Coalescent trees of birth–death models
&
Applications to phylogenetics

Amaury Lambert
(with R.S. Etienne, M. Manceau, H. Morlon, T. Stadler)

Journée ANR MANEGE du 26 novembre 2013

Outline

Evolutionary Biology and Math

- Evolutionary biologists...
  - collect data : contemporary geno/phenotypes
  - identify patterns
  - postulate evolutionary processes responsible for these patterns.

  ...But since those processes can not be reproduced *in vivo*...

- We (mathematicians)...

  - propose simple models underlying the evolutionary processes
  - predict the patterns generated by these models
  - quantify the ability of competing processes to generate the observed patterns.
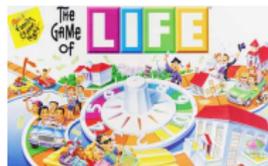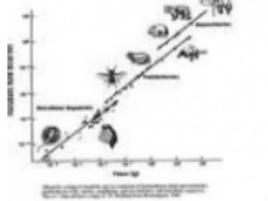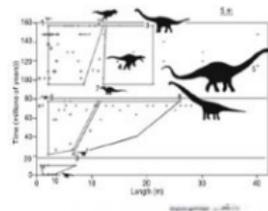
Understanding biodiversity patterns
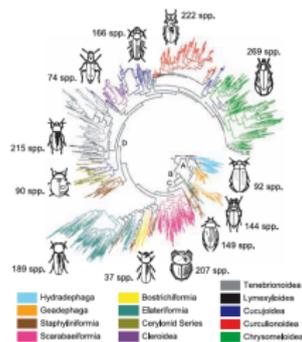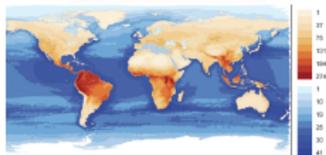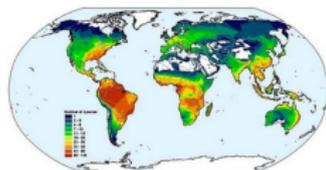
- Why are there so many species in the tropics ?

- Why are there so few species in the oceans ?

- Why are some taxonomic groups so much richer than others ?

- Infer diversification processes to see how these processes vary across time, geographic regions, habitats, and taxonomic groups.

Birth–death models of genealogies/phylogenies

- We use birth–death models of diversification

- 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

- N_t := nb particles at time t

- Particles may bear some trait i, and rates may depend on t, N_t, i,...

complete phylogeny
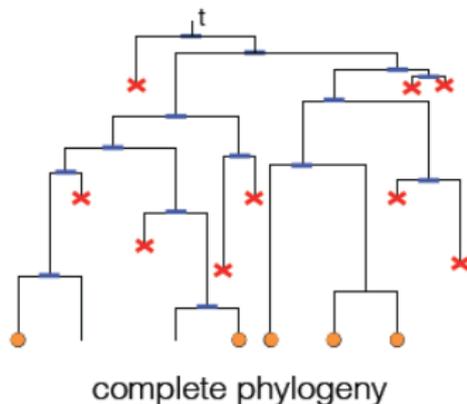
Assumptions on rates

Rates $b(t,n,a,i)$ and $d(t,n,a,i)$ may depend upon :
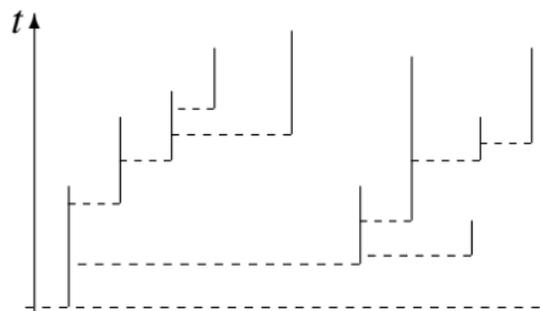
- **time** t
- **number** n of standing particles
- a **non-heritable trait** a (e.g., age)
- a **heritable trait** i
- Traits behave as iid Markov processes on each lineage
- Asymmetric birth =
  Mother keeps her trait
- Orientation =
  Daughter sprouts to the right

Outline

Reconstructed (or reduced) tree

Reconstructed tree = start with one particle at time 0, stop at time T, remove all lineages extinct by T.

complete phylogeny   reconstructed phylogeny
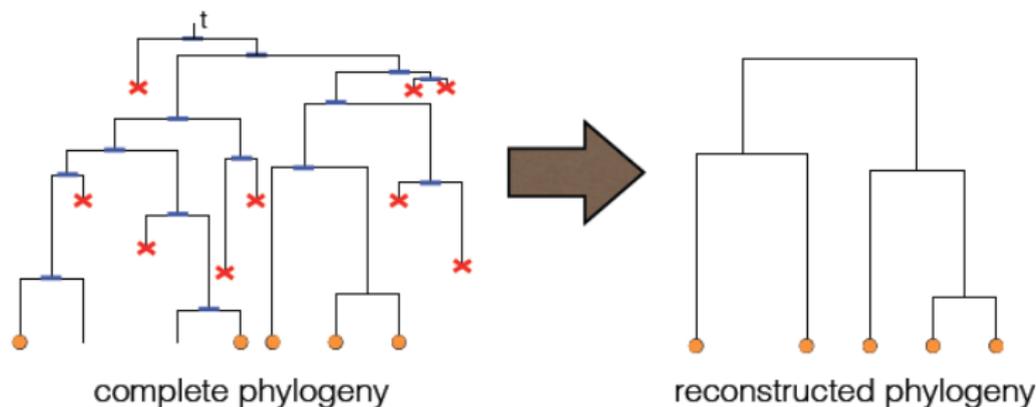
- Use time-calibrated phylogenies to infer diversification processes
- By computing the likelihood of reconstructed trees (ML, MCMC) and estimating div rates
- **Q1 :** <u>Can we characterize the distribution of the reconstructed tree under a generalized birth–death model of diversification ?</u>

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

A reference distribution on ultrametric, oriented trees with edge lengths

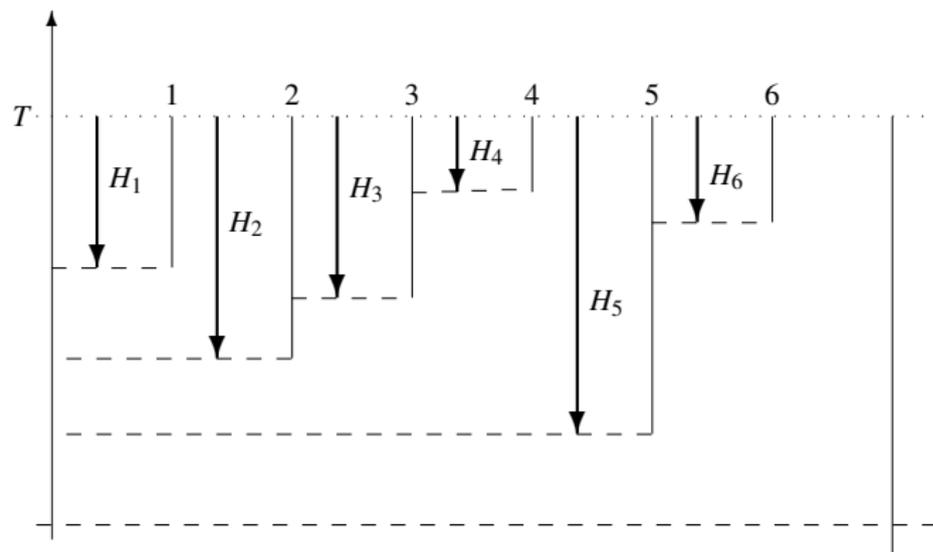**CPP = Coalescent Point Process** = Oriented tree whose node depths $H_1, H_2, \ldots$, form a sequence of **iid random variables** killed at its first value larger than T.

The likelihood of a tree with node depths $h_1, \ldots, h_{n-1}$ can be factorized as a product
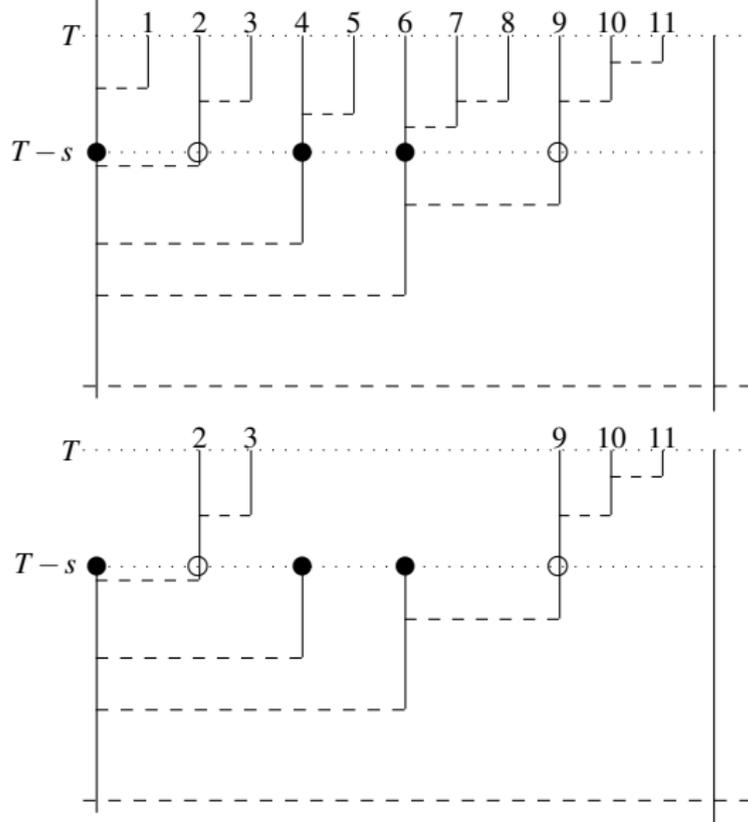
$$L(h_1, \ldots, h_{n-1}) = P(H > T) \prod_{i=1}^{n-1} f(h_i),$$

where f is the density of H.

Simulating CPPs

CPP with one bottleneck

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

Recall that t is time and a is any non-heritable trait.

Theorem (L. & Stadler 2013)

*If $b = b(t)$ **and** $d = d(t, a)$, then **the reconstructed oriented tree is a CPP** with typical node depth H whose distribution is given by*

$$P(H > t) = \exp\left(-\int_{T-t}^{T} b(s)\, p(s)\, ds\right) \qquad t \in [0, T],$$

where $p(t)$ denotes the probability that a species born at time t has extant descendance by time T.

*This still holds in the presence of **bottlenecks** = mass extinction events (fixed times, fixed probabilities).*

Law of H from model parameters (1)

For any $s \geq t$, let $g(t,s)$ **be the density at time s of the extinction time of a species born at time t.**

If $\mathbb{E}_{t,x}$ denote the expectation associated to the trait X started at time t in state x, then

$$g(t,s) = \int_{\mathbb{R}} \nu_t(dx) \, u_s(t,x) \qquad s \geq t,$$

where ν_t is the initial trait ditribution for sp born at t and

$$u_s(t,x) := \mathbb{E}_{t,x}\left(\mu(s,X_s) \, e^{-\int_t^s dr \, \mu(r,X_r)} \right) \qquad s \geq t.$$

If X is a Markov process with generator L_t at time t, then by Feynman–Kac formula, u_s is the unique solution to

$$\frac{\partial u_s}{\partial t}(t,x) + L_t u_s(t,x) = \mu(t,x) \, u_s(t,x),$$

with terminal condition $u_s(s,x) = \mu(s,x)$.

Law of H from model parameters (2)

Recall $g(t,s)$ is the density at time s of the extinction time of a species born at time t.

Proposition (L. & Stadler 2013)

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 $F(0) = 1$.

Two special cases

- If $b = b(t)$ and $d = d(t)$, then

$$F(t) = 1 + \int_{T-t}^{T} ds\, b(s)\, e^{\int_s^T du\, (b-d)(u)}.$$

- If b is constant and $d = d(a)$, then $g(s,t) = g(t-s)$
  ($g(a) = d(a)\, e^{-\int_0^a ds\, d(s)}$ if a the age), and

$$F' = b\,(F - F \star g),$$

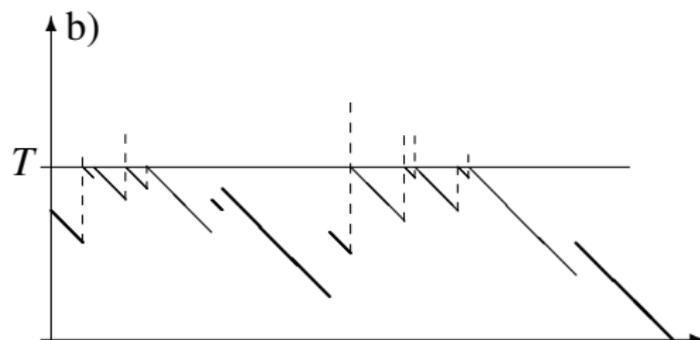with $F(0) = 1$. Equivalently, F is the unique non-negative function with Laplace transform
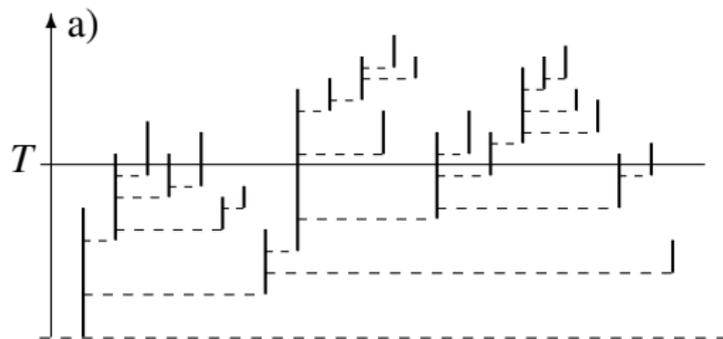
$$\int_0^\infty F(t)\, e^{-tx}\, dt = \frac{1}{\psi(x)},$$

where ψ is the Lévy exponent

$$\psi(\lambda) = \lambda - \int_0^\infty b\,g(t)\,\left(1 - e^{-\lambda t}\right) dt \qquad x \geq 0.$$

Jumping contour of a tree

a) Binary tree with edge lengths and b) Jumping contour process of its truncation below time T.
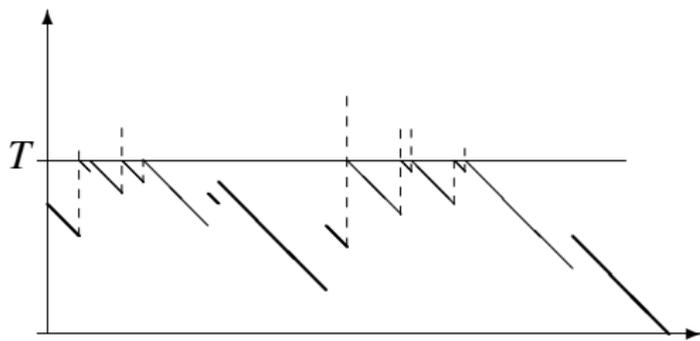
Contour of a splitting tree

Theorem (L. (2010))

The jumping contour process of a splitting tree truncated below T is a strong Markov process.

*In the time-homogeneous case, it is a **Lévy process with Lévy density** $bg(\cdot)$, **without negative jumps and drift** -1, reflected below T and killed upon hitting 0.*

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)
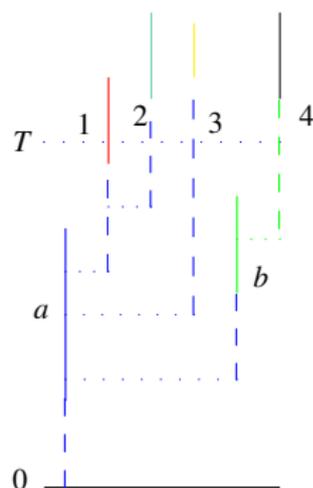
- 4 extant populations at time T

- 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.
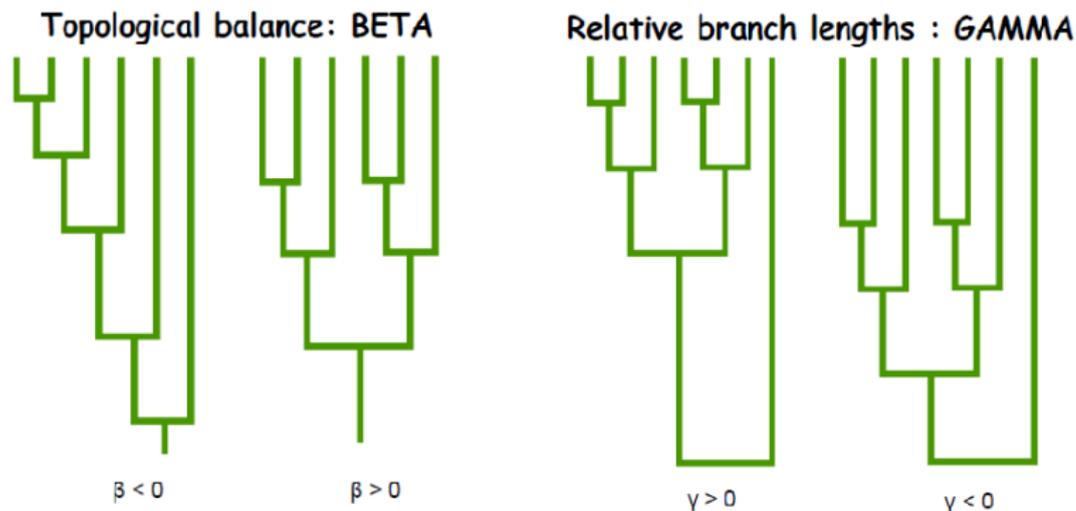
Protracted speciation (4)
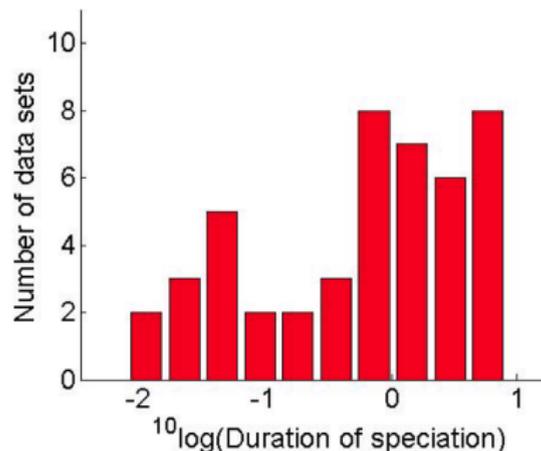
- Test on simulations : poor ML inference for each individual parameter

- 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

More real data : Two statistics

Topological balance: BETA

$\beta < 0$ $\beta > 0$

Relative branch lengths : GAMMA

$\gamma > 0$ $\gamma < 0$

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

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

The URT distribution

- Protracted speciation models produce $\gamma < 0$, BUT $\beta \approx 0$

- CPP = Fast simulation + fast inference, BUT always $\beta \approx 0$

- All CPPs have the same topology in distribution

- This topology is called **URT = uniform distribution** on ranked **oriented** trees after ignoring the orientation
  (The Kingman coalescent tree follows URT but not CPP)

- **Q2 :** <u>What conditions on rates are necessary and sufficient to produce CPP trees ? URT trees ?</u>
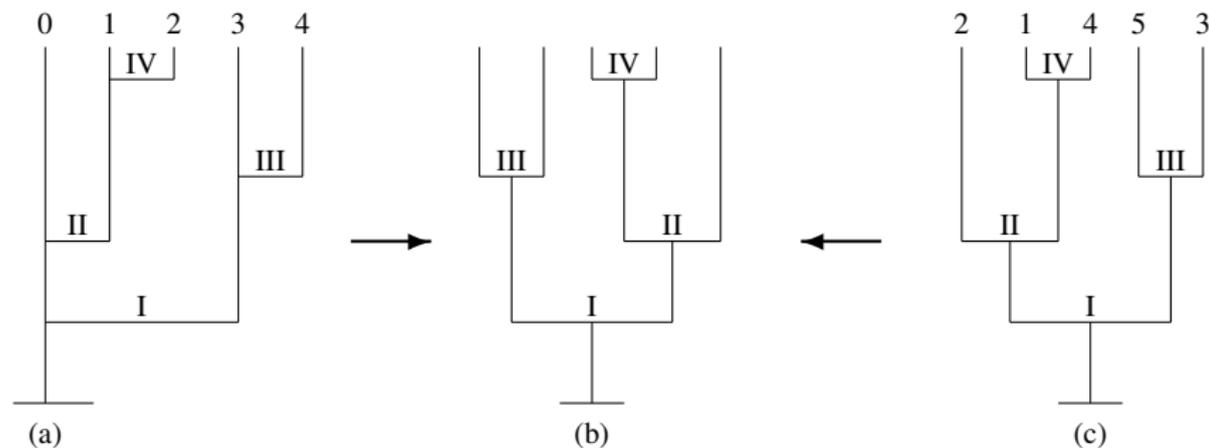
FIGURE : Under the uniform distribution on ranked oriented trees, the probability of the tree in (a) is $1/(n-1)! = 1/24$; under the uniform distribution on ranked labelled trees, the probability of the tree in (c) is $2^{n-1}/n!(n-1)! = 1/180$; under URT, the probability of the tree τ in (b) is $2^{n-1-c}/(n-1)! = 1/6$.

Answer to Q2

Proposition (L. & Stadler 2013)

1. *Reconstructed trees always follow CPP*
   IFF $b = b(t)$ and $d = d(t,a)$

2. *Reconstructed tree shapes always follow URT*
   IFF $b = b(t,n)$ and $d = d(t,n,a)$

Q3 : <u>Can we design tractable, trait-dependent models of diversification with both $\beta < 0$ and $\gamma < 0$?</u>

Outline

Speciation by genetic differentiation (1)
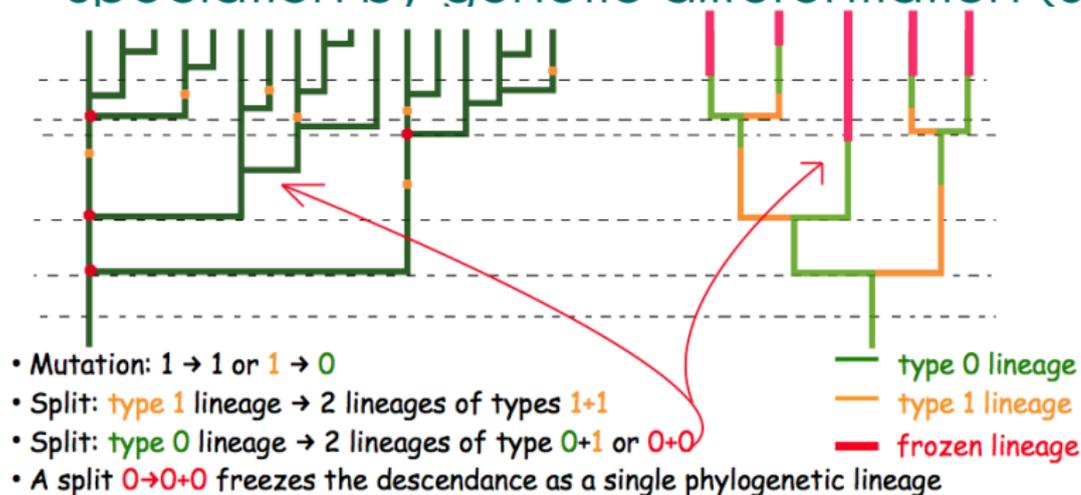## Work in progress with M. Manceau and H. Morlon

- 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)

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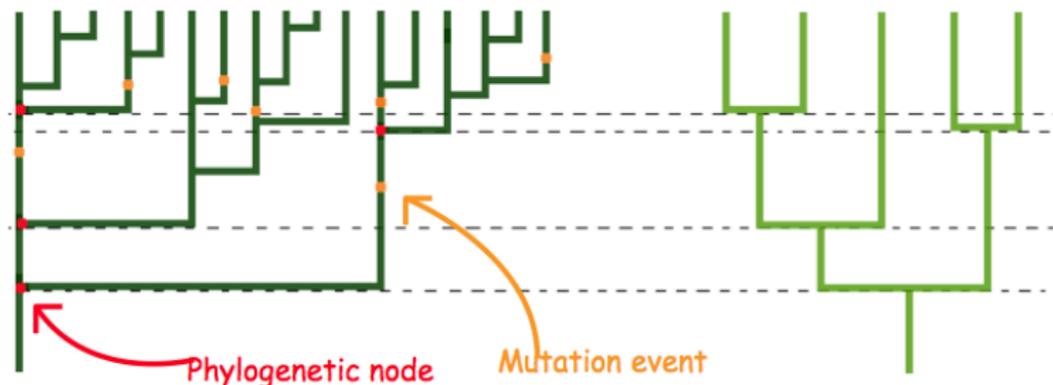
- A node on the genealogy is phylogenetic (= appears on the phylogeny) if

  (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)

- Mutation: 1 → 1 or 1 → 0
- Split: type 1 lineage → 2 lineages of types 1+1
- Split: type 0 lineage → 2 lineages of type 0+1 or 0+0
- A split 0→0+0 freezes the descendance as a single phylogenetic lineage

— type 0 lineage
— type 1 lineage
— frozen lineage

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

- Tests by simulations : accurate ML estimates of θ and $b - d$

- Values inferred from Cetaceans generate realistic values of β, γ
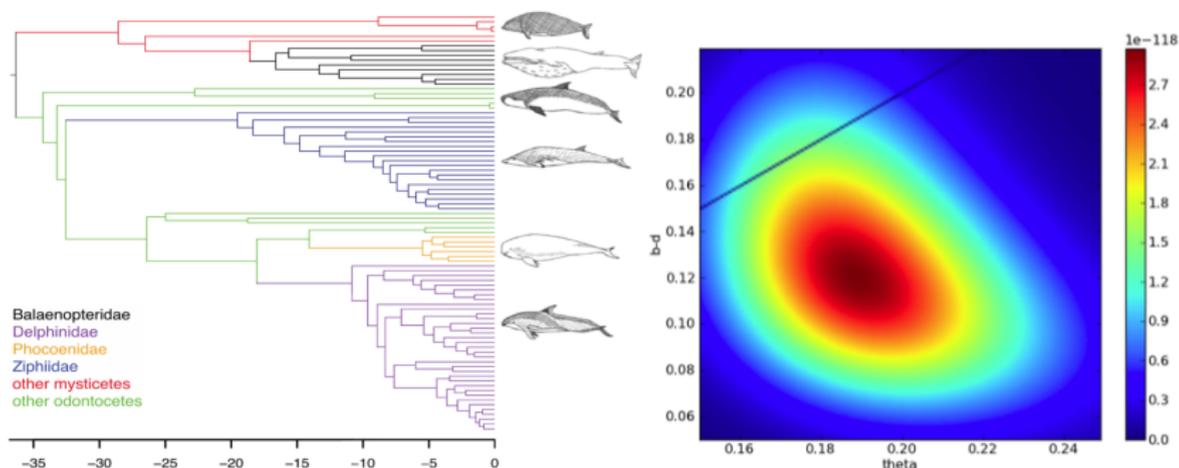
Speciation by genetic differentiation (5)
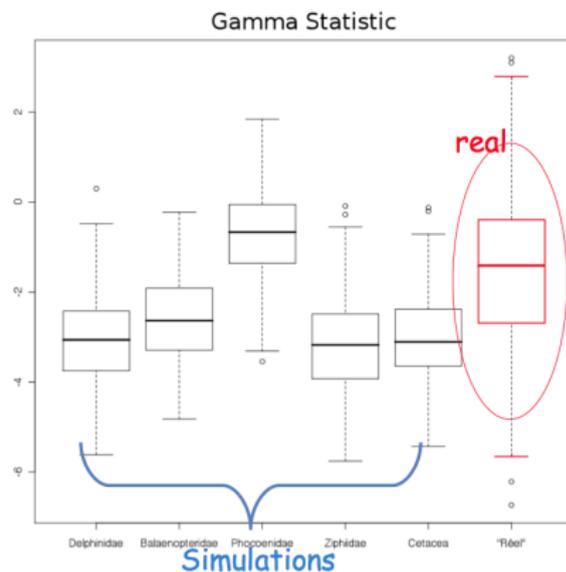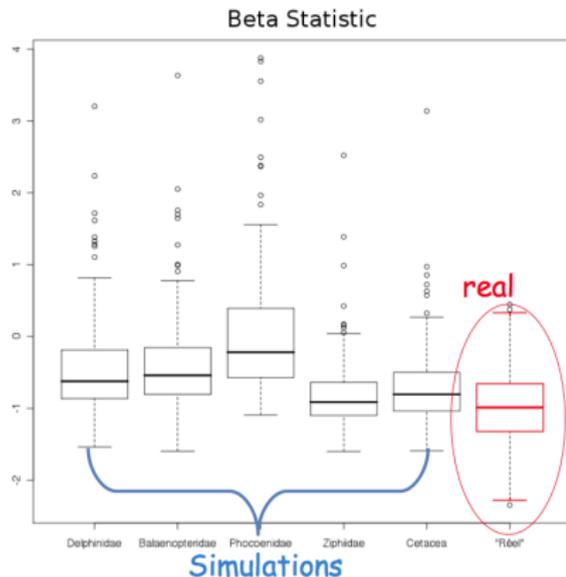
Institutions

- *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

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

SMILE : A cross-disciplinary group in CIRB

- CIRB = Center for Interdisciplinary Research in Biology (Collège de France)

- SMILE = Stochastic Models for the Inference of Life Evolution
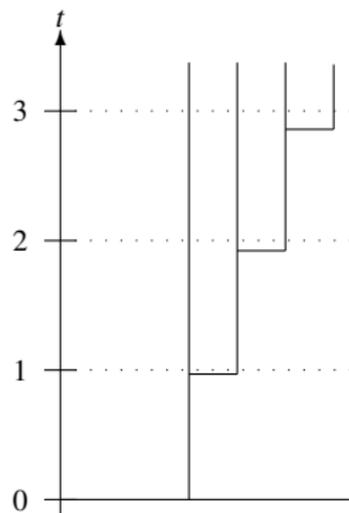
$b = b(a)$ fails to produce URT

FIGURE : Here, $d = 0$ and $b(a) = \mathbb{1}_{[1-\varepsilon, 1]}(a)$, where a is the age.

$\implies$ **Age-dependent speciation rates** can produce caterpillar trees w.h.p., and so **do NOT produce URT** trees in general.
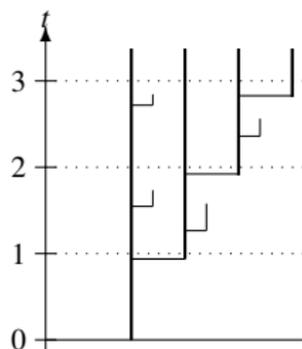
$d = d(i, a)$ fails to produce URT

FIGURE : Species can bear the heritable trait $i = 0$ or 1. *All sp bear trait 0 except (the anc and) when born from a sp with trait 1 and age a in $[1 - \varepsilon, 1]$.* Here $b = 1$, $d(1) = 0$ and $d(0) \gg -\log(\varepsilon)/\varepsilon$.

$\Longrightarrow$ **Heritable trait-dependent extinction rates** can produce caterpillar trees w.h.p. and so **do NOT produce URT** trees in general.

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

- If $b = b(t, n)$ and $d = d(t, n, a)$, then the law of the oriented tree is invariant under regrafting of subtrees (same time, different edge)

- The law of the oriented reconstructed tree is invariant under permutation of edges

- **The reconstructed tree shape always follows URT**.
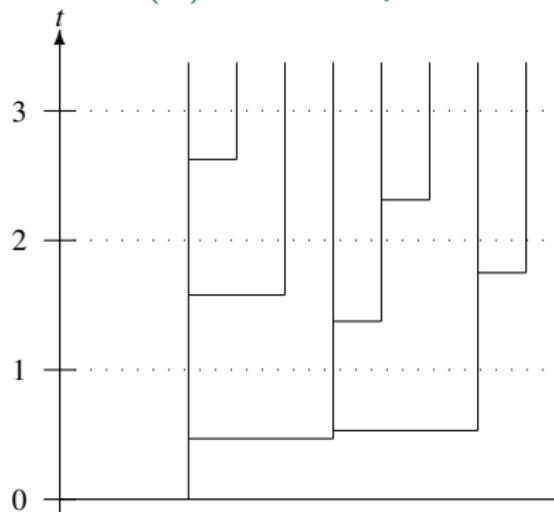
$b = b(n)$ or $d = d(n)$ fail to produce CPP

FIGURE : Here, $d = 0$ and $b = 1 + B\mathbb{1}_{n=2}$ with $B \gg 1$.
Alternatively $b = 1$ and $d(n) = D\mathbb{1}_{n=2}$, with $D \gg 1$.

$\implies$ **Rates dependent on the nb of species** can produce trees where the 1st and 2nd speciations are arbitrarily close w.h.p., and so do **NOT produce CPP** trees in general.