A Non-Exchangeable Coalescent Arising in Phylogenetics

Amaury Lambert

LPMA

CIRB

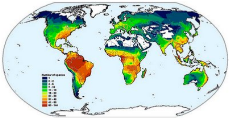
UPMC SORBONNE UNIVERSITÉS

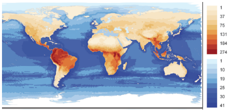
COLLÈGE DE FRANCE 1530

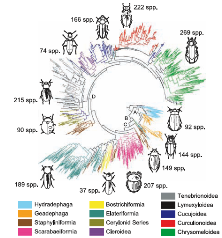
Journées MAS
Grenoble, 30 août 2016

Outline

Pattern & Process

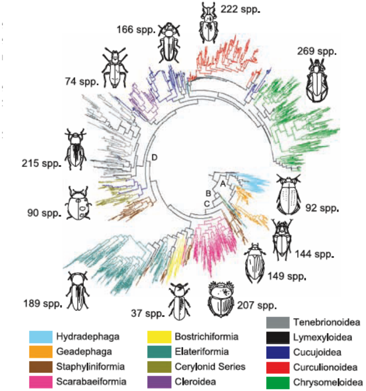


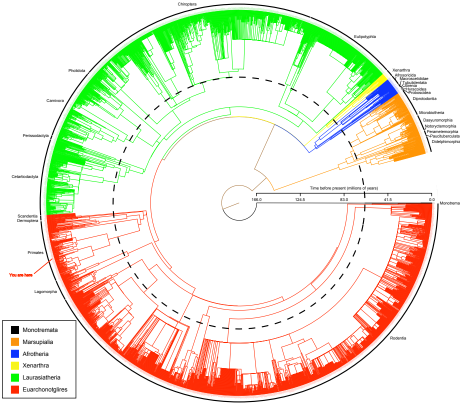




- Design probabilistic models of evolutionary processes...
- ...Generating similar patterns as those observed in nature, and...
- ...Allowing for the inference of these processes from real data...
- ...Assuming the data is a phylogeny (gene tree, species tree,...) already inferred from MSA.

Phylogenetic Trees





Two Questions About Macroevolution

Reconstruct the past of biodiversity : What processes underpin the observed macro-evolutionary patterns ?

- Q1 : "Can we test the possibility that some aspects of the evolutionary record behave as stochastic variables ?" (Raup et al 1973)

  Example of phylogenetic trees = Most basic pattern left by macroevolutionary history

Two Questions About Macroevolution

Reconstruct the past of biodiversity : What processes underpin the observed macro-evolutionary patterns ?

- Q1 : "Can we test the possibility that some aspects of the evolutionary record behave as stochastic variables ?" (Raup et al 1973)

  Example of phylogenetic trees = Most basic pattern left by macroevolutionary history

- Q2 : "Are there mathematically simple or biologically plausible stochastic models for phylogenetic trees whose realizations mimic actual trees ?" (Aldous 2001)

Two Questions About Macroevolution

Reconstruct the past of biodiversity : What processes underpin the observed macro-evolutionary patterns ?

- Q1 : "Can we test the possibility that some aspects of the evolutionary record behave as stochastic variables ?" (Raup et al 1973)

  Example of phylogenetic trees = Most basic pattern left by macroevolutionary history

- Q2 : "Are there mathematically simple or biologically plausible stochastic models for phylogenetic trees whose realizations mimic actual trees ?" (Aldous 2001)

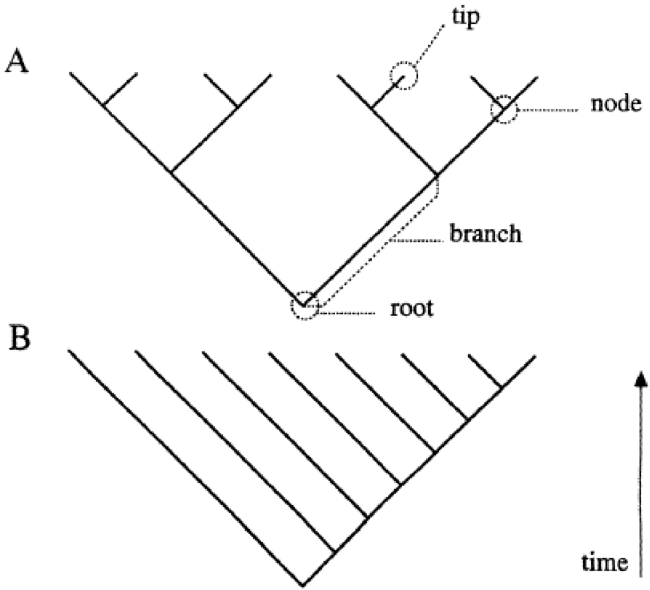
- Alternatively Q2' : Can we infer the most likely evolutionary process to have generated the tree ?

Difficulty of characterizing trees

- Comparing two trees : distance ? Robinson-Foulds, Gromov-Hausdorff, Billera-Holmes-Vogtman...
- Characterizing one tree : distance to some reference tree ?
- A distribution of trees : average tree ?
- Real functions of trees = statistic, likelihood
- Requires stochastic models of trees
  - Compare statistic to its distribution under null model (Q1)
  - Fit a non-null model (Q2)

Outline

Perfectly Balanced Tree (A) vs Caterpillar Tree (B)



Model-free statistics of trees I : Using topology only

See Shao & Sokal (1990), Kirkpatrick & Slatkin (1993), Mooers & Heard (1997)

Local statistics

- $c_i = \#$ nodes on the path from the root to tip i
- $s_{\min}(v) = \#$ tips in smaller daughter clade of node v
- Balance of node $v = s_{\min}(v)/s_{\max}(v)$

Global statistics

- Sackin index (Sackin 1972)

$$\frac{1}{n}\sum_i c_i$$

- Colless index (Colless 1982)

$$\frac{2}{(n-1)(n-2)}\sum_v \left(s_{\max}(v) - s_{\min}(v)\right)$$

Model-free statistics of trees II : Using branch lengths also

Local statistics

- 'Distinctiveness' $=$ length of external edge of tip i (Redding et al 2008)
- Local Branching Index (Luksza & Laessig 2014, Neher et al 2014)

$$= \int_{\text{tree}} e^{-d(x,y)/\delta}\, dy$$

Global statistics

- *Phylogenetic Diversity PD* $=$ Total Length of Tree $= \sum_{k=2}^{n} k g_k$ with $g_k =$ internode duration (Vane-Wright et al 1991, Faith 1992)
- Lineage-Through-Time plot
- Gamma (Cox & Lewis 1966, Pybus & Harvey 2000)

$$\gamma = \frac{\frac{1}{n-2}\sum_{i=2}^{n-1}\sum_{k=2}^{i} k g_k - \frac{PD}{2}}{PD/\sqrt{12(n-2)}}$$

Understanding the origin of patterns

- 1960's : Root imbalance = radiation undergone by the larger daughter subclade ?

Understanding the origin of patterns

- 1960’s : Root imbalance = radiation undergone by the larger daughter subclade ?

- The ‘Woods Hole’ group (Raup, Gould, Schopf, Simberloff) advocates for "a clearer separation of stochastic and deterministic elements in the evolutionary record" (Raup et al 1973)

Understanding the origin of patterns

- 1960's : Root imbalance = radiation undergone by the larger daughter subclade ?

- The 'Woods Hole' group (Raup, Gould, Schopf, Simberloff) advocates for "a clearer separation of stochastic and deterministic elements in the evolutionary record" (Raup et al 1973)

- H_0 : 'pattern is not distinguishable from that generated by a Yule pure birth process'... vs key adaptations, adaptive radiations, etc.

Understanding the origin of patterns

- 1960's : Root imbalance = radiation undergone by the larger daughter subclade ?

- The 'Woods Hole' group (Raup, Gould, Schopf, Simberloff) advocates for "a clearer separation of stochastic and deterministic elements in the evolutionary record" (Raup et al 1973)

- H_0 : 'pattern is not distinguishable from that generated by a Yule pure birth process'... vs key adaptations, adaptive radiations, etc.

- Root balance under the Yule model is uniform !
  "How different, then, is the real world from the stochastic system ? The answer would seem to be 'not very' — the outstanding feature of real and random clades is their basic similarity" (Gould et al 1977, Savage 1983)

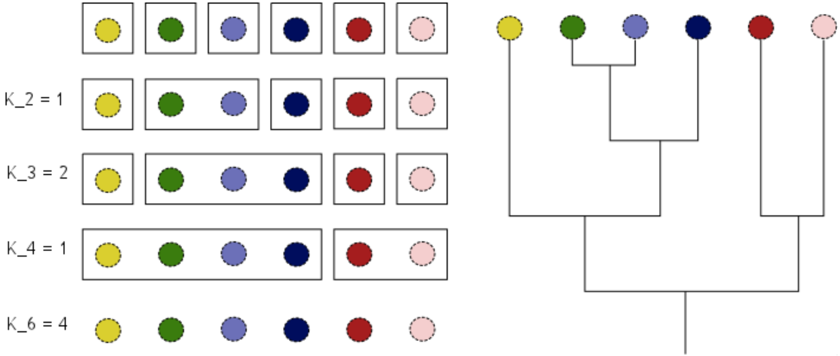
Understanding the origin of patterns

- 1960's : Root imbalance = radiation undergone by the larger daughter subclade ?
- The 'Woods Hole' group (Raup, Gould, Schopf, Simberloff) advocates for "a clearer separation of stochastic and deterministic elements in the evolutionary record" (Raup et al 1973)
- H_0 : 'pattern is not distinguishable from that generated by a Yule pure birth process'... vs key adaptations, adaptive radiations, etc.
- Root balance under the Yule model is uniform !
  "How different, then, is the real world from the stochastic system ? The answer would seem to be 'not very' — the outstanding feature of real and random clades is their basic similarity" (Gould et al 1977, Savage 1983)
- Empirical root balance $\neq$ uniform (Slowinski 1990, Guyer & Slowinski 1991, 93)

Aldous' Markov branching model on binary tree shapes

Aldous (1996, 2001)

- Assume we are given distributions q_n on $\{1, \ldots, n-1\}$, $n \geq 2$
- Recursively split each subset of n balls according to q_n (r.v.'s K_n below)



- q_n uniform yields the same tree shape as a Yule tree

Sampling consistency

- A tree model is a family of probability distributions (P_n) on (exchangeably labelled) tree shapes with n tips
- Call T_n a random tree with law P_n
- Call T'_n the tree obtained by removing one tip from T_{n+1} (say the tip labelled $n+1$)
- The model is said **sampling consistent** if T_n and T'_n have the same distribution.
- Example : Kingman coalescent.

Aldous' Markov branching model

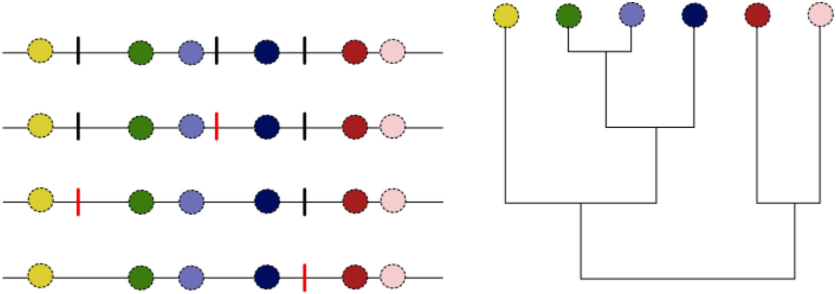
Theorem (Haas et al 2008, Lambert 2016)

A MB tree model is sampling-consistent iff *it there is a function f s.t.*

$$q_n(i) = a_n(f)^{-1} \binom{n}{i} \int_0^1 x^i (1-x)^{n-i} f(x)\, dx$$

Construction

- Color dots are uniformly distributed in the interval
- Intervals are fragmented by r.v. with density $\sim f$



Aldous' Markov β-splitting model on binary tree shapes

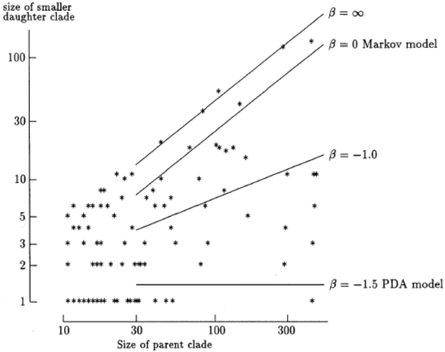
Aldous (1996, 2001), Maliet, Gascuel & Lambert (2016)

- The β-splitting model is for $\beta \in (-2, \infty) : f(x) = cx^{\beta}(1-x)^{\beta}$
- Imbalance decreases with β
- $\beta = 0$ under the Yule model (= Q1)

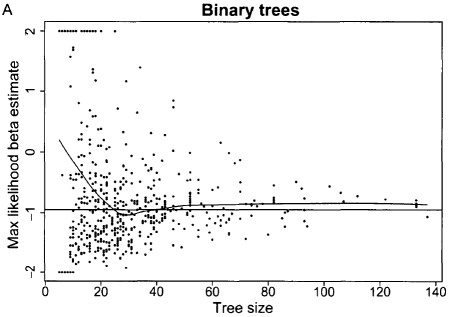
β	Description	Median split
-2	Completely unbalanced	1
-1.5	PDA model	1.5
-1	Unnamed	$\sqrt{m}$
0	Markov model	$m/4$
∞	An almost completely balanced model	$m/2$

Estimating β in real phylogenies

$S_{\min}$ vs $S_{\min} + S_{\max}$ (Aldous 2001)



MLE of β (Blum & François 2006)



$$\Longrightarrow \beta \approx -1$$

Q2 : "Are there mathematically simple/biologically plausible stochastic models for phylogenetic trees whose realizations mimic actual trees ?" (Aldous 2001) or "Why $\beta \approx -1$?"

Outline

1. Introduction

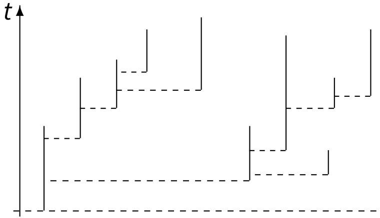
2. Characterizing Trees

3. Lineage-Based Models

4. A Simple Individual-Based Model

Birth-Death Models of Macroevolution (Nee 2006)

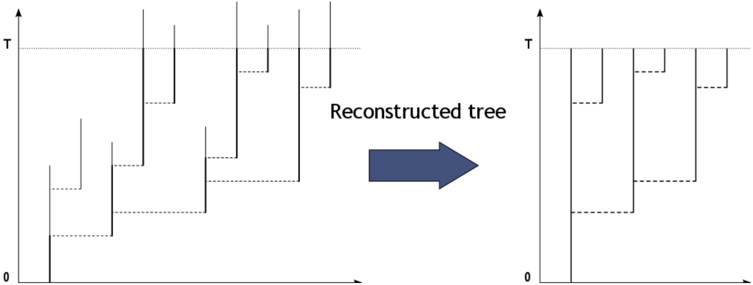
- Species seen as particles that can split (speciation) and die (extinction)
- Rates $b(t, n, a, i)$ and $d(t, n, a, i)$ may depend upon :



- **time** t
- **number** n of standing particles
- a **non-heritable trait** a (e.g., age)
- a **heritable trait** i

Yule model : $b =$ constant, $d = 0$.

Reconstructed Tree



- 'Reconstructed tree' or 'reduced tree' at height T
  = remove all lineages extinct by T (fixed time).

- Q2 : Are there universal conditions on the rates for which the reconstructed tree has $\beta \approx -1$?

- Q2' : What is the law of the reconstructed tree under the model ?
  Can we compute the likelihood of a given ultrametric (clock-like) phylogenetic tree under the model ?

Classifying Lineage-Based Models

Lambert (2010), Lambert & Stadler (2013)

- **A (partial) positive answer to Q2'** :

  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 a **'coalescent point process'** [...]

- **A (partial) negative answer to Q2** :

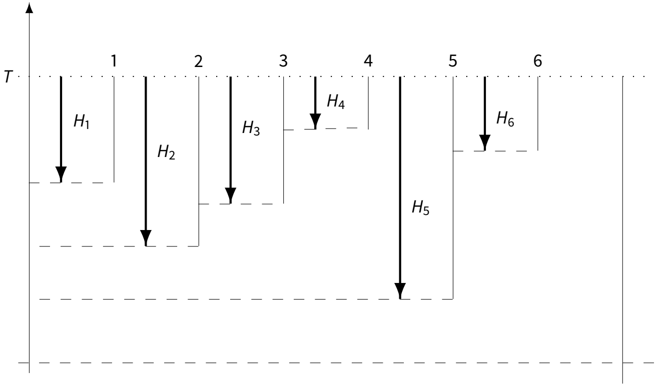
  Reconstructed trees always have the same topology in distribution as Yule trees ($\beta = 0$) IFF $b = b(t, n)$ and $d = d(t, n, a)$

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

The CPP distribution

Rannala (1997), Popovic (2004), Aldous & Popovic (2005)

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** $d = d(t, a)$.
Set $g(t, s)$ the density at time s of the extinction time of a species born at time t.

Theorem (Lambert & Stadler 2013)

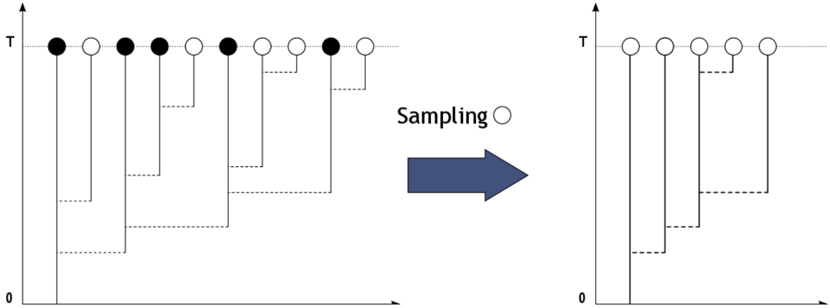
The ***reconstructed (oriented) tree is a CPP*** *with typical node depth H, where the function $F = 1/P(H > \cdot)$ is the unique solution to the following linear integro-differential equation*

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

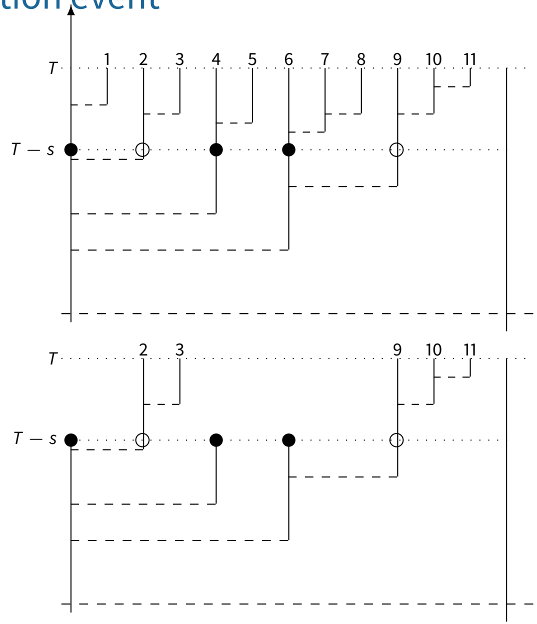
with initial condition $F(0) = 1$.

The result still holds with missing species/mass extinction events : each species is removed independently with the same probability p.

Missing species



Mass extinction event



Special cases

- If $b = b(t)$ and $d = d(t)$ (Kendall 1948, Nee et al 1994)

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

- If b is constant and $d = d(a)$, then $g(s,t) = g(t-s)$ [if a the age $g(a) = d(a)\, e^{-\int_0^a ds\, d(s)}$] (Lambert 2010)

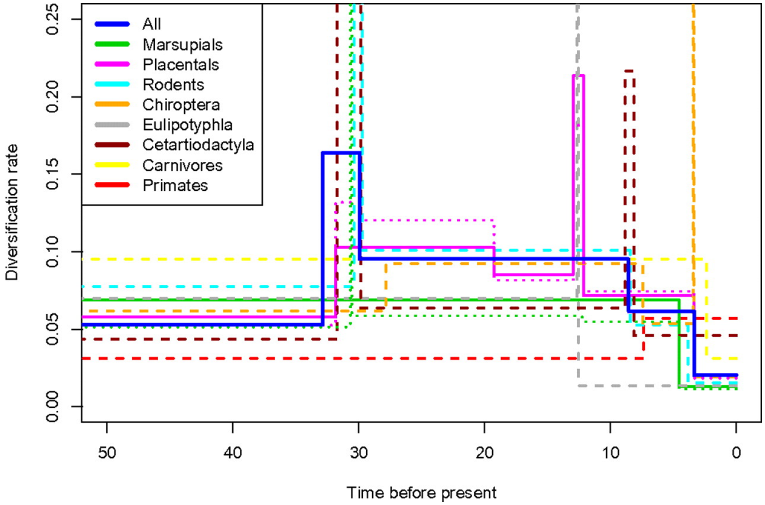
$$F' = b\, \left(F - F \star g\right),$$

- Mass extinction event with survival probability p at time $T - s$

$$F_p(t) = \begin{cases} F(t) & \text{if } 0 \le t \le s \\ (1-p)F(s) + pF(t) & \text{if } s \le t \le T, \end{cases}$$

Appl.1 Diversification of Mammals

Stadler "Mammalian Phylogeny Reveals Recent Diversification Rate Shifts" *PNAS* (2011)



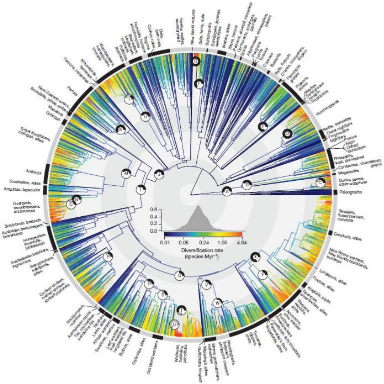
Appl.2 Do species age ?

Alexander, Lambert & Stadler "Quantifying Age-dependent Extinction from Species Phylogenies"
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 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 $b = 0.108\ My^{-1}$

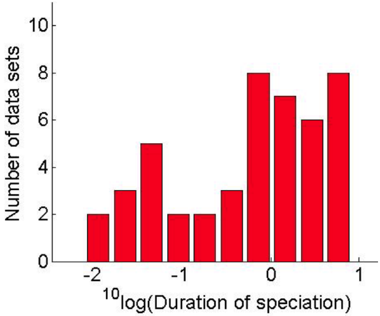


Appl.3 How long does speciation take ?

Etienne, Morlon, Lambert "Estimating the Duration of Speciation from Phylogenies" *Evolution* (2014)

Model of **Protracted Speciation** (Rosindell et al 2010, Etienne & Rosindell 2012)

- Species are ensembles of populations, each population gradually diverges from mother species
- Newborn populations are incipient, become good after some random time = new species
- Speciation stage = non-heritable trait



- Duration of speciation = Time before a good sp appears in the pop genealogy
- Test on simulations : efficient inference of duration of speciation
- Left : duration of speciation inferred in 46 bird clades (in My)

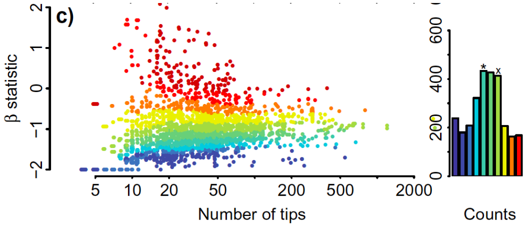
A positive answer to Q2 ?

Hagen, Hartmann, Steel, Stadler "Age-Dependent Speciation Can Explain the Shape of Empirical Phylogenies" *Systematic Biology (2015)*

- $b = b(a)$ parameterized by

$$b(a) = ca^{\phi-1}$$

- Estimates of ϕ lie in $(0, 1)$: speciation rate decreases with age



For $\phi = 0.6$, the reconstructed tree has $\beta \approx -1$.

Q2 : "Why $\beta \approx -1$?"

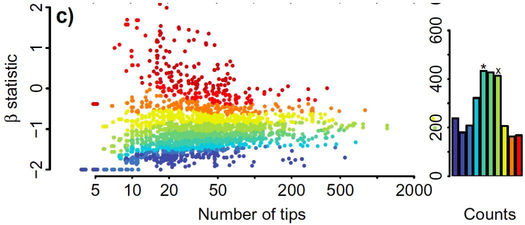
A positive answer to Q2 ?

Hagen, Hartmann, Steel, Stadler "Age-Dependent Speciation Can Explain the Shape of Empirical Phylogenies" *Systematic Biology (2015)*

- $b = b(a)$ parameterized by

$$b(a) = ca^{\phi-1}$$

- Estimates of ϕ lie in $(0, 1)$: speciation rate decreases with age



For $\phi = 0.6$, the reconstructed tree has $\beta \approx -1$.

Q2 : "Why $\beta \approx -1$?"

— "Because $\phi \approx 0.6$" ;-)

Outline

Goal

In this section, 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"
- Key innovations, niche invasions, co-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, Lambert 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)

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

  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 (Champagnat 2006) as $K \to \infty$ each new mutant arises

- 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 $1/K\varepsilon_K$,

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

Statement

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*

$$\rho_n = \lambda \left(1 - \frac{\mu}{\lambda}\right) \sum_{i=1}^{n} \overline{x}_i$$

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* $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_{N_t - 1}, 0, \ldots, 0)$.

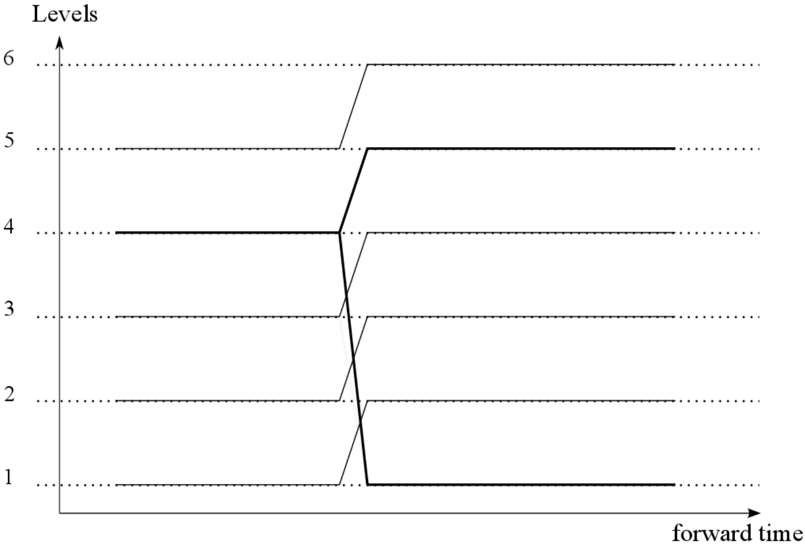
A Non-Exchangeable Coalescent Process

In the new timescale at stationarity, at constant rate

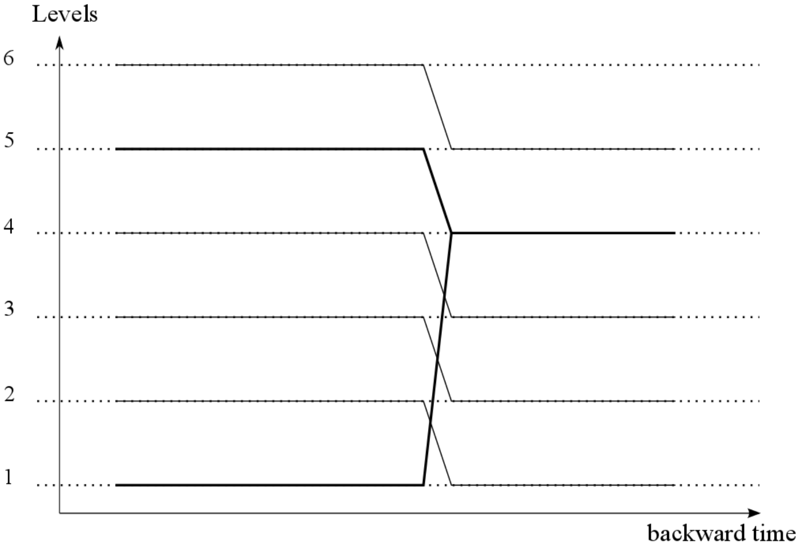
$$\rho = \frac{\kappa}{1-\alpha}\left(1 - \frac{\mu}{\lambda}\right)$$

- 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

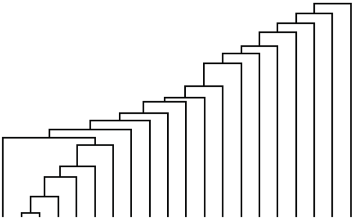
Speciation in forward time...

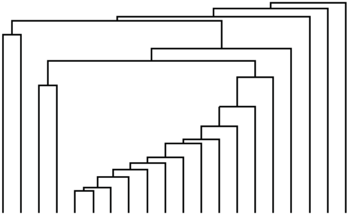


...Coalescence in backward time



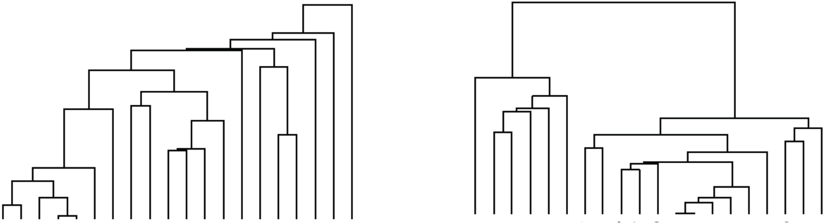
Simulated trees with 20 tips





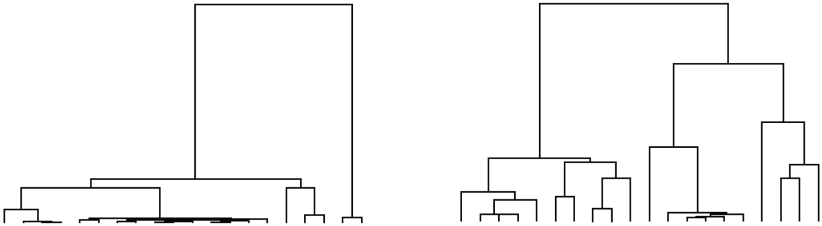
$$\alpha = 0.1$$

Simulated trees with 20 tips



$\alpha = 0.7$

Simulated trees with 20 tips



$\alpha = 0.99$

Intertwining (Rogers & Pitman 1981)

Let $((X_t, Y_t), t \geq 0)$ be 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)

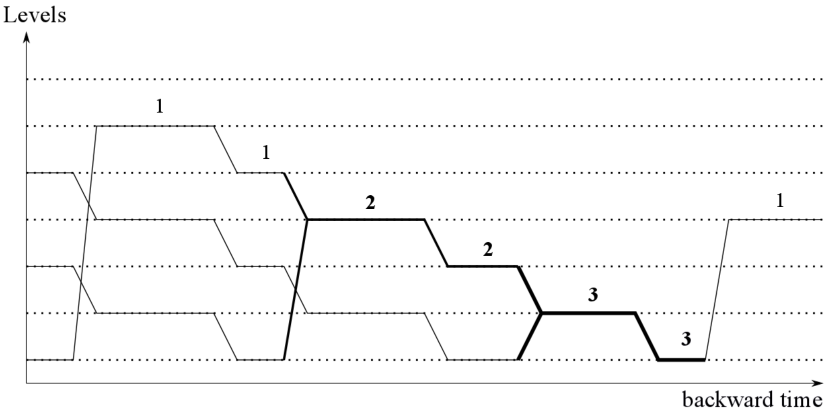
If there exists a generator G of a Markov process in E such that for each $f : E \times F \to \mathbb{R}$ in the domain of $\hat{G}$,

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

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 with generator G.*

The weight measure



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 levels 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.*

Convergence to the Kingman coalescent

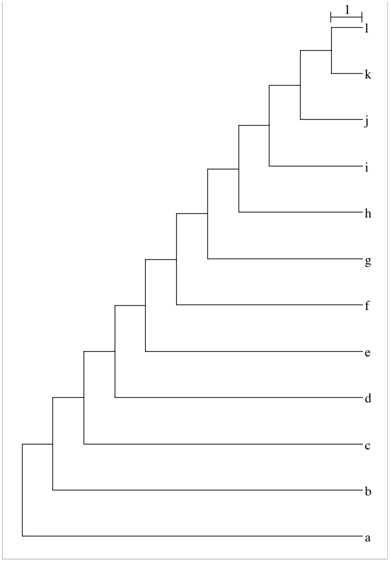
Recall $\alpha = 1 - d/c$ and $\kappa = (\lambda - \mu)/c =$ abundance of youngest species.

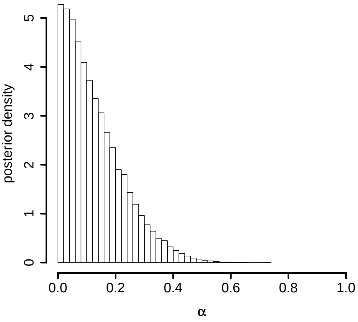
Theorem

As $\alpha \to 1$, the process $(B_{t/(1-\alpha)}; t \geq 0)$ converges (fdd) to $Z_t \delta_1$, where

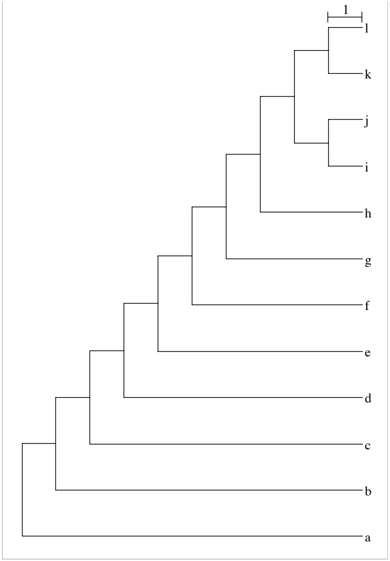
- *$(Z_t; t \geq 0)$ is a pure-death process with death rate $Cn(n-1)/2$*
- *$C = (1 - \mu/\lambda)\kappa$ (replacement rate).*

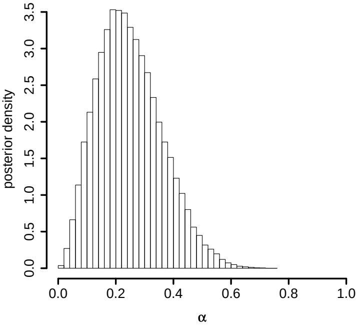
MCMC inference (1) : Caterpillar tree



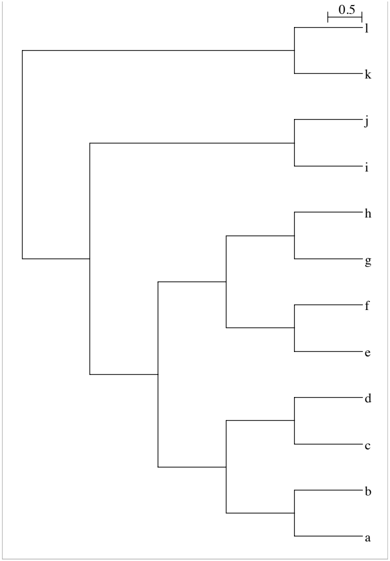


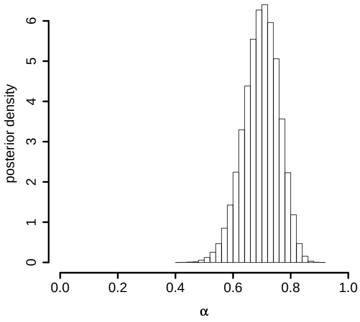
MCMC inference (2) : Very imbalanced tree





MCMC inference (3) : Balanced tree

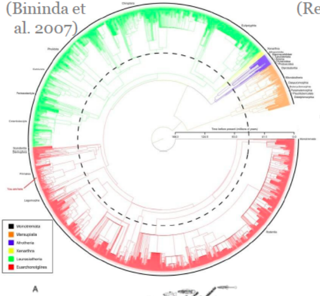




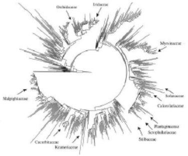
Conclusions

- One-parameter model of phylogeny based on a non-neutral, individual-based model of evolution see also Chisholm & O'Dwyer (2014)
- Relaxing neutrality fails to reproduce universal pattern
- So why $\beta = -1$?
- $\exists$ numerical methods for likelihood computation for general diversification processes
- But all mathematical methods known only work for $\beta = 0$ trees

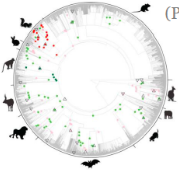
(Bininda et al. 2007)



(Renner & Schaefer 2010)

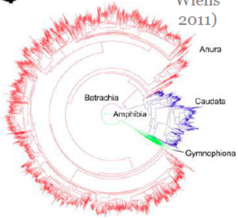


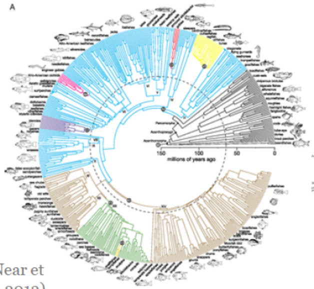
(Purvis et al. 2011)



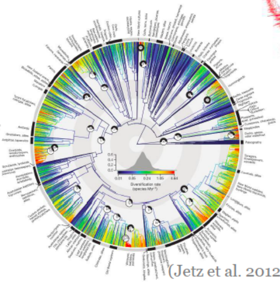
Thank you !

(Pyron & Wiens 2011)

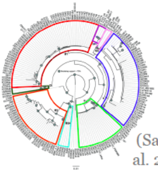




(Near et al. 2013)



(Jetz et al. 2012)



(Sahl et al. 2012)

Collaborators

Guillaume Achaz (UPMC) ....

Helen Alexander (Oxford) ....

Rampal Etienne (U Groningen) ....

Nicolas Lartillot (CNRS) ....

Hélène Morlon (CNRS) ....

Todd L. Parsons (CNRS) ....

Tanja Stadler (ETHZ) ....

SMILE : an interdisciplinary group in Paris

UPMC SORBONNE UNIVERSITÉS

COLLÈGE DE FRANCE 1530

S.M.I.L.E

Stochastic Models for the Inference of Life Evolution

SMILE = Stochastic Models for the Inference of Life Evolution

Other lineage-based models of macro-evolution

- Diversity-dependent diversification (Etienne et al *Proc B* 2012)

- Trait-dependent diversification : BiSSE, QuaSSE, GeoSSE... (Maddison et al *Syst Biol* 2007, FitzJohn *MEE* 2012...)

  But pb with false positives...

- Reviews...
  - Ricklefs *TREE* (2007)
  - Pyron & Burbrink *TREE* (2013)
  - Stadler *JEB* (2013)
  - Morlon *Eco Lett* (2014)

Alternative Answers to Q2

- Phylogenetic reconstruction artifact ? (Huelsenbeck & Kirkpatrick *Evolution* 1996)
- Protracted/age-dependent speciation ? (Rosindell et al *Eco Lett* 2010, Hagen et al *Syst Biol* 2015)
- Neutral Biodiversity Theory ? (Jabot & Chave *Eco Lett* 2009, Davies et al *Evolution* 2012, Manceau, Lambert & Morlon *Eco Lett* 2015)