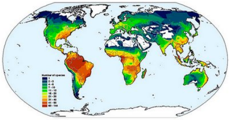
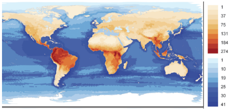
A Non-Exchangeable Coalescent Arising in Phylogenetics

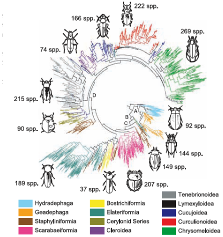
Amaury Lambert
(joint work with G. Achaz, N. Lartillot, T.L. Parsons)

CIRM
Luminy, June 17, 2015

Pattern & Process





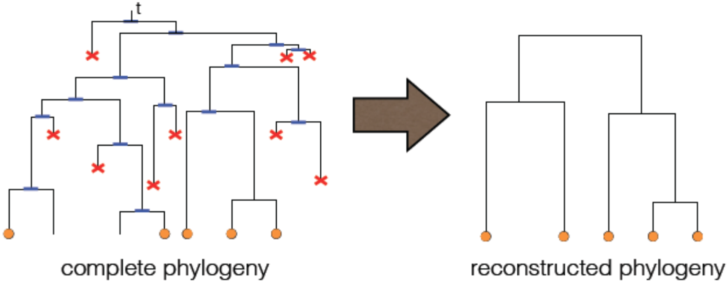


- Design probabilistic models of evolutionary processes...
- ...Generating similar patterns as those observed in nature, and...
- ...Allowing for the inference of these processes from real data...
- ...Assuming the data is a phylogeny (gene tree, species tree,...) already inferred from MSA.

Outline

1. Properties of « Ultrametric Trees »
2. Examples & applications
3. A Non-Exchangeable, Individual-Based Model of Phylogeny
4. Simulations and Inference

Reconstructed tree



- **« Reconstructed tree » or « reduced tree »** at height T = remove all lineages extinct by T (fixed time).

- The reduced tree is one-to-one with...

- ...The **sphere of radius** T $\{x : d(\text{root}, x) = T\}$ = particles alive at time T (yellow dots)

- The sphere is ultrametric : $d(x,z) \leq \max(d(x,y), d(y,z))$.

Comb metric (1)

Let I be a compact interval and $f : I \to \mathbb{R}_+$.

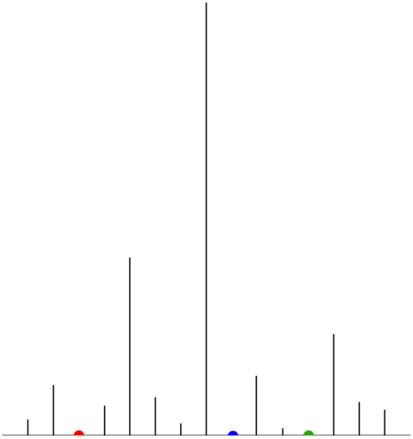
Definition

The mapping f is called a ***comb*** *if for any $\varepsilon > 0$, $\{f \geq \varepsilon\}$ is finite.*
For any $s, t \in I$, define d_f by

$$d_f(s,t) = 2 \sup_{(s \wedge t, s \vee t)} f.$$

Then d_f is an ultrametric distance on $\{f = 0\}$ (properly quotiented) called the comb metric.

Comb metric (2)

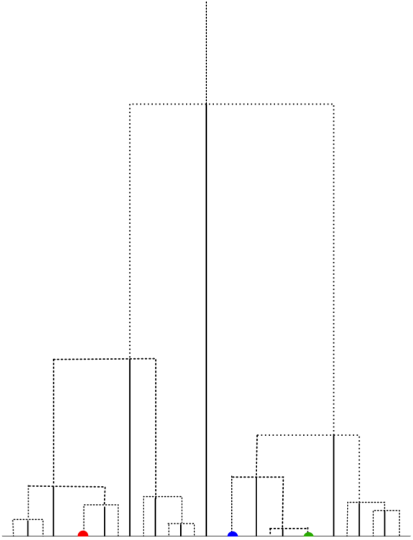


When the comb has finite support,

the comb metric space

is one-to-one with...

Comb metric (3)



When the comb has finite support,

the comb metric space

is one-to-one with...

An « ultrametric tree »

What about the general case ?

A representation theorem

Theorem (L. 2015)

Any compact, ultrametric space with no isolated point is isometric to a (properly completed) comb metric space.

In particular, any sphere $\{x \in t : d(root, x) = T\}$ *of a locally compact real tree* (t, d) *having no isolated point, is isometric to a comb metric space.*

The spheres of the Brownian tree can be represented by a comb whose graph is a Poisson point process with intensity $dx\, y^{-2}\, dy$ (properly stopped).

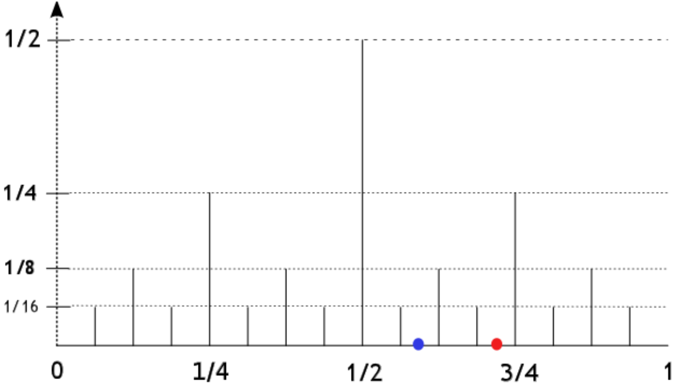
For Lévy trees, see L. & Popovic, *Ann. Appl. Prob.* (2013).

Outline

1. Properties of « Ultrametric Trees »
2. Examples & applications
3. A Non-Exchangeable, Individual-Based Model of Phylogeny
4. Simulations and Inference

Expl1. The $p-$adic comb

- $U :=$ Non stationary sequences of 0's and 1's with Hamming distance
$$d_H(x,y) = 2^{-\min\{n: x_n \neq y_n\}}$$
- $(x_n) \mapsto \sum x_n 2^{-n}$ maps (U, d_H) to the dyadic comb (see fig)
- Blue dot $= (1,0,0,1,\ldots)$ Red dot $= (1,0,1,1,\ldots)$



Expl 2. Exchangeable coalescents

...and Aldous' construction

Let f be a comb on $[0,1]$ and (V_i) i.i.d. random variables uniform in $(0,1)$. Define the partition $R_f(t)$ on $\mathbb{N}$ induced by the equivalence relation $\sim_t$

$$i \sim_t j \Leftrightarrow d_f(V_i, V_j) \leq t.$$

The process $(R_f(t); t > 0)$ is an exchangeable coalescent process.

For example, take

$$f = \sum_{j \geq 1} \tau_j \mathbb{1}_{U_j},$$

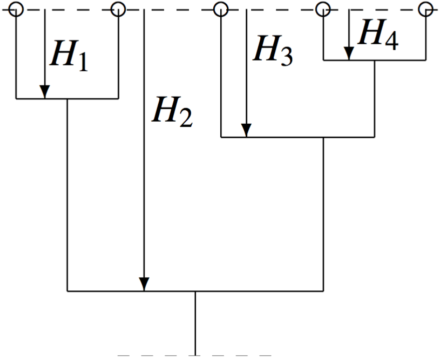
where the (U_j) are i.i.d. uniform on $(0,1)$ and $\tau_j = \sum_{k \geq j+1} e_k$, where e_k are independent exponential r.v. with parameter $k(k-1)/2$, then the process $(R_f(2t); t \geq 0)$ has the same law as the Kingman coalescent.

Expl3. The coalescent point process

(Popovic 2004, Aldous & Popovic 2005)

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

- More general definition via Poisson point processes (cf Brownian tree)



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

L.& Stadler, *TPB*, 2013

Consider a birth–death process started at time 0 with 1 particle and

- Birth rate $b = b(t)$, where t is time
- Death rate $d = d(t,a)$, where a is any non-heritable trait (e.g. age).

Theorem (L. & Stadler 2013)

The reconstructed tree at time T is a CPP with typical node depth H, where the function $F = 1/P(H > \cdot)$ is the unique solution to a linear integro-differential equation with initial condition $F(0) = 1$.

If b and d are time-homogeneous, F can also be obtained by inverting an explicit Laplace transform.

The result still holds with bottlenecks/partially sampled tips.

$\Rightarrow$ Likelihoods in product form $\Rightarrow$ Applications...

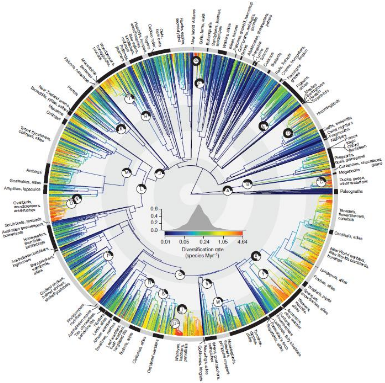
Appl. 1 « Do species age ? »

Alexander, L., Stadler, *Systematic Biology* (2015 ?)

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

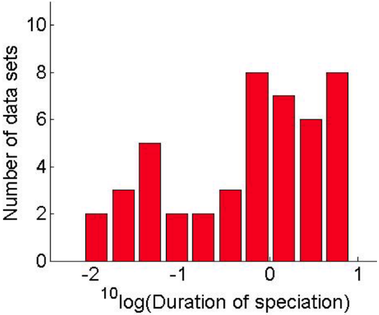
$$g(a) = \Gamma(k)^{-1} s^{-k} a^{k-1} e^{-a/s}$$

- Test on simulations : accurate MLEs of b and m
- MLE on *Aves* phylogeny = 9993 extant bird sp (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 $b = 0.108\ My^{-1}$



Appl.2 *« How long does speciation take ? »*

Etienne, Morlon, L., *Evolution* (2014)



- Speciation takes time

  = new populations take time to diverge from mother pop until total reproductive isolation

- Test on simulations : efficient inference of duration of speciation

- Left : duration of speciation inferred in 46 bird clades (in My)

Other models of reconstructed trees ?

- Advantages of CPP as models of phylogenies :
  - Process-based
  - Mathematically tractable
  - Likelihood-based methods available Stadler (2011), Morlon, Parsons & Plotkin (2011), L. & Stadler (2013), Etienne, Morlon & L. (2014), L., Morlon & Etienne (2015), Alexander, L. & Stadler (2015)...
- Shortcomings :
  - Lineage-based : No insight at the ind level, no predictions at the population level
  - Topology always equivalent to Yule tree = Uniform over trees with ranked node depths

Outline

1. Properties of « Ultrametric Trees »
2. Examples & applications
3. A Non-Exchangeable, Individual-Based Model of Phylogeny
4. Simulations and Inference

Goal

In this second part, our goal is to propose :

- A biologically reasonable model of phylogeny
  - Individual-based
  - Where species play different roles
- Mathematically tractable
- Fitting empirical patterns

The Red Queen Hypothesis

- “Old species are continually replaced by younger, fitter species”
- Examples
  - Key innovations, niche invasions
  - Evolutionary arms races
- No parameterization of fitness = fitness mediated by order of appearance

Asymmetric multispecies model

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

- Individual-based model with n species = multitype logistic branching process (Ethier & Kurtz 1980, L. 2005)
- *Per capita* birth rate λ, death rate μ
- Death by competition at **rate** c_{ij} **felt by** each ind of sp i, **from** each ind of sp j, where **sp i is *younger* than sp j** and

$$\begin{cases} c_{ij} & = & 0 \\ c_{ii} & = & c/K \\ c_{ji} & = & d/K \end{cases}$$

Large population limit

Now species have **levels** :
Species at level 1 = youngest species,
Species at level 2 = 2nd youngest species,...

If $K^{-1}X_i(0)$ converge as $K \to \infty$, then $K^{-1}(X_i) \Rightarrow (x_i)$ (Kurtz 1981)

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

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

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

⇒ Younger species are more abundant.

Speciation by point mutation

Each newborn is a mutant with probability ε_K, where for all $V > 0$,

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

Separation of timescales as $K \to \infty$:

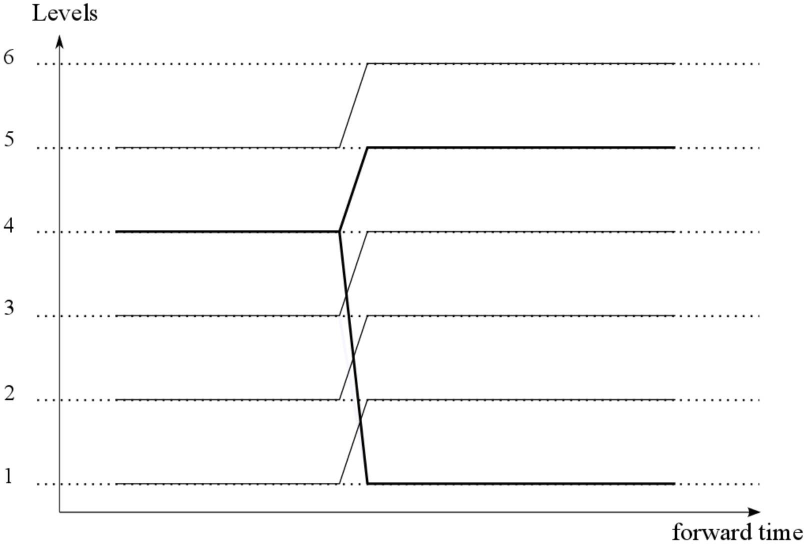
Theorem

Set $T_N :=$ first time when the number of species exceeds N.

Let $(N_t; t \geq 0)$ be a pure-birth process with birth rate $\rho_n = \lambda \left(1 - \frac{\mu}{\lambda}\right) \sum_{i=1}^{n} \bar{x}_i$.

Then, as $K \to \infty$, the process $K^{-1}(X_i)\left(\frac{1}{K\varepsilon_K}(t \wedge T_N)\right)$ converges (fdd) to the process $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{N_t - 1}, 0, \ldots, 0)$.

Speciation in forward time...



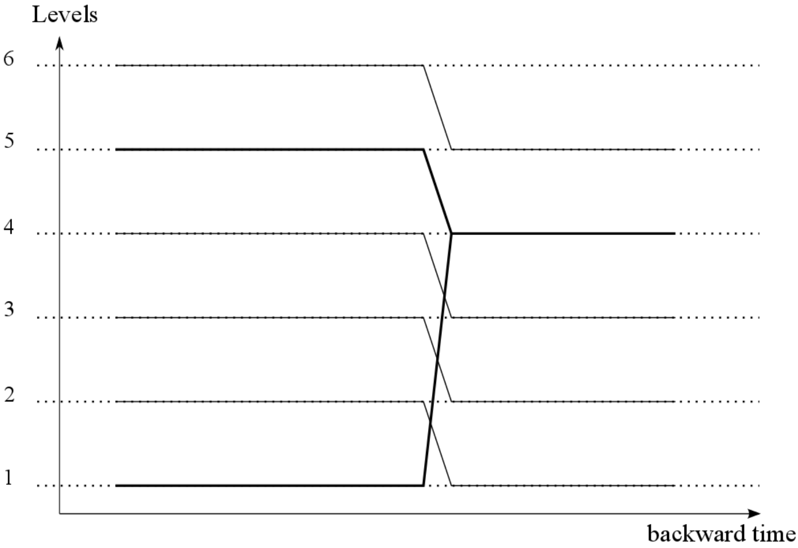
A non-exchangeable coalescent process

In the new timescale, at constant rate

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

- 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

...Coalescence in backward time



Intertwining (Rogers & Pitman 1981)

Let $((X_t, Y_t), t \geq 0)$ a Markov process with state-space $E \times F$ with generator $\hat{G}$ and K a probability kernel from E to F with associated operator

$$Kf(x) = \int_F K(x, dy) f(x, y).$$

Theorem (Rogers & Pitman 1981)

If there exists a generator G of a Markov process in E such that for each $f : E \times F \to \mathbb{R}$ in the domain of $\hat{G}$,

$$K\hat{G}(f)(x) = GK(f)(x) \quad x \in E,$$

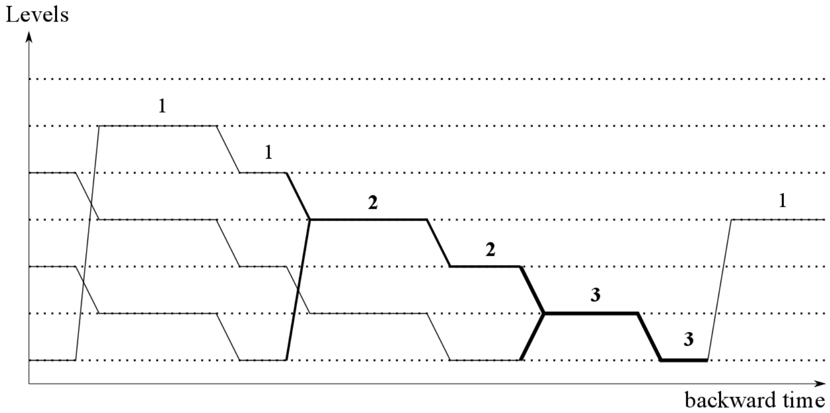
then

1. $\mathbb{P}(Y_0 \in dy | X_0) = K(X_0, dy)$ *a.s. implies that for each $t > 0$,*

$$P(Y_t \in dy | (X_s, 0 \leq s \leq t)) = K(X_t, dy) \quad a.s.$$

2. $(X_t, t \geq 0)$ *is a Markov process.*

The weight measure (1)



Weight = 1 + Number of coalescences ‘from below’ since last visit of level 1

= Number of ‘delayed’ lineages (i.e., coal. only when leaving level 1)

Intertwining (1)

$W_t(\ell) =$ weight of level ℓ = number of 'delayed' lineages at level ℓ

$N_t := W_t(\mathbb{N}) =$ number of 'delayed' lineages.

Theorem

$(N_t; t \geq 0)$ is a $\delta_{1-\alpha}$ coalescent process and conditional on $(N_s; 0 \leq s \leq t)$,

$$W_t = \sum_{i=1}^{N_t} \delta_{G_i},$$

where the G_i's are i.i.d. Geom(α) random variables.

Intertwining (2)

$W_t(\ell) =$ weight of level ℓ = number of 'delayed' lineages at level ℓ

$B_t(w) =$ number of lineages with weight w.

Theorem

$(B_t;t \geq 0)$ is a Markov process and conditional on $(B_s;0 \leq s \leq t)$,

$$W_t = \sum_{w \geq 1} \sum_{i=1}^{B_t(w)} \delta_{Y_{wi}},$$

where the Y_{wi}'s are independent Geom(α^w) random variables, conditioned to be pairwise distinct.

Outline

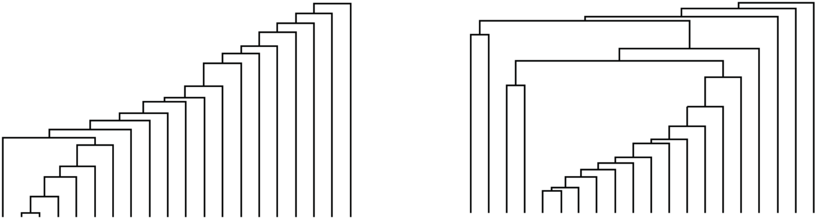
1. Properties of « Ultrametric Trees »

2. Examples & applications

3. A Non-Exchangeable, Individual-Based Model of Phylogeny

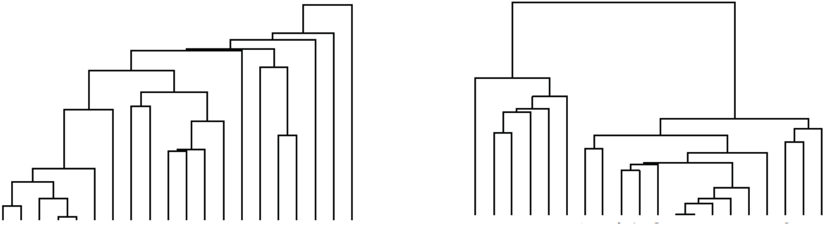
4. Simulations and Inference

Simulated trees with 20 tips



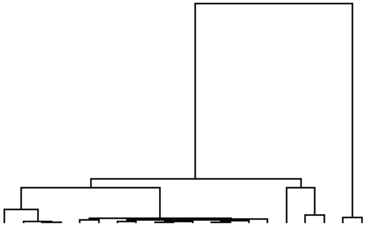
$$\alpha = 0.1$$

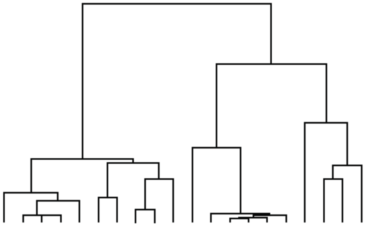
Simulated trees with 20 tips



$$\alpha = 0.7$$

Simulated trees with 20 tips





$$\alpha = 0.99$$

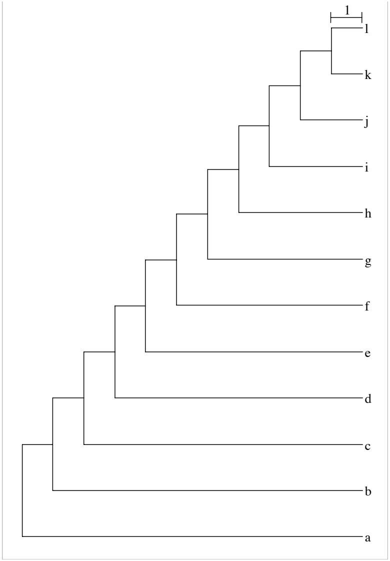
Convergence to the Kingman coalescent

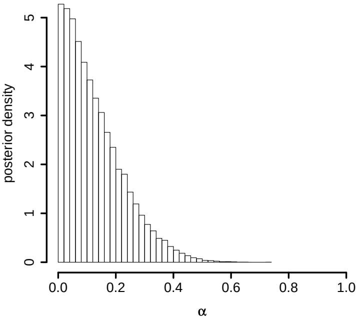
Recall $\alpha = 1 - d/c$ and $\kappa = (\lambda - \mu)/c$ = abundance of youngest species.

Theorem

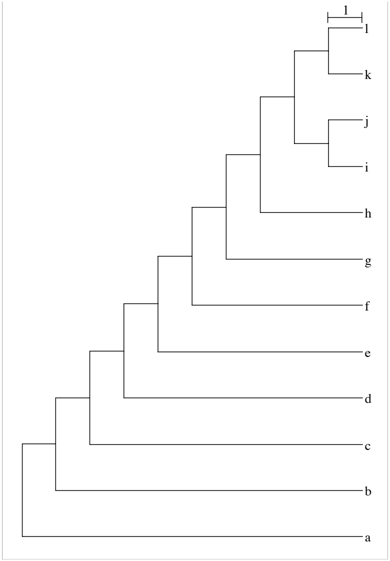
As $\alpha \to 1$, the process $(B_{t/(1-\alpha)}; t \geq 0)$ converges (fdd) to $N_t \delta_1$, where $(N_t; t \geq 0)$ is a pure-death process with death rate $Cn(n-1)/2$, where $C = (1 - \mu/\lambda)\kappa$ (replacement rate).

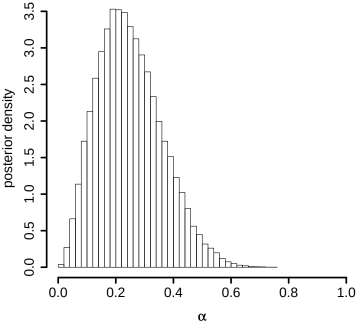
MCMC inference (1) : Caterpillar tree



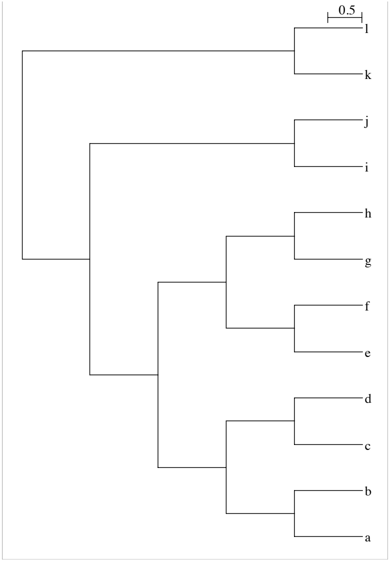


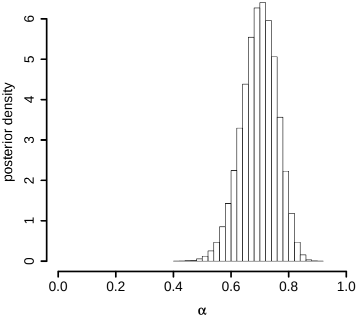
MCMC inference (2) : Very imbalanced tree



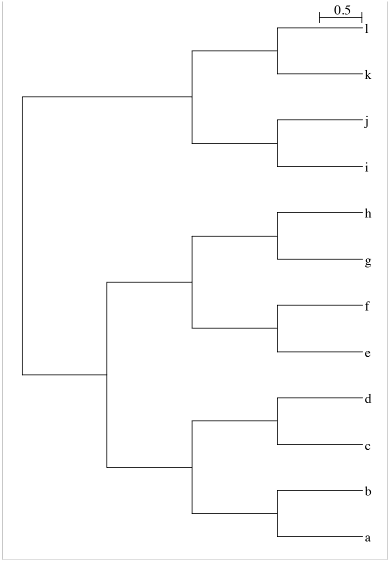


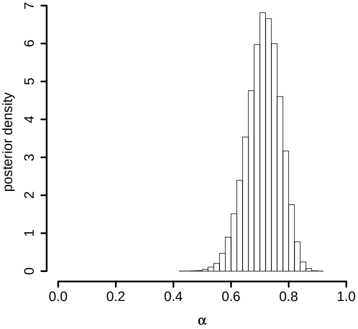
MCMC inference (3) : Balanced tree





MCMC inference (4) : Very balanced tree





Conclusion and perspectives

- A simple model of phylogeny based on an individual-based model of evolution under the Red Queen hypothesis see also Chisholm & O'Dwyer (2014)
- Reduction of state-space for fast simulation of the phylogeny of a sample of species
- Convergence to Kingman coalescent as $\alpha \to 1$
- Likelihood computation after data augmentation : MCMC inference algorithm
- WIP : Distributions of β and γ vs α
- WIP : Inference in the transient phase, inference under models of niche colonisation (Verónica Miró Pina)

Institutions

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

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

Acknowledgements

- Thanks to my co-authors
  - G. Achaz (UPMC & SMILE, Paris)
  - N. Lartillot (CNRS & U Lyon)
  - T.L. Parsons (CNRS & SMILE, Paris)
- Thanks to the members of the SMILE group

SMILE group in May 2015

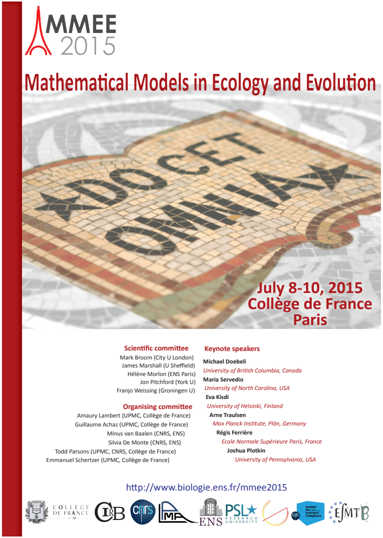
Conference announcement

Mathematical Models in Ecology & Evolution

Collège de France, Paris, France

July 8–10, 2015

`http://www.biologie.ens.fr/mmee2015/`


DOCET OMNIA

MMEE 2015

Mathematical Models in Ecology and Evolution

July 8-10, 2015
Collège de France
Paris

Scientific committee

Mark Broom (City U London)
James Marshall (U Sheffield)
Hélène Morlon (ENS Paris)
Jon Pitchford (York U)
Franjo Weissing (Groningen U)

Organising committee

Amaury Lambert (UPMC, Collège de France)
Guillaume Achaz (UPMC, Collège de France)
Minus van Baalen (CNRS, ENS)
Silvia De Monte (CNRS, ENS)
Todd Parsons (UPMC, CNRS, Collège de France)
Emmanuel Schertzer (UPMC, Collège de France)

Keynote speakers

Michael Doebeli
University of British Columbia, Canada
Maria Servedio
University of North Carolina, USA
Eva Kisdi
University of Helsinki, Finland
Arne Traulsen
Max Planck Institute, Plön, Germany
Régis Ferrière
Ecole Normale Supérieure Paris, France
Joshua Plotkin
University of Pennsylvania, USA

http://www.biologie.ens.fr/mmee2015