Time-calibrated phylogenies & Coalescent point processes

Amaury Lambert

Mathematical & Computational Evolutionary Biology
19 juin 2012, Hameau de l'Étoile (34)

General setting & goal

- Given a time-calibrated phylogenetic tree (= chronogram)
- And a (class of) forward models of diversification
- Infer past evolutionary dynamics from the tree
- By choosing the most likely model to have generated the observed pattern.

Models of random chronograms

- We use lineage-based models of phylogenies (but see last slides)
- Models with fixed (total) number of species produce bad phylogenies (Hey 1992)
- Widespread use of birth-death branching models
- Question : Law of the reconstructed tree ? (Nee et al 1994)
  = start at 0, condition on being alive at T, erase dead branches.

Biological motivations

- Nee 2006 : 'Familiarity with the patterns that random processes create is an essential piece of a scientist's mental furniture'
- H_0 : 'pattern is not distinguishable from that generated by a random process'... vs key adaptations, adaptive radiations, etc.
- Estimate speciation and extinction rates
- Understand how and why they vary across time, geographic regions, habitats and taxonomic groups.

Product likelihood

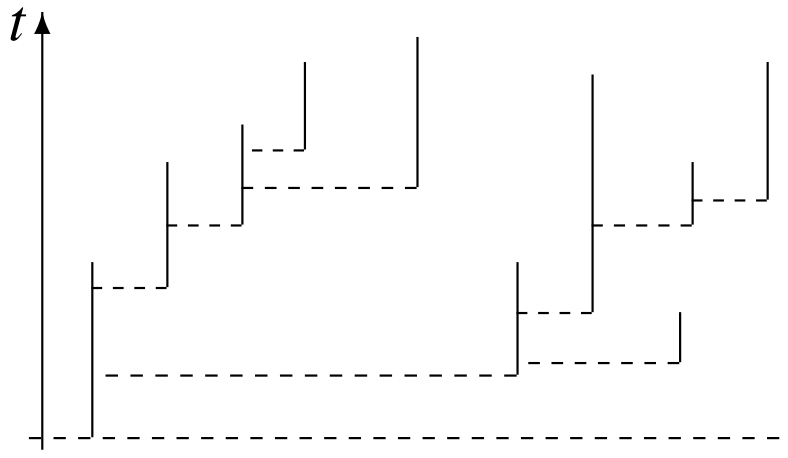
We will show that for a very general class of commonly used tree models, the likelihood of the reconstructed tree with splitting times $0 = t_1 < t_2 < \cdots < t_n < T$ can be factorized as a product

$$L(t_2, \ldots, t_n; T) = \prod_{i=2}^{n} f_T(t_i),$$

- easy inference of evolutionary past from the knowledge of the reconstructed tree
- $\neq$ Markovian coalescents (Kingman, Λ Lambda)

Splitting tree in forward time (Geiger & Kersting 97)

We consider a population of particles where

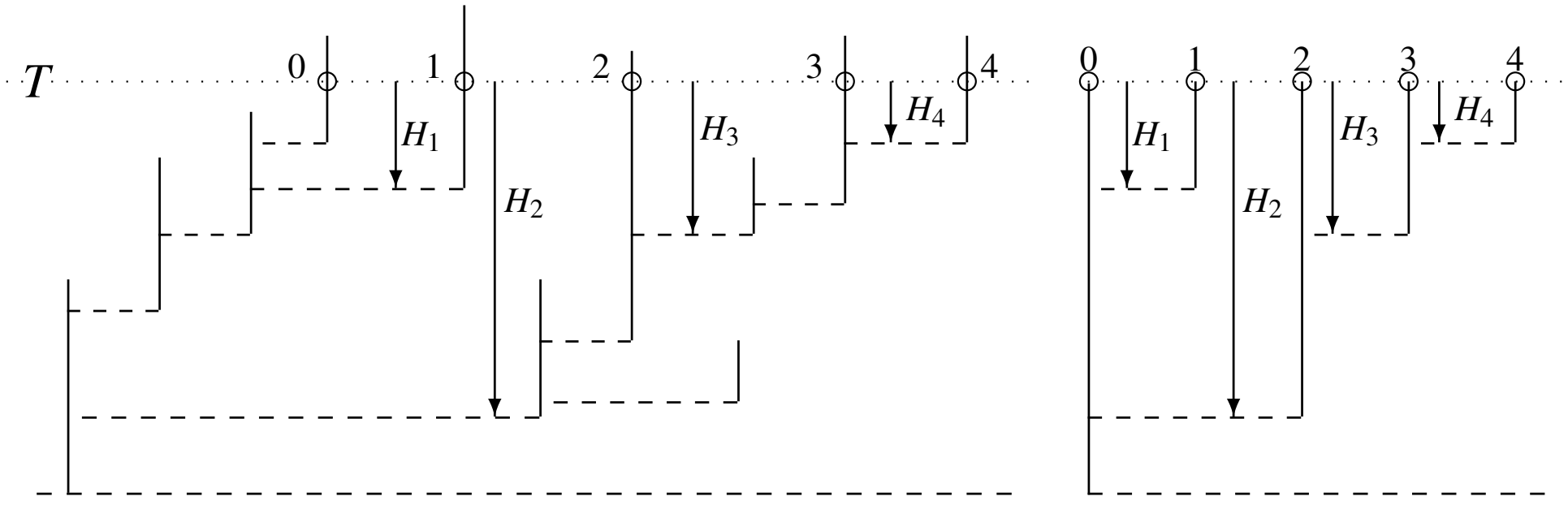


- particles reproduce independently
- particles may change type, provided type at birth does not depend on mother's type
- the death rate $\mu(t,a,i)$ may depend on absolute time t, age a and type i of particles
- the birth rate $\lambda(t)$ may depend on absolute time t (only)

This class of trees are called splitting trees (possibly multi-type and time-inhomogeneous).

Reconstructed tree

Tips can be labelled from left to right...



...and the times $H_1, H_2, H_3 \ldots$ are called coalescence times in population genetics and node depths in phylogenetics...

First result

Theorem (Lambert 2010)

Under any splitting tree model, there is a random variable $H^T > 0$ *such that, conditional on survival at time* T*, the reconstructed tree seen from* T *is a coalescent point process*
= T *the coalescence times form a sequence of independent r.v., all distributed as* H^T*, stopped at its first value larger than* T*.*

$\Rightarrow$ Notation :

$$F_T(s) := \frac{1}{P(H^T \geq s)}.$$

Coalescent point processes : Popovic (2004), Aldous & Popovic (2005), Lambert & Popovic (2012). See also Gernhard (2008), Stadler (2009).

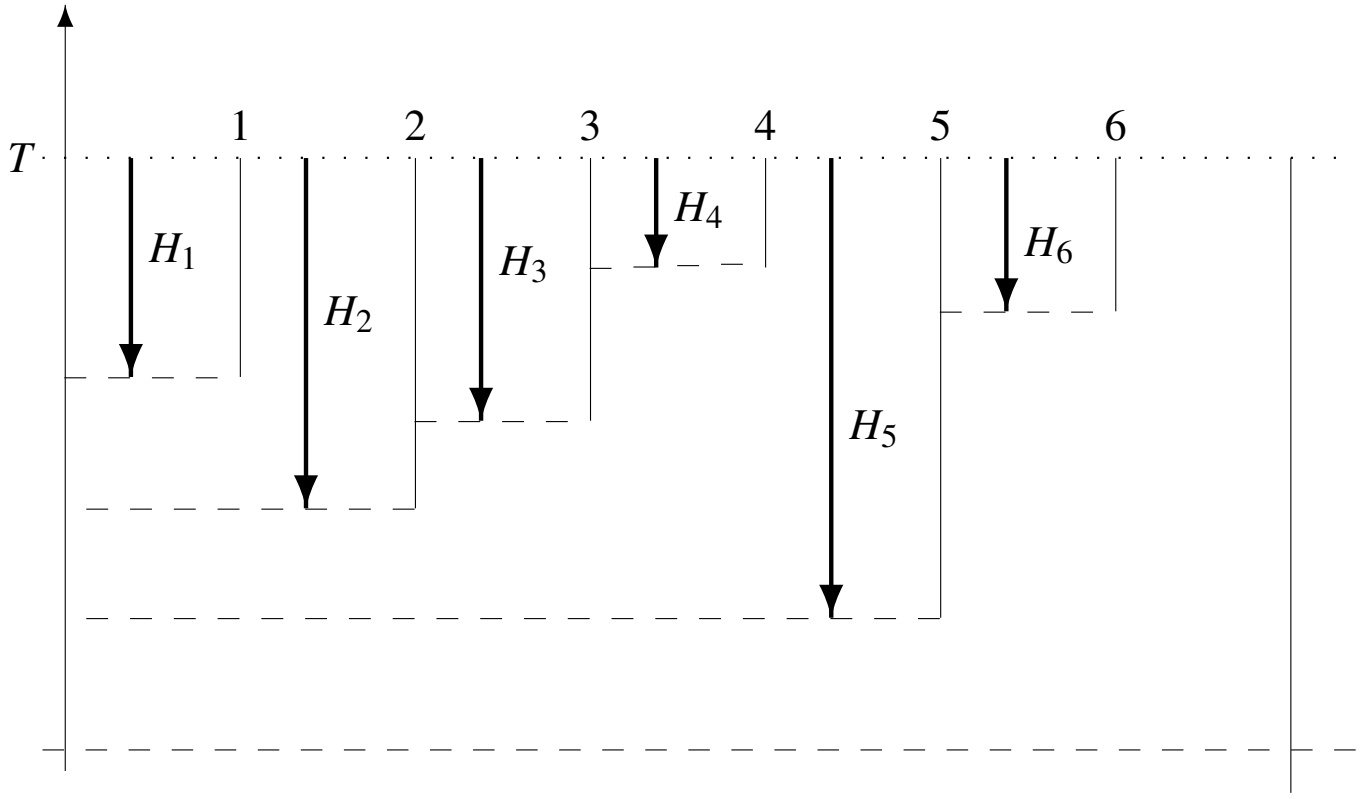
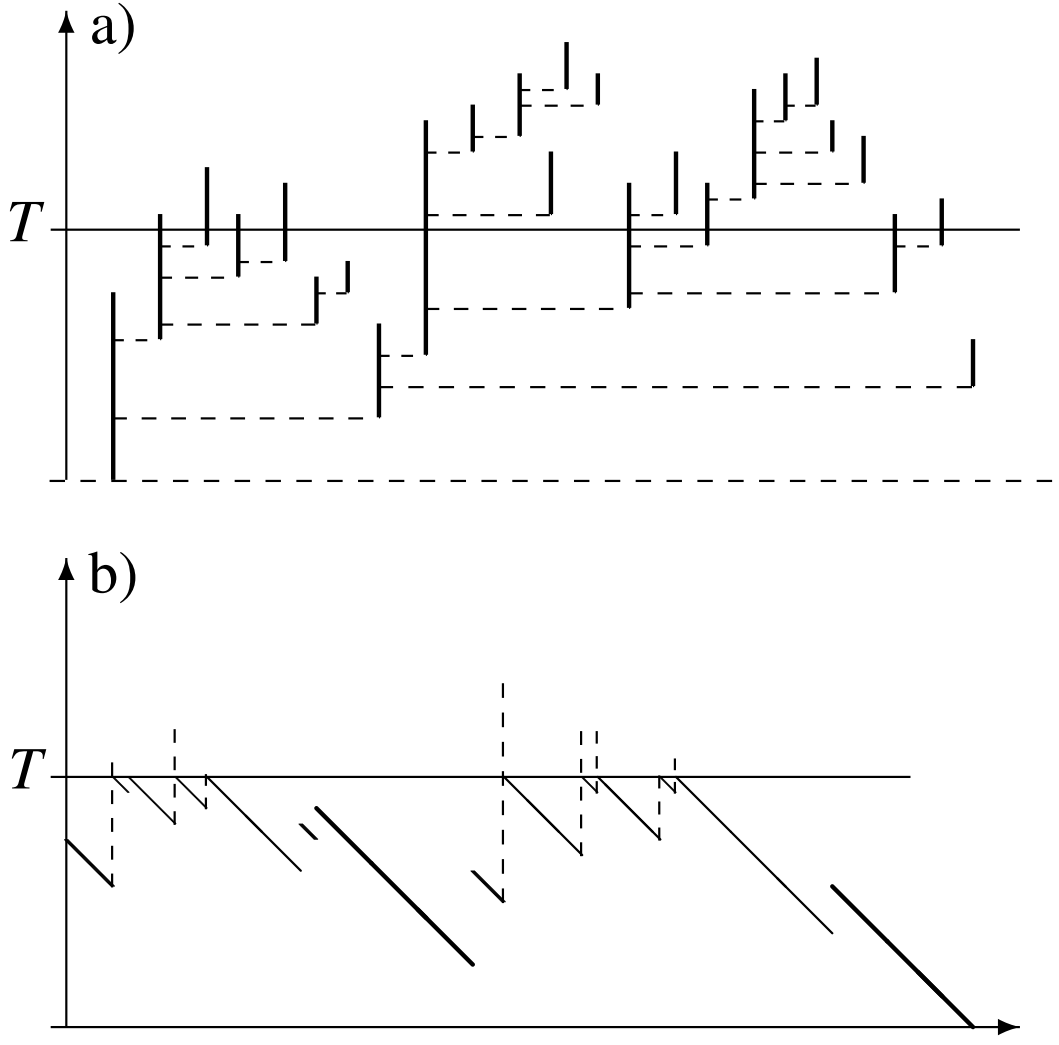


FIGURE: Illustration of a coalescent point process showing the node depths $H_1, \ldots, H_6$ for each of the 6 consecutive pairs of tips. The node depth H_7 is the first one which is larger than T.

Contour of a splitting tree

a) Splitting tree and b) Jumping contour process of its truncation below time T.



Three special cases

1. Time-homogeneous case (Lambert 2010) $\equiv \lambda$ and μ do NOT depend on t ...And then F_T does not depend on T...
2. Markovian case (Nee, May & Harvey 1994) $\equiv \mu(t)$ does NOT depend on age or type

$$F_T(t) = 1 + \int_{T-t}^{T} dx\, \lambda(x)\, e^{\int_x^T dy\, r(y)},$$

where $r(t) := \lambda(t) - \mu(t)$ (instantaneous growth rate).

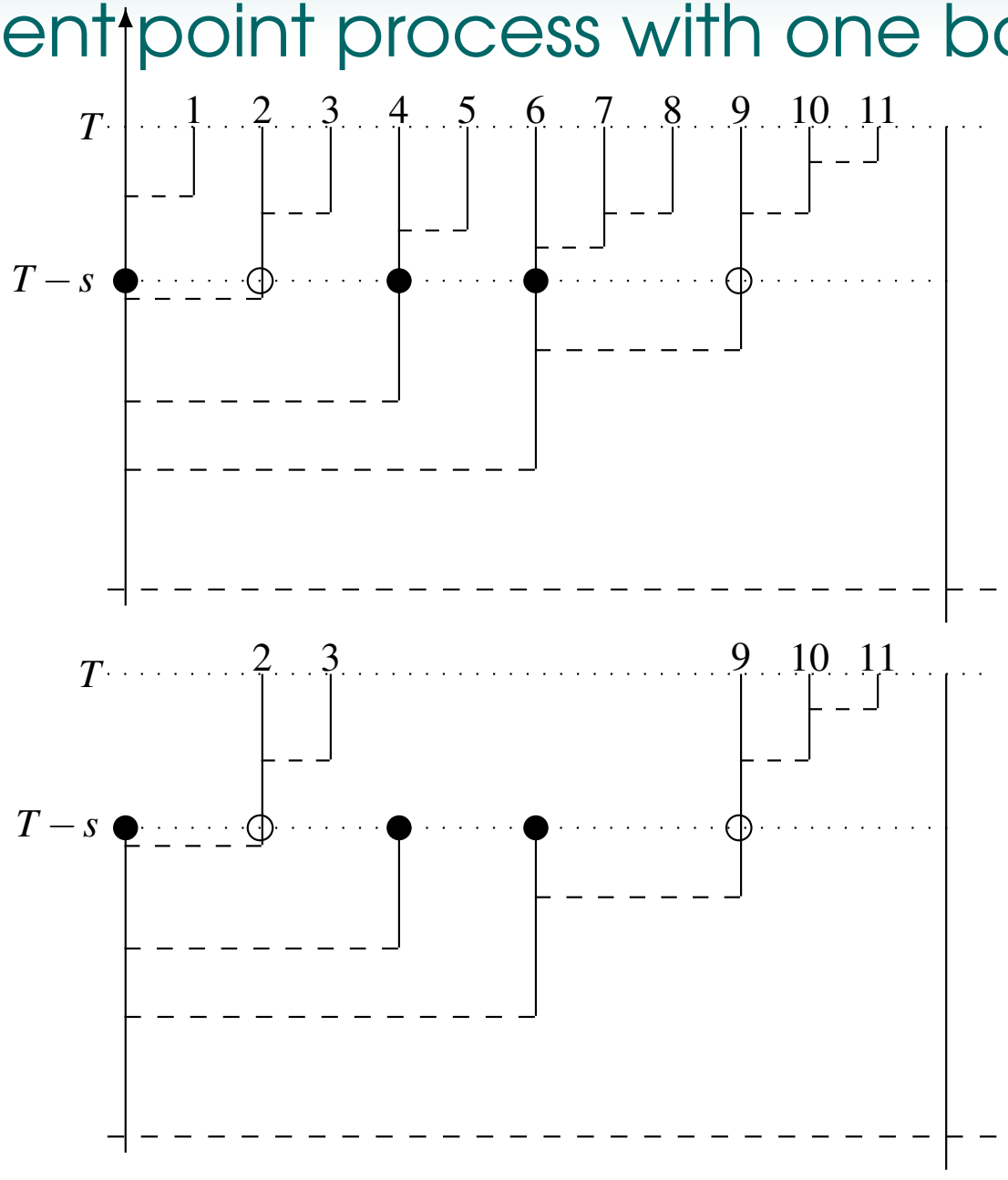
3. Time-homogeneous + Markov (Rannala, 1997) $\equiv \lambda$ and μ are constant $\equiv$ linear birth–death process

$$F_T(t) = 1 + \frac{\lambda}{r}(e^{rt} - 1).$$

Bottleneck : definition

- Start with a coalescent point process
- Add a bottleneck with survival probability ε at time s backwards, i.e., all lineages crossing this time section are independently deleted with probability $1-\varepsilon$
- Special case $s=0$ corresponds to sampling.
- Set $B_\varepsilon^T :=$ coalescence time between two consecutive survivors,

Coalescent point process with one bottleneck



Bottleneck : result

- With probability $P(H^T < s)$, B_ε^T is distributed as H^T conditional on $H^T < s$
- With probability $P(H^T \geq s)$,

$$B_\varepsilon^T \overset{(d)}{=} \max\{A_1, \ldots, A_K\},$$

where the A_i's are i.i.d. distributed as H^T conditional on $H^T \geq s$ and

$$\mathbb{P}(K = j) = \varepsilon(1-\varepsilon)^{j-1}.$$

- This yields

$$F_\varepsilon(t) := \frac{1}{P(B_\varepsilon^T \geq t)} = \begin{cases} F_T(t) & \text{if} \quad t < s \\ \varepsilon F_T(t) + (1-\varepsilon)F_T(s) & \text{if} \quad t \geq s \end{cases}$$

More bottlenecks

Start with a coalescent point process and add extra bottlenecks with survival probabilities $\varepsilon_1, \ldots, \varepsilon_k$ at times $T - s_1 > \ldots > T - s_k$ (where $s_1 \geq 0$ and $s_k < T$).

Proposition (Lambert & Stadler (2012))

Conditional on survival, the new reconstructed tree is again a coalescent point process with inverse tail distribution F_ε given by

$$F_\varepsilon(t) = \varepsilon_1 \cdots \varepsilon_m F_T(t) + \sum_{j=1}^{m} (1 - \varepsilon_j)\, \varepsilon_1 \cdots \varepsilon_{j-1} F_T(s_j) \qquad t \in [s_m, s_{m+1}],$$

for each $m \in \{0, 1, \ldots, k\}$, with $s_0 := 0$ and $s_{k+1} := T$.

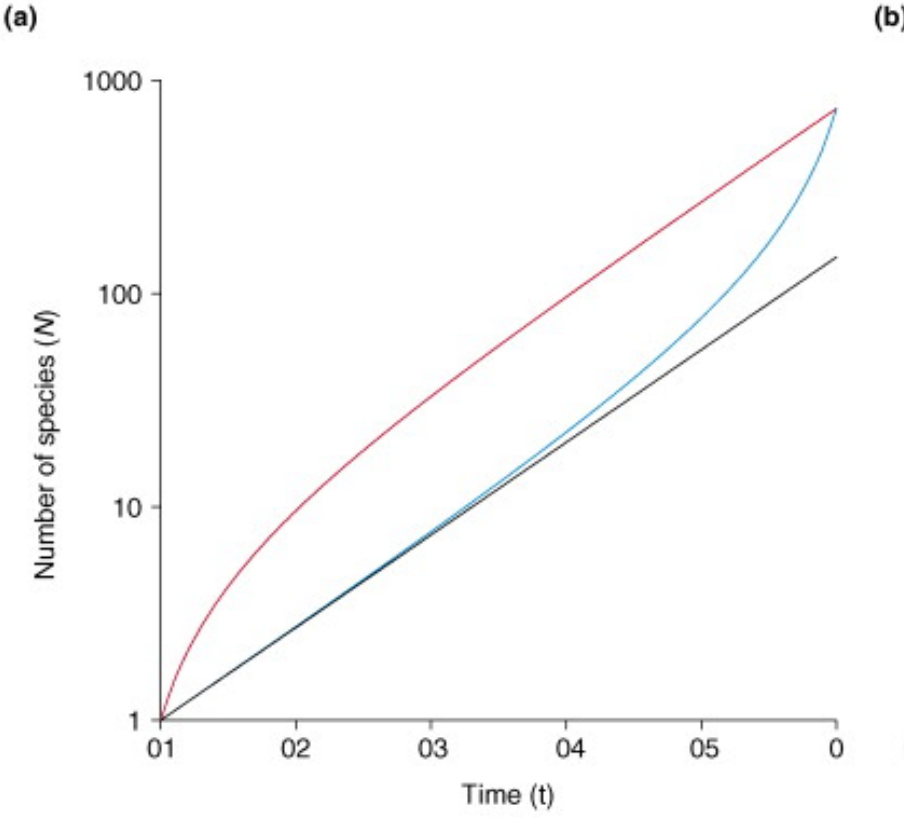
Pull of the present (1)

- Set $N_t^\star :=$ # lineages at time t in the reconstructed tree seen from T
- In our general setting,

$$\mathbb{E}(N_t^\star) = \frac{\mathbb{P}(H^T > T-t)}{\mathbb{P}(H^T > T)} = \frac{be^{rT}-d}{be^{r(T-t)}-d}$$

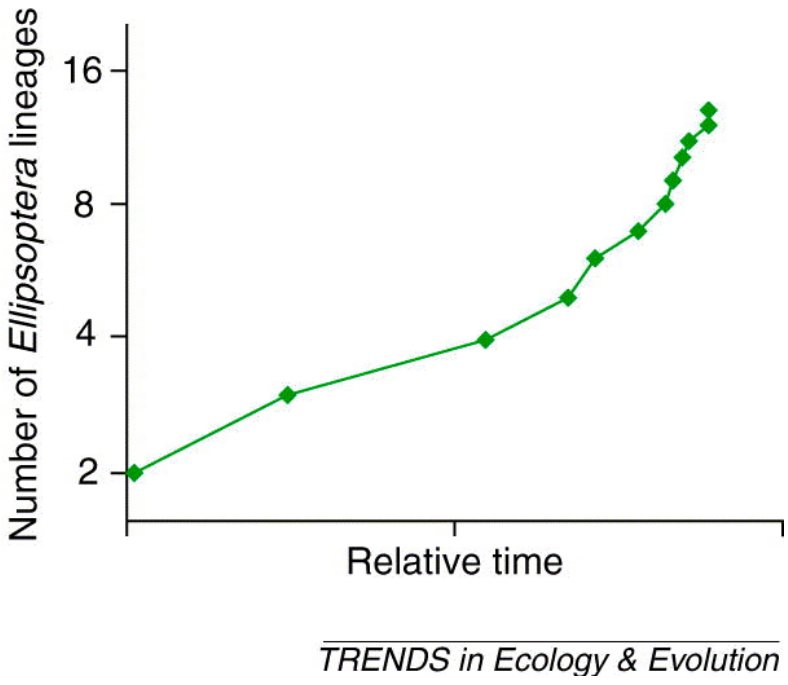
in the case of constant rates

- $\mathbb{E}(N_t \mid N_T \neq 0)$ has a different formula (except if $d=0$)

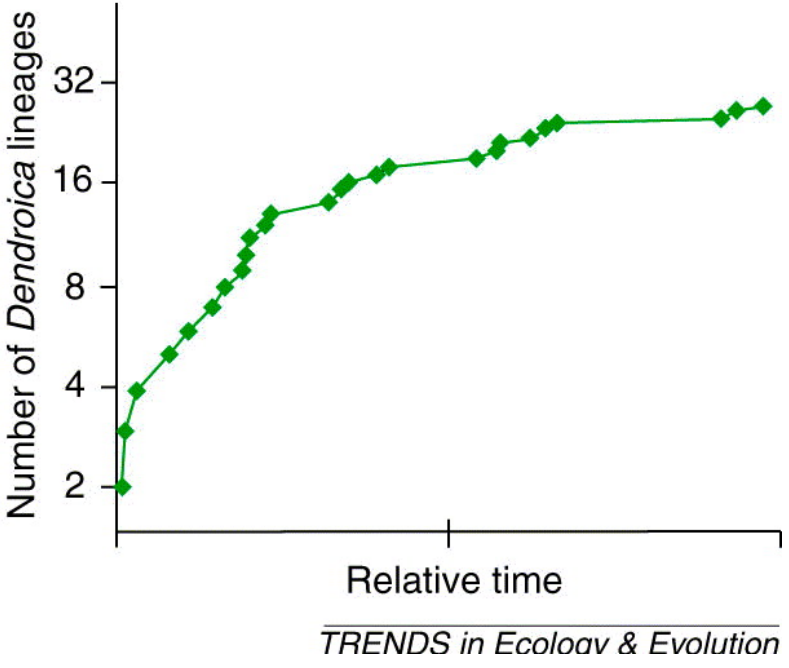


Pull of the present (2)

We sometimes see a recent upturn (pull of the present)...



...But more often see a recent slowdown...



Protracted speciation (with H. Morlon, R.S. Etienne, B. Haegeman)

Model of protracted speciation (Etienne & Rosindell 2011) :

1. Two types of species : New born species are incipient, and turn good after a random time
2. Speciation rate is the same for both species types
3. Extinction rates can be different according to species status
4. Extant incipient species are counted as good never/if they are the youngest descendant species of some extinct species.

$\Rightarrow$ The reconstructed phylogenetic tree of good species is a coalescent point process reproducing recent slowdowns.

Speciation by genetic differentiation (with H. Morlon, M. Manceau)

Model of speciation by genetic differentiation and point mutation (Hubbell 2001) :

1. Individual-based model...
2. ...with Poisson mutations at rate θ on individual lineages
3. Monophyletic definition of species : two individuals are in the same species if their MRCA point is on some unmutated geodesic path between two tips

$\Rightarrow$ The reconstructed phylogenetic tree is NOT a coalescent point process, and (so) reproduces imbalance, and even branch lengths of real phylogenies.

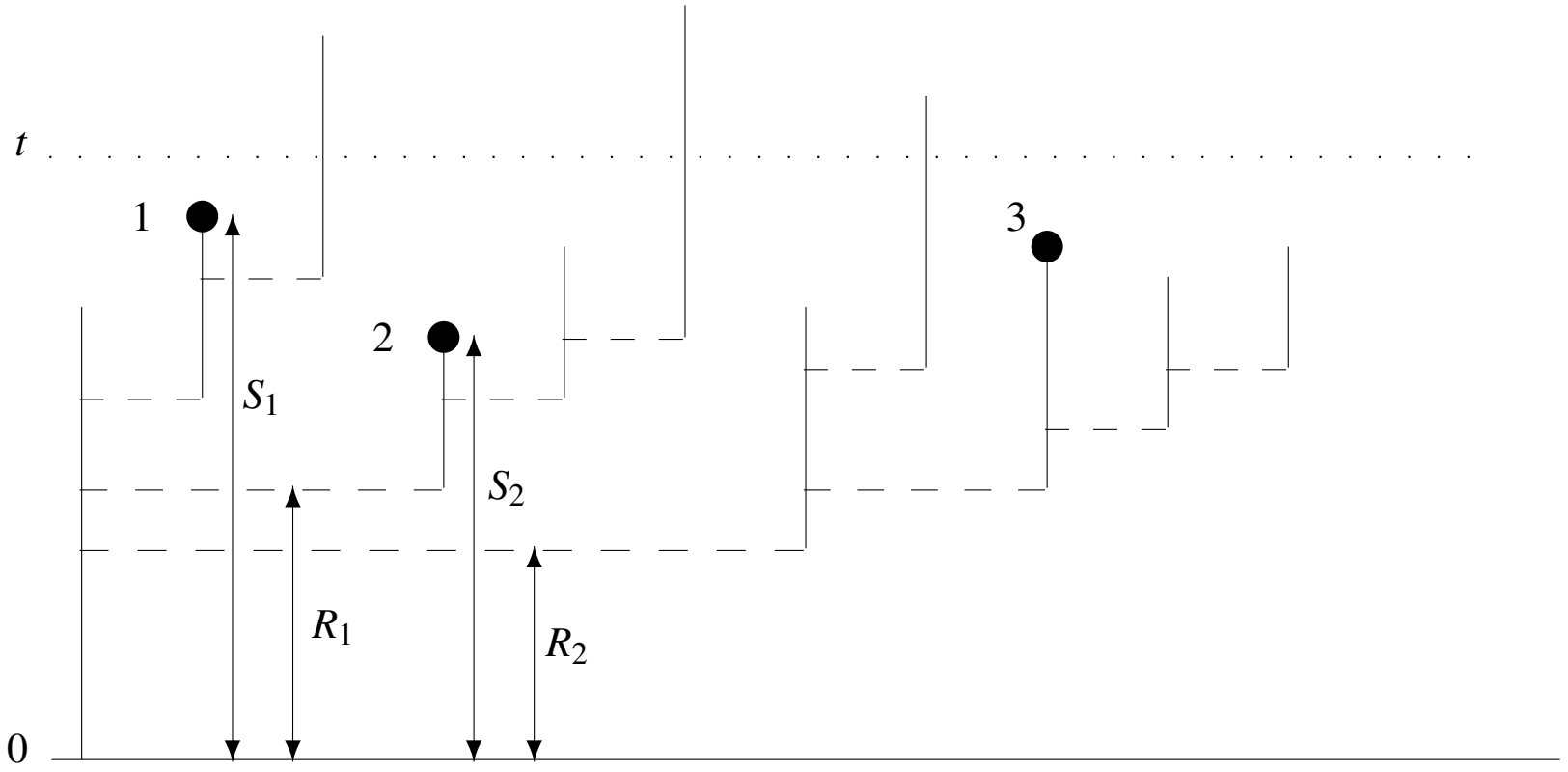
Epidemiology with sampling at constant rate

Epidemic model with sampling through time :

- Epidemics modelled by a splitting tree : constant *per capita* transmission rate, possibly age-dependent death rates, but no density-dependence
- Each infected individual independently, is sampled after an exponential time with parameter δ initialized at birth (birth = transmission)
- A sampled individual immediately leaves the infective population.

A splitting tree with exponential sampling clocks

Black dots = sampled individuals



Temporally-spaced epidemiological data (with Tanja Stadler)

- A sampled individual immediately leaves the infective population.
- $S_i :=$ sampling time of individual i
- $R_i :=$ coalescence time between individuals $i-1$ and i.

By the contour technique, the (S_i, R_i) is a Markov chain with explicit transitions.
$\Rightarrow$ inference of model parameters from viral phylogenies (HIV, flu).

Acknowledgements

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

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

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