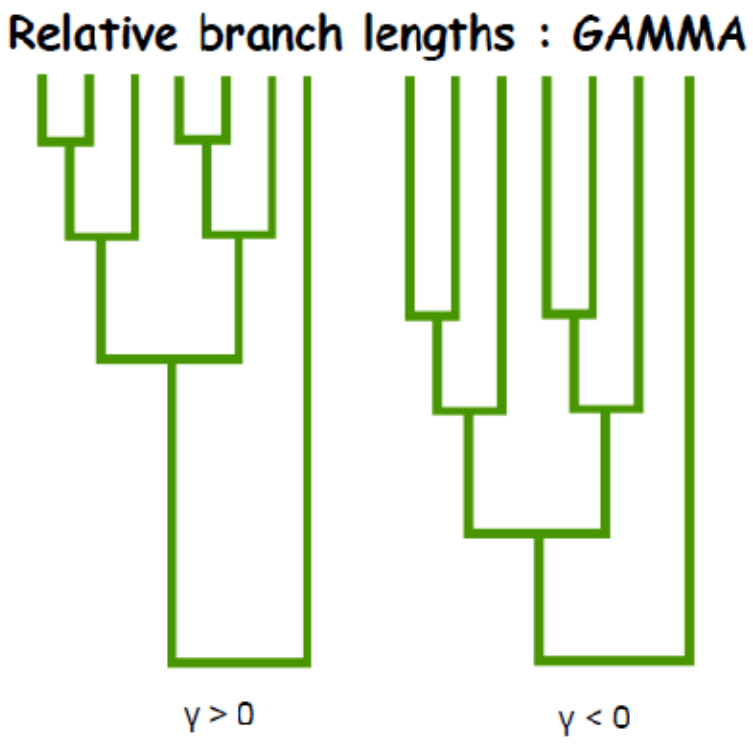
- Design probabilistic models of evolutionary processes...
- ...generating similar patterns as those observed in nature, or/and...
- ...allowing for inference of these processes from real data.

In this talk, my goal is to take time-calibrated phylogenies as the raw data,

- Propose models of speciation producing phylogenetic trees...
- ... sharing common features with real phylogenies, or/and...
- ...whose likelihood can be computed.

2 examples of observable statistics





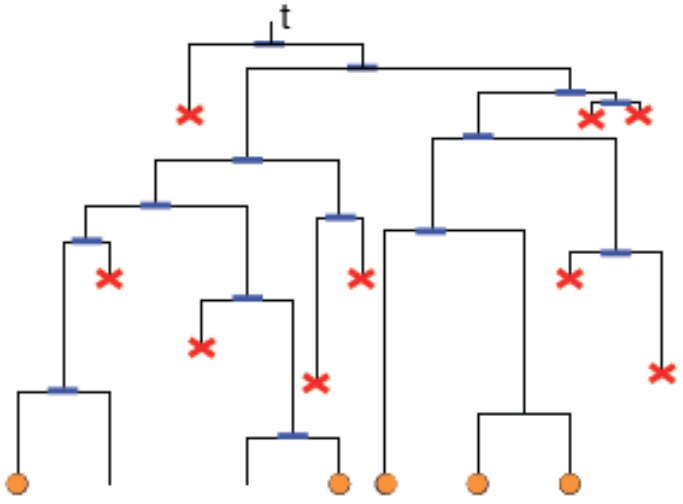
- MLE of Beta-splitting (Aldous 1996)
- Yule tree (pure birth) : $\beta = 0$
- Real trees are imbalanced : $\beta < 0$ (Blum & François 2006)

- Yule tree (pure birth) : $\gamma = 0$
- Kingman coalescent has nodes closer to tips : $\gamma > 0$
- Real trees have nodes closer to the root : $\gamma < 0$ (McPeek 2008)

Outline

Birth–death models of genealogies/phylogenies

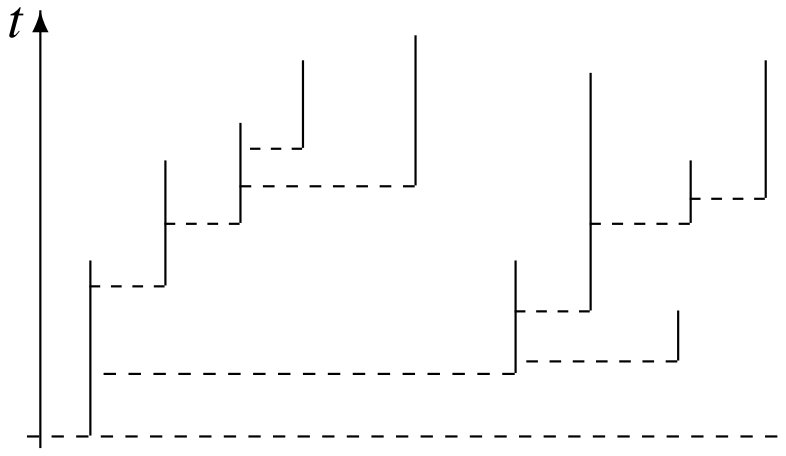
- Lineage-based model = birth–death model
- Where particles can be individuals or species (Nee et al *PNAS* 1992)
- Particles split into two new particles at rate b = birth (or speciation) rate
- Particles die at rate d = death (or extinction) rate
- Particles may bear some trait (evolving as branching Markov)



complete phylogeny

Assumptions on rates

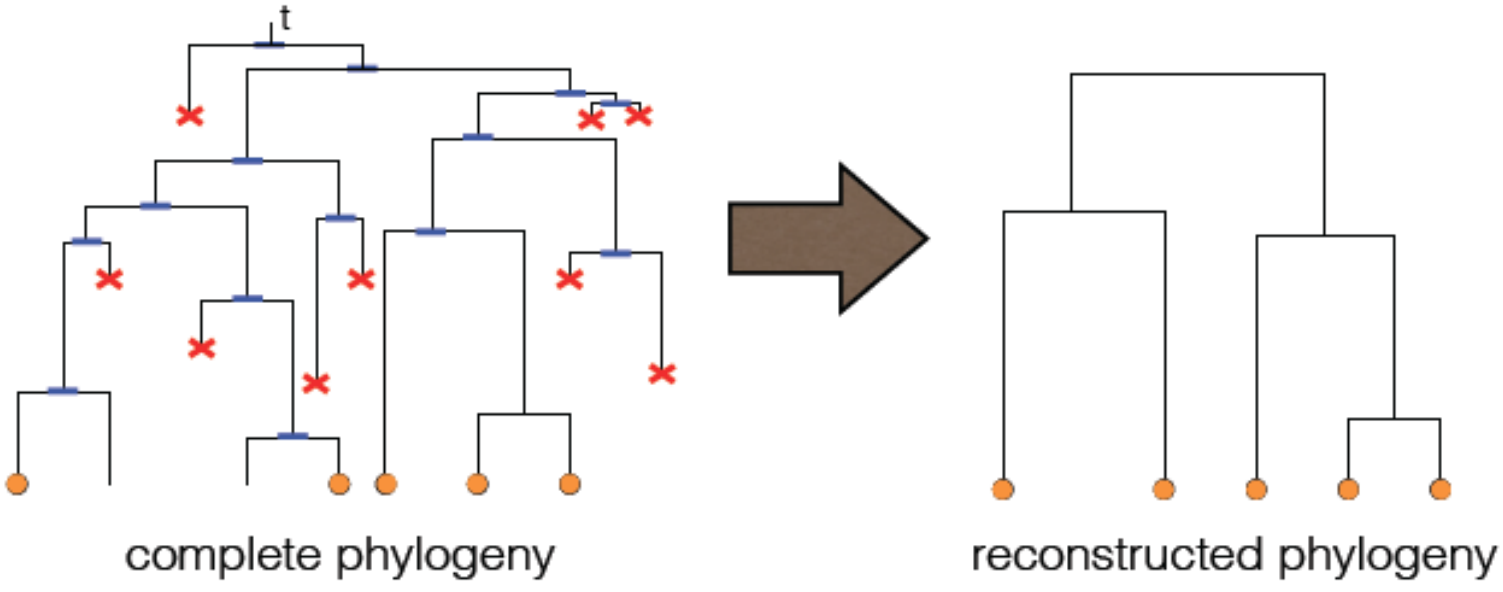
Rates $b(t,n,a,i)$ and $d(t,n,a,i)$ may depend upon :



- **time** t
- **number** n of standing particles
- a **non-heritable trait** a (e.g., age)
- a **heritable trait** i
- Asymmetric birth =
  Mother keeps her trait
- Orientation =
  Daughter sprouts to the right

Reconstructed tree

Reconstructed tree = remove all lineages extinct by T (fixed time).



Characterizing lineage-based models

Proposition (L. & Stadler 2013)

Under these (lineage-based) models of diversification,

1. *Reconstructed trees always have the same topology as Yule trees IFF* $b = b(t,n)$ *and* $d = d(t,n,a)$

   $\Longrightarrow$ *As soon as* $b = b(t,n)$ *and* $d = d(t,n,a)$, *estimate* $\beta \approx 0$, *BUT*

2. *The likelihood of reconstructed trees always has an explicit product form IFF* $b = b(t)$ *and* $d = d(t,a)$.

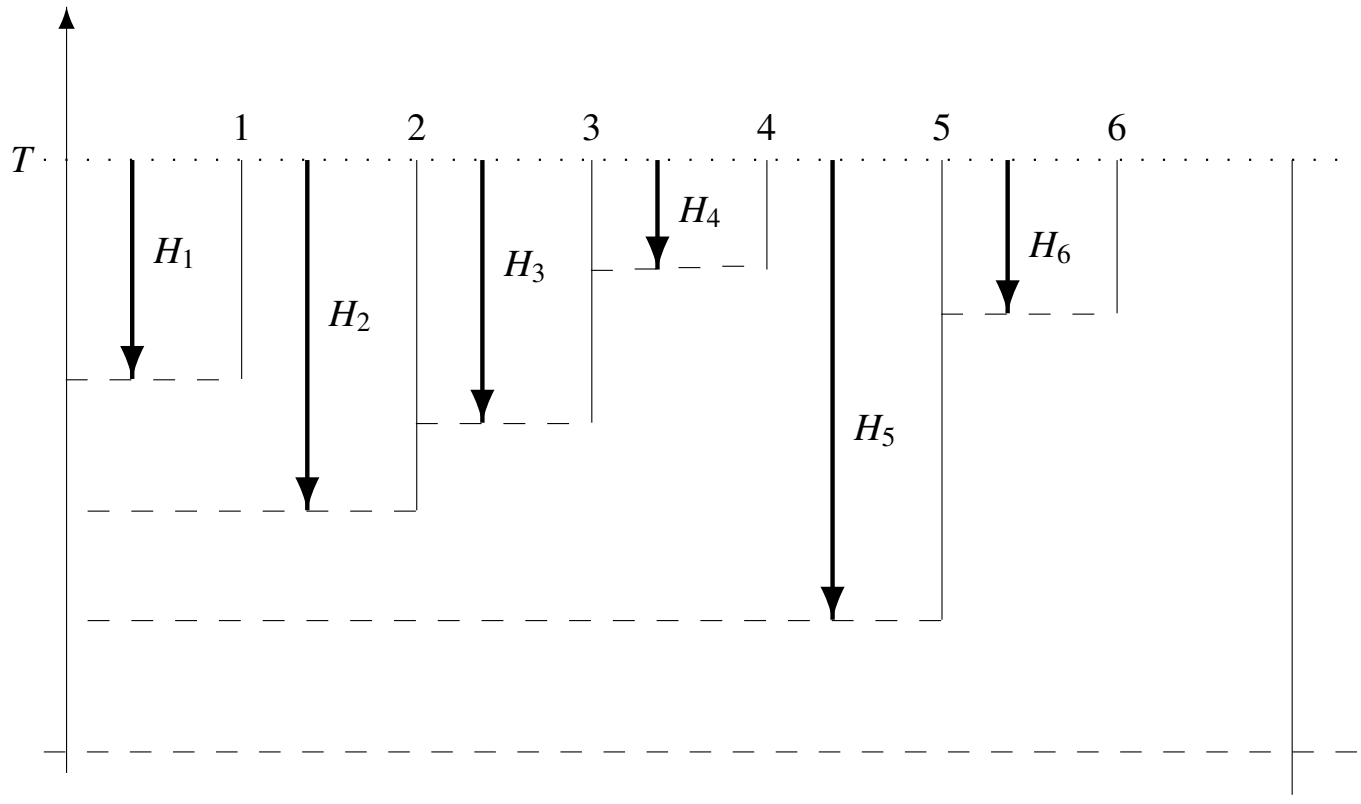
   $\Longrightarrow$ *The reconstructed tree is called a coalescent point process...*

Outline

The CPP distribution (Popovic 2004, Aldous & Popovic 2005)

A reference distribution on ultrametric, oriented trees with edge lengths

CPP = Coalescent Point Process = Oriented tree whose node depths $H_1, H_2, \ldots,$ form a sequence of **iid random variables** killed at its first value larger than T.



$b = b(t)$ and $d = d(t,a)$ always produce CPP

Assume that $b = b(t)$ **and** $d = d(t,a)$, where t is time and a is any non-heritable trait.
Set $g(t,s)$ the density at time s of the extinction time of a species born at time t.

Theorem (L. & Stadler 2013)

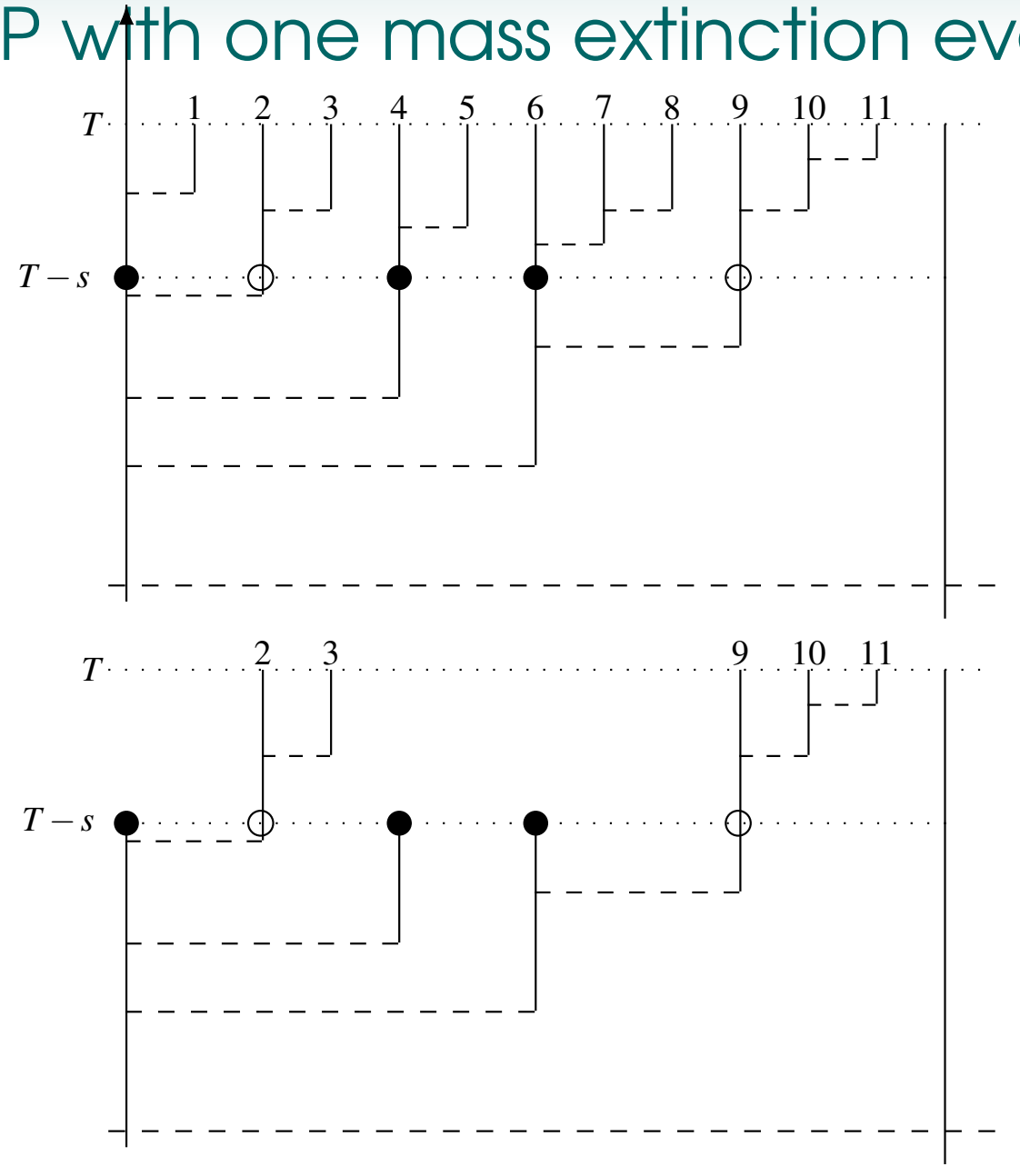
*The **reconstructed (oriented) tree is a CPP** with typical node depth H, where the function $F = 1/P(H > \cdot)$ is the unique solution to the following linear integro-differential equation*

$$F'(t) = b(t)\left(F(t) - \int_{T-t}^{T} ds\, F(s)\, g(t,s)\right) \qquad t \geq 0,$$

with initial condition $F(0) = 1$.

The result still holds with mass extinction events/missing species.

CPP with one mass extinction event



Age-dependent extinction in the bird phylogeny

With T. Stadler and H.K. Alexander

- Gamma distributed lifetime ($k, s > 0$), with mean $m := ks$

$$g(a) = \Gamma(k)^{-1} s^{-k} a^{k-1} e^{-a/s}$$

- Exponential distribution is $k = 1$: age-independent ext rate
- Test on simulations : accurate ML estimates of b and m
- MLE on *Aves* phylogeny = 9993 extant bird species (Jetz et al *Nature* 2012)
- Exponential model rejected ($p = 10^{-15}$)
- Shape parameter $k \gg 1$: extinction rate increases with age
- Average lifetime $m = 15.26\ My$
- Speciation rate $b = 0.108\ My^{-1}$

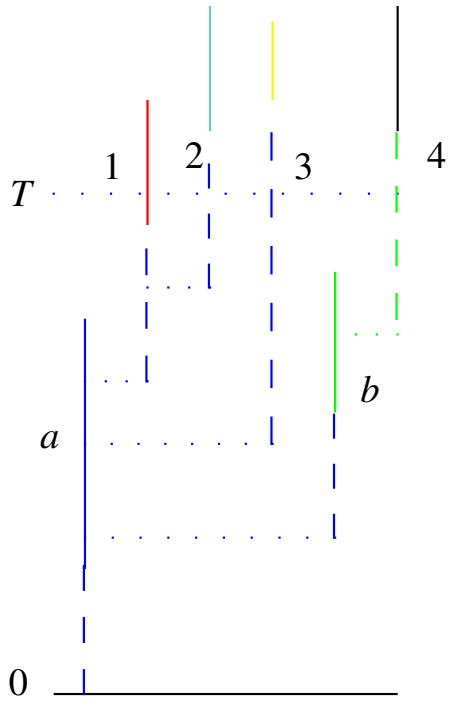
Outline

Protracted speciation (Rosindell et al 2010, Etienne & Rosindell 2012)

With R.S. Etienne and H. Morlon

- Particles = Populations
- **Speciation stage** = non-heritable trait = Each population gradually diverges from mother species
  - Newborn populations are **incipient** = same species as mother population
  - Become **good** after some random time = new species
- Each species is represented by a single population

Protracted speciation (2)



- 4 extant populations at time T
- 3 extant species
- Species b is represented by Population 4
- Species a is represented by Population 2.

Protracted speciation (3)

Assume that the birth rate b does not depend on speciation stage.

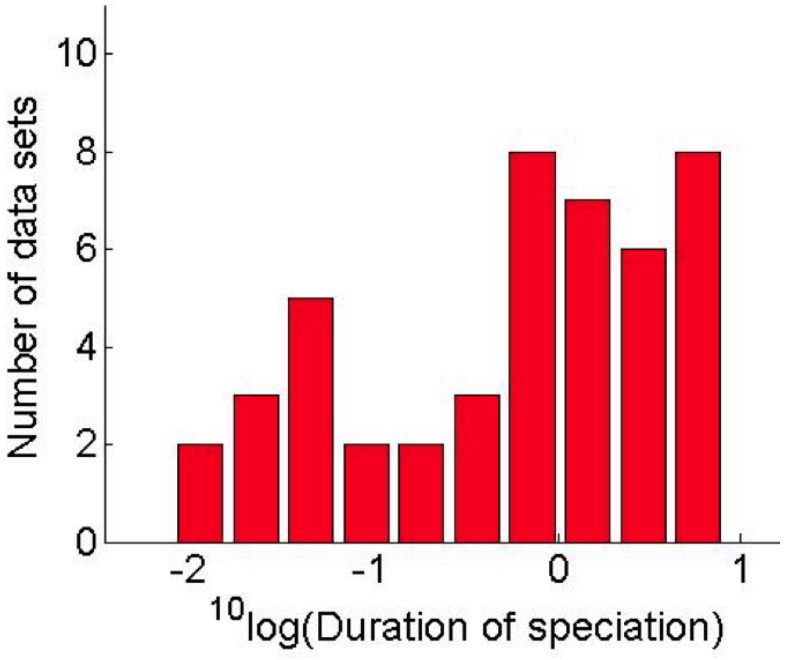
Theorem (Etienne, L. & Morlon 2013)

The reconstructed tree spanned by extant ***representative*** *populations at T is a* ***coalescent point process*** *with node depth H^r, where*

$$P(H^r > t) = \exp\left(-\int_{T-t}^{T} b(s)\,(1-p_1^r(s))\,ds\right)$$

and $p_1^r(t)$ is the probability that a species born at time t does not have any good descending species that has extant descendance at time T.

Protracted speciation (4)



- Test on simulations : poor ML inference for each individual parameter
- Efficient inference of **duration of speciation** = waiting time before **first descending good** population
- Left : duration of speciation inferred in 46 bird clades (in My)

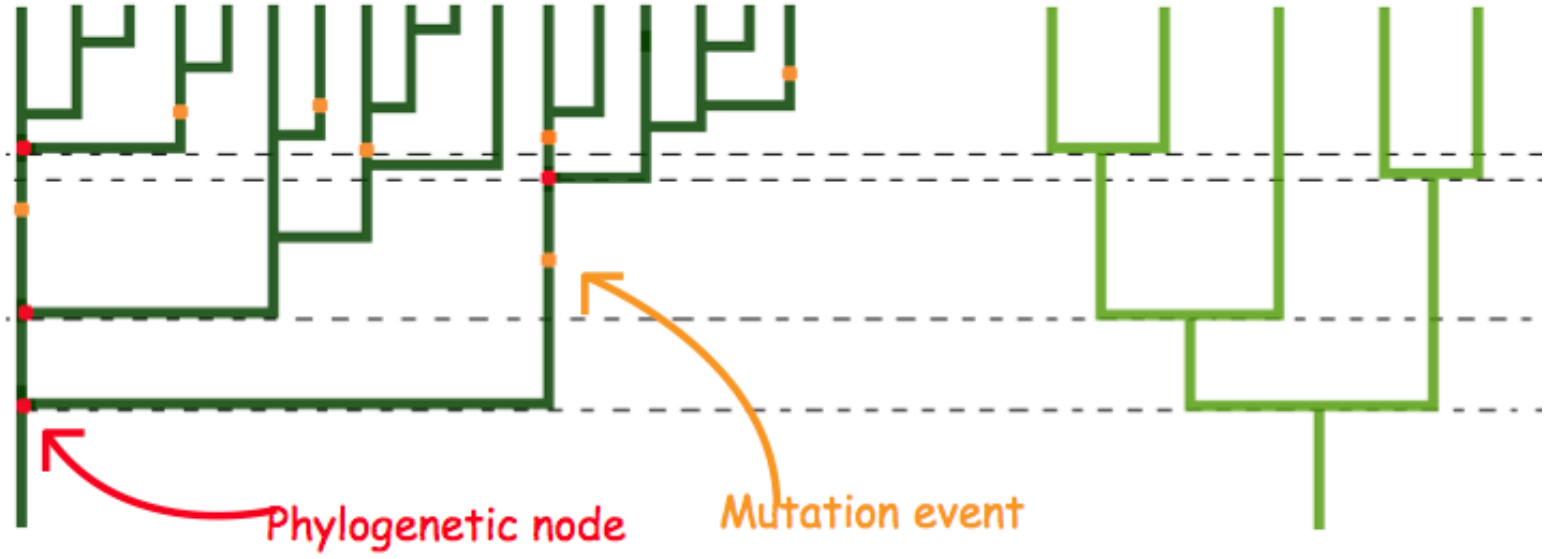
Outline

Speciation by genetic differentiation (1)

Work in progress with M. Manceau and H. Morlon

- Start with a birth–death tree (constant rates b and d, but...)
- Add Poissonian mutations rate θ, infinite-allele model
- **Species = minimal monophyletic taxon** such that any 2 tips with the same allele belong to the same species
- SGD = Speciation by genetic differentiation = individual-based version of protracted speciation

Speciation by genetic differentiation (2)

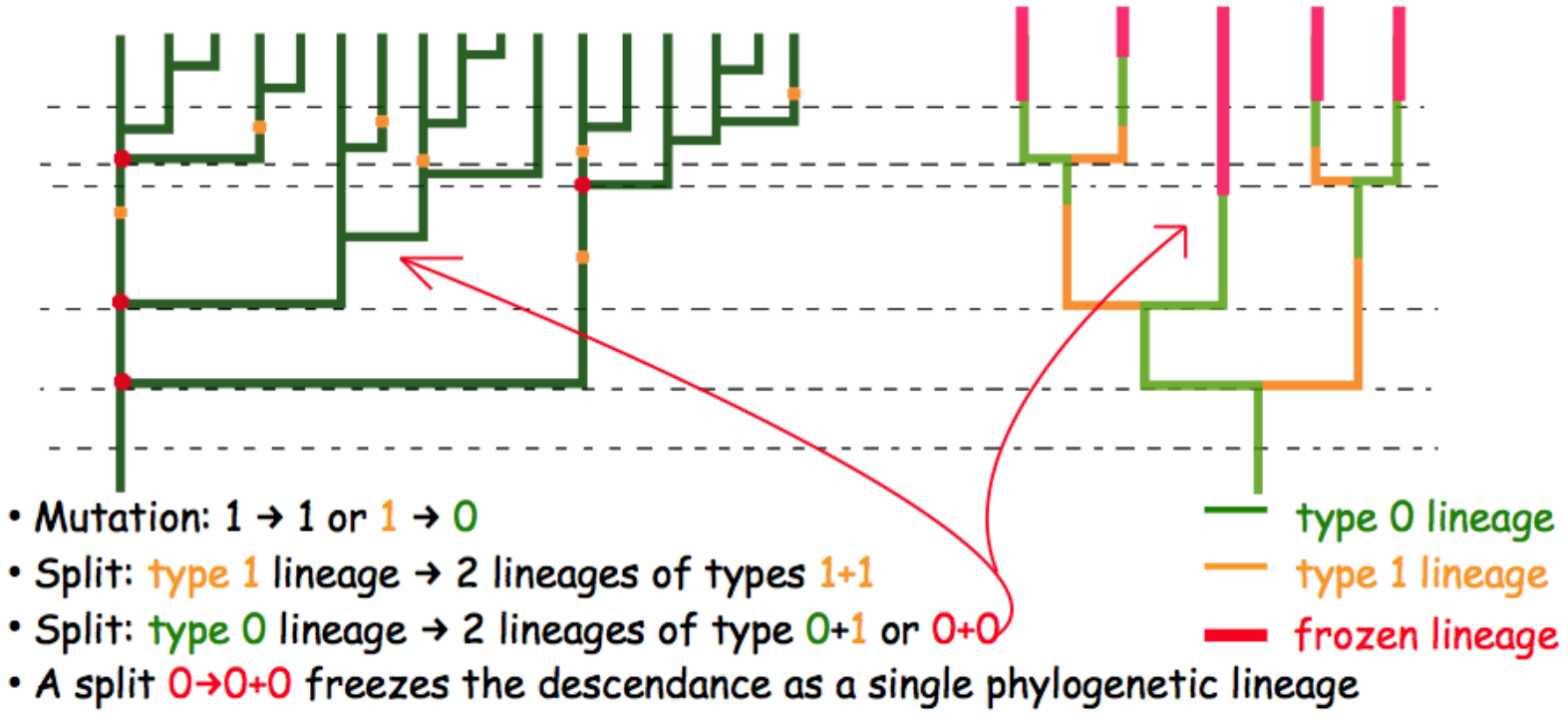


- A node on the genealogy is phylogenetic (= appears on the phylogeny) if

  (i) The previous node is phylogenetic
  (ii) All tips separated by this node carry different alleles

- The first node is phylogenetic if it satisfies (ii)

Speciation by genetic differentiation (3)

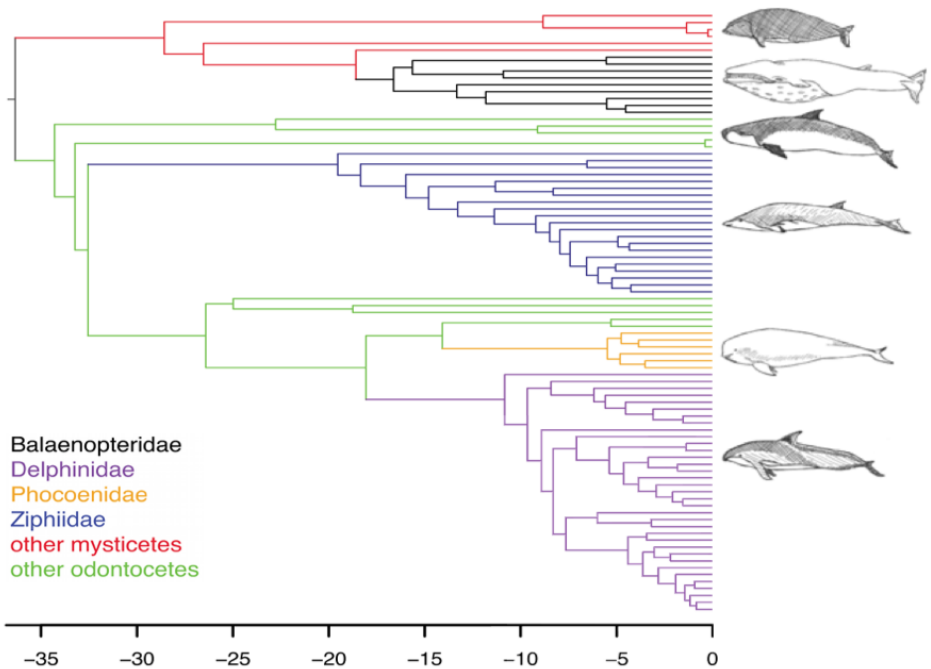


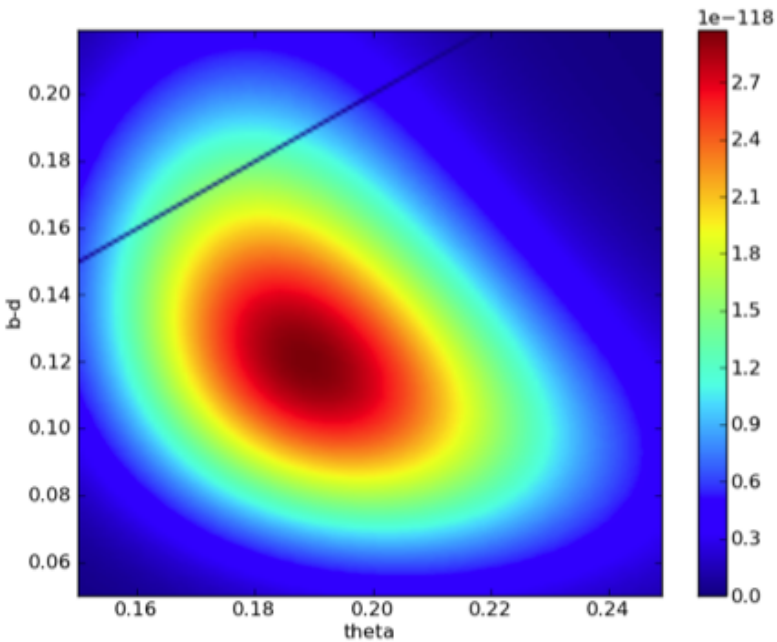
The phylogeny is generated by a 3-type time-inhomogeneous branching process

- a lineage is in state 1 if the allele it is carrying is NOT represented at T
- a lineage is in state 0 if the allele it is carrying is represented at T
- a lineage in state 0 gets frozen into one single phylogenetic lineage when it splits into two 0-lineages

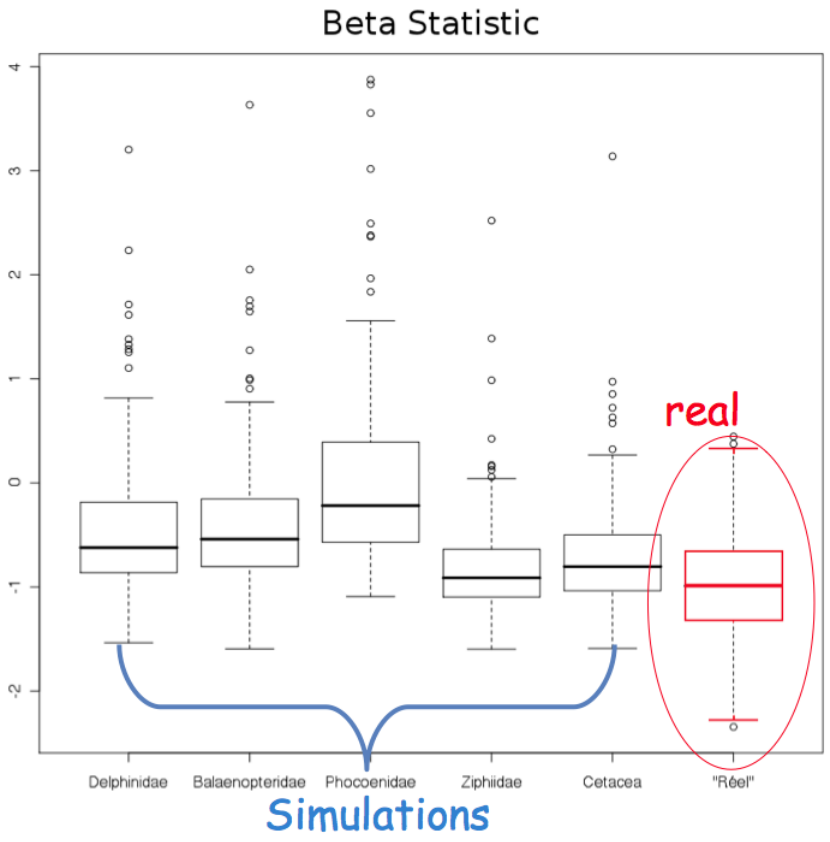
Speciation by genetic differentiation (4)

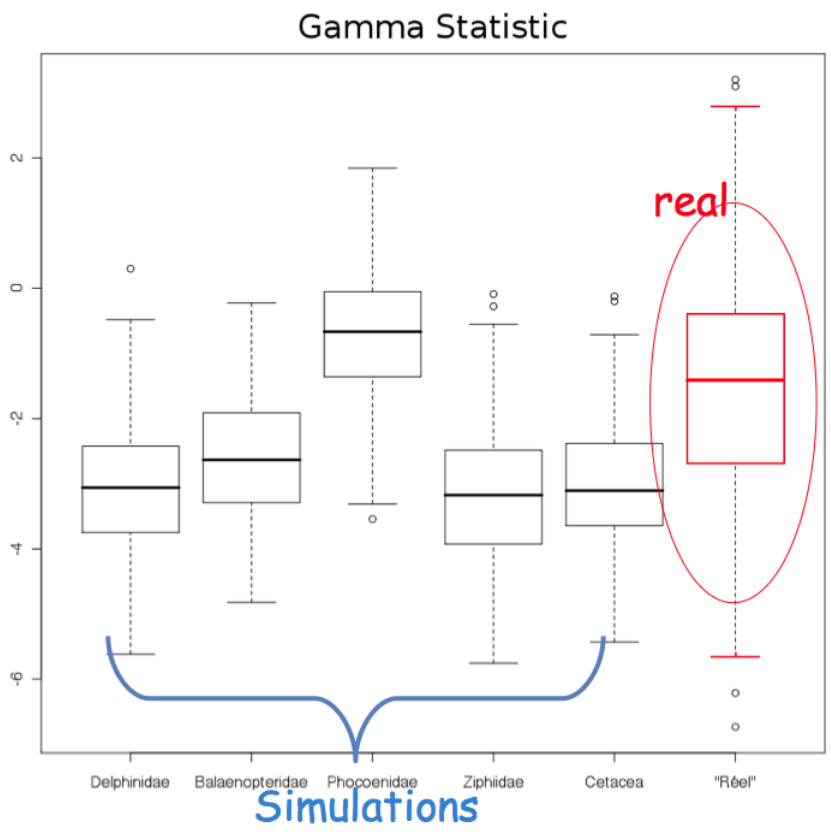
- Branching process representation : fast simulation
- Likelihood computation by peeling algorithm, including the case of missing species
- Tests by simulations : accurate ML estimates of θ and $b-d$
- Inference from Cetaceans (Steeman et al *Syst Biol* 2009) generates realistic values of β, γ





Speciation by genetic differentiation (5)





Outline

Speciation by ecological release

Work in progress with G. Achaz, N. Lartillot, T.L. Parsons

Let $\lambda > \mu > 0$, $c > d > 0$, and $K =$ scaling parameter.

- Start with an individual-based, multitype logistic branching process (Lambert 2005)
- Each ind gives birth
  - at rate λ to an ind belonging to the same species
  - at rate ε_K to an ind belonging to a new species (infinite-allele model)
- Each ind belonging to species i, having abundance X_i, dies at rate

$$\mu + \frac{c(X_i - 1)}{K} + \frac{dX_i'}{K},$$

where $X_i' =$ **total abundance of all YOUNGER species**.

Large population limit

Now assume labels are **levels** :
Species 1 = youngest species,
Species 2 = 2nd youngest species,...

In the absence of mutations, if $K^{-1}X_i(0)$ converge as $K \to \infty$, then $K^{-1}(X_i) \Rightarrow (x_i)$ (Kurtz 1980) where the (x_i) satisfy the system of ODE

$$\dot{x}_i = \left(\lambda - \mu - cx_i - d\sum_{j<i} x_j\right) x_i$$

which, letting $\kappa = \frac{\lambda-\mu}{c}$ and $\alpha = 1 - \frac{d}{c}$ has equilibrium state

$$\lim_{t\to\infty} x_i(t) =: \bar{x}_i = \kappa\alpha^{i-1}.$$

Separation of timescales (Champagnat 2006)

If the mutation rate ε_K is such that

$$e^{-VK} \ll \varepsilon_K \ll \frac{1}{K \ln K}$$

for all $V > 0$, then as $K \to \infty$, subsequent mutants appear

- after the populations have reached their deterministic equilibrium
- before macroscopic departure from this equilibrium.

In the mutation timescale, i.e., when time is accelerated by a factor $1/K\varepsilon_K$,

- $X_i \approx K\bar{x}_i$
- Species i produces a mutant at rate $\varepsilon_K(K\bar{x}_i)/K\varepsilon_K = \bar{x}_i$
- The descendance of a mutant reaches macroscopic abundance with probability $1 - \mu/\lambda$.

A non-exchangeable coalescent process

In the new timescale, at constant rate

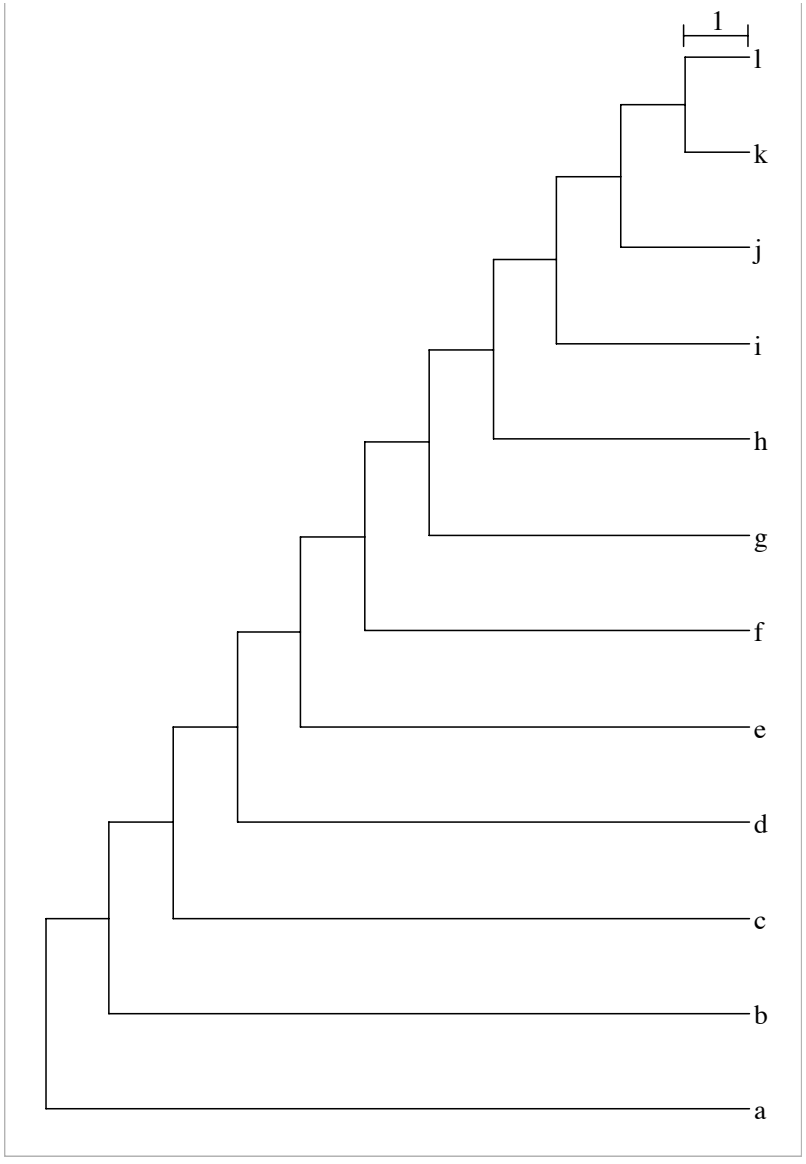
$$\rho = \frac{\kappa}{1-\alpha}\left(1-\frac{\mu}{\lambda}\right) = \frac{(\lambda-\mu)^2}{\lambda d}$$

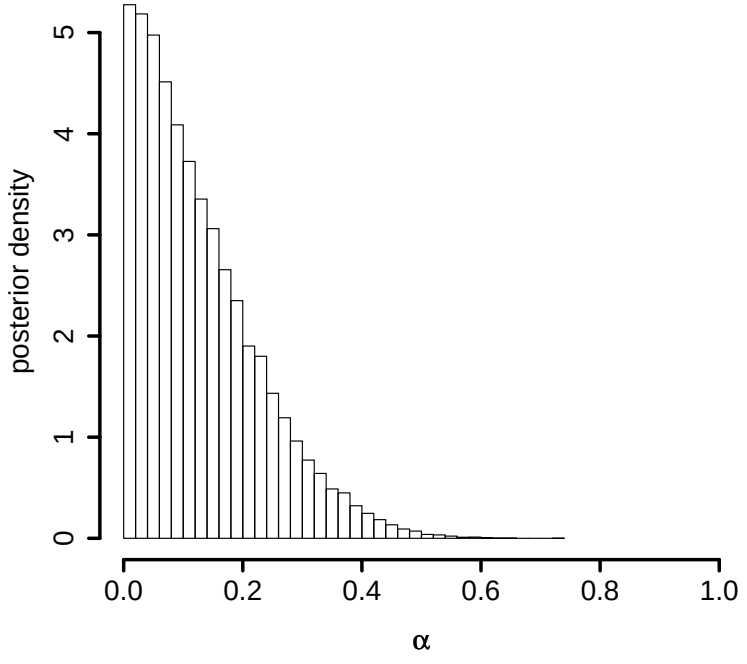
- Speciation occurs from the sp at level i, with proba $(1-\alpha)\,\alpha^{i-1}$
- All species simultaneously "shift up" their level by $+1$
- The new species occupies the newly vacated bottom level = youngest species.
- Backwards-in-time picture = Shift-Down/Look-Up Coalescent

Work done and perspectives

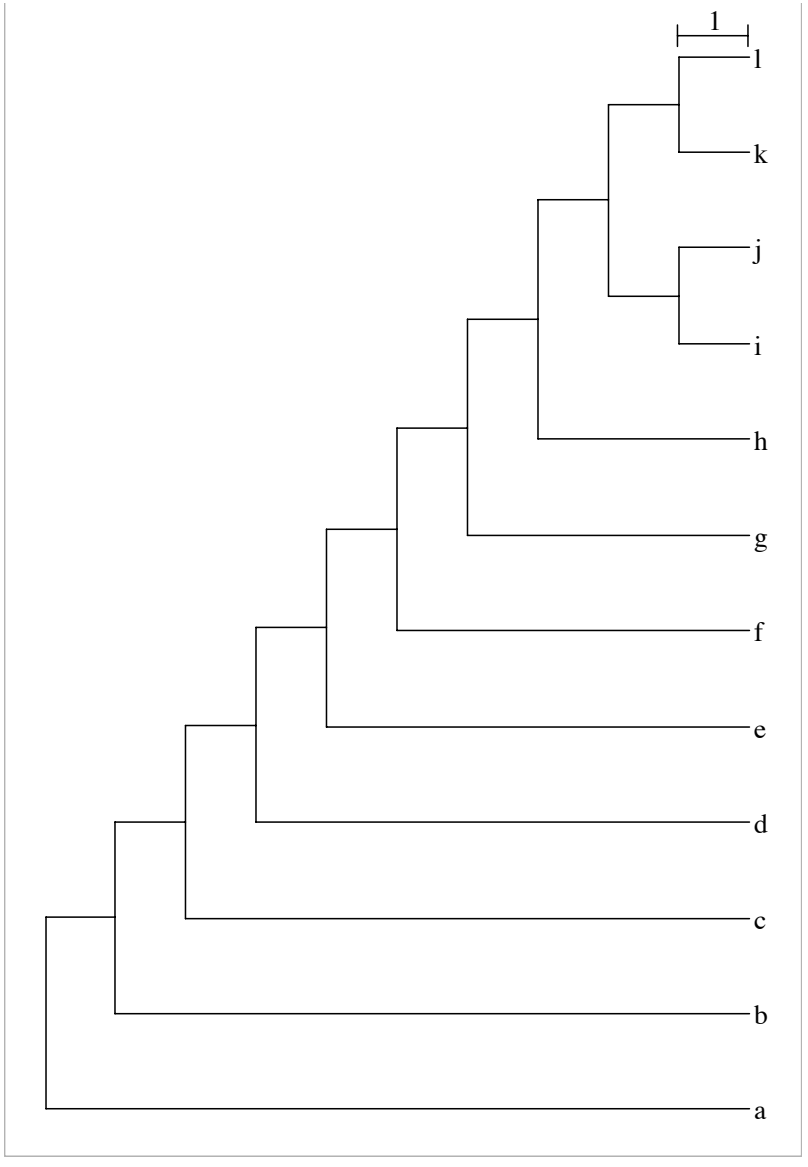
- Reduction of state-space for fast simulation of the phylogeny of a sample of species
- Likelihood computation after data augmentation : MCMC inference algorithm
- Perspectives : tests by simulations, distribution of β and γ vs α
- Other perspective : ecological release = competition suffered from a subset of younger species

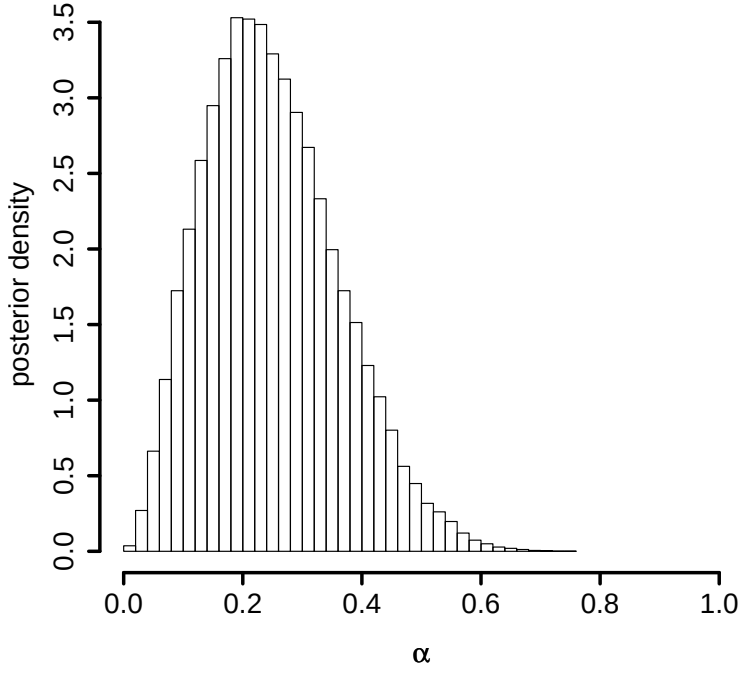
MCMC inference (1) : Caterpillar tree



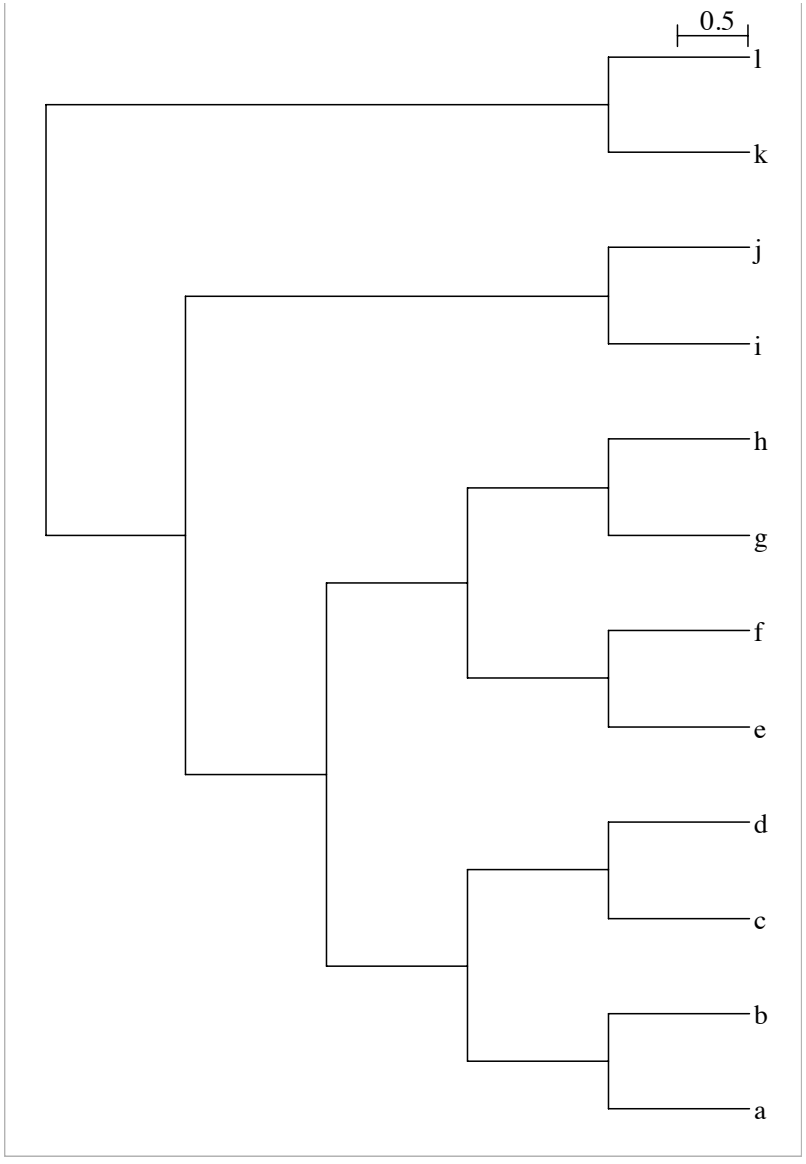


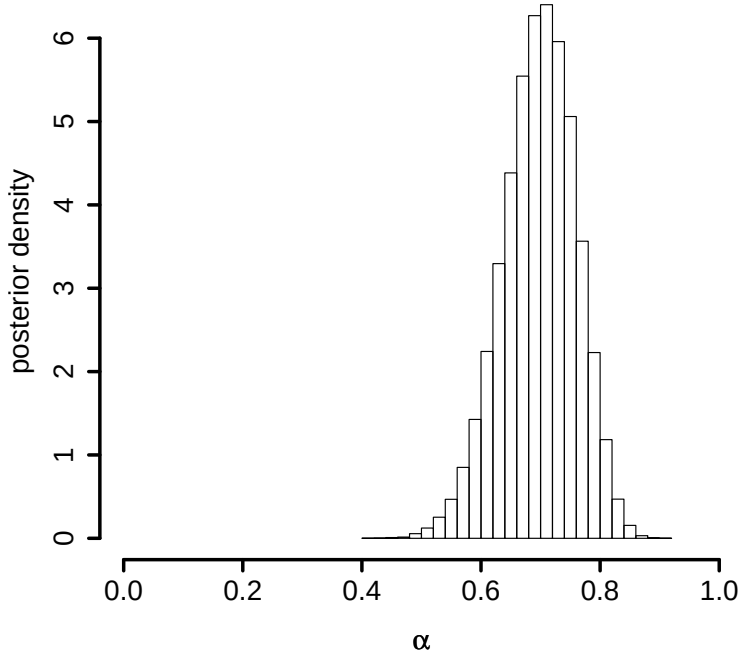
MCMC inference (2) : Very imbalanced tree



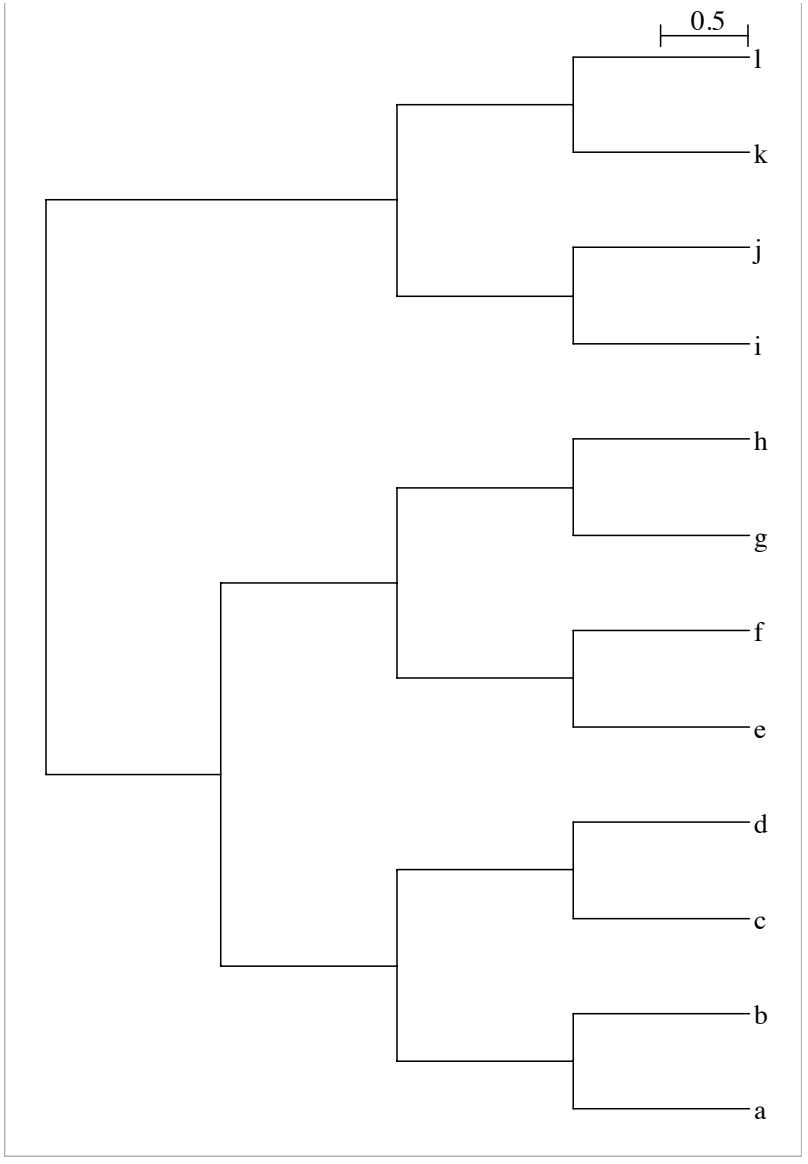


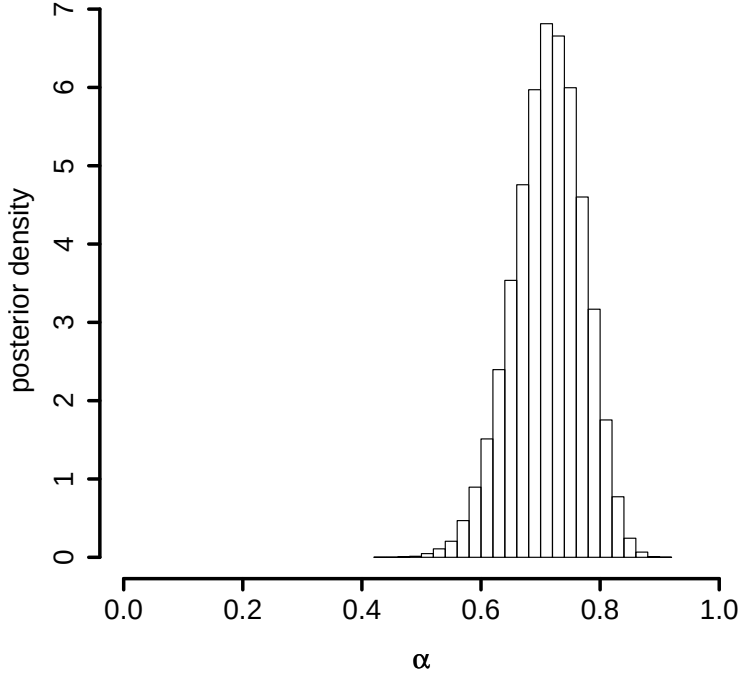
MCMC inference (3) : Balanced tree





MCMC inference (4) : Very balanced tree





Acknowledgements

- Thanks to my co-authors
  - G. Achaz (UPMC & SMILE, Paris)
  - H.K. Alexander (ETH Zürich)
  - R.S. Etienne (U Groningen)
  - N. Lartillot (CNRS & U Lyon)
  - H. Morlon (CNRS & École Normale Supérieure, Paris)
  - T.L. Parsons (CNRS & SMILE, Paris)
  - T. Stadler (ETH Zürich)
- Thanks to the members of the SMILE group

Institutions

- ***Stochastic Models for the Inference of Life Evolution*** **(SMILE)**
  ⊂ Center for Interdisciplinary Research in Biology
  ⊂ Collège de France

- ***Stochastics & Biology group***
  ⊂ Laboratoire de Probabilités et Modèles Aléatoires
  ⊂ UPMC University Paris 06

- **ANR** ***Modèles Aléatoires eN Écologie, Génétique, Évolution*** **(MANEGE)**