Three Lectures about :
"Evolutionary Processes and Patterns of Biodiversity"

Lecture 2/3 : The formation of new species

Amaury Lambert

amaury_upmc

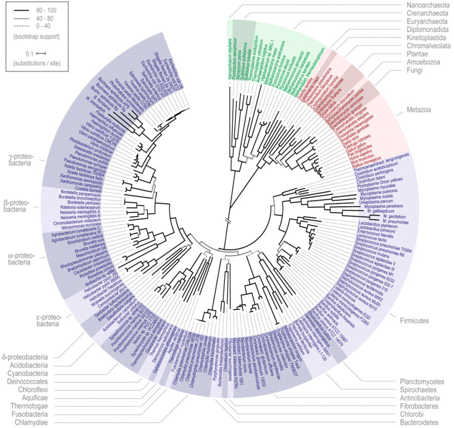
IICD & Probability and Society Initiative Joint Seminar Series
Columbia University
October 16, 2020

Outline

Introduction : speciation, reproductive isolation, phylogeny

Dobzhansky, Mayr...

- In systematics, species are groups of individuals which

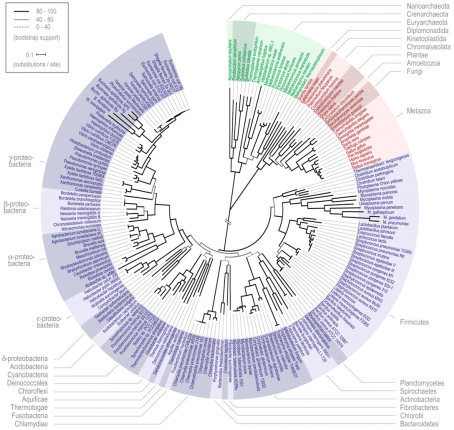


A subtree of the Tree of Life
Ciccarelli et al *Science* 2006

Introduction : speciation, reproductive isolation, phylogeny

Dobzhansky, Mayr...

- In systematics, species are groups of individuals which
  - can interbreed/mate/hybridize

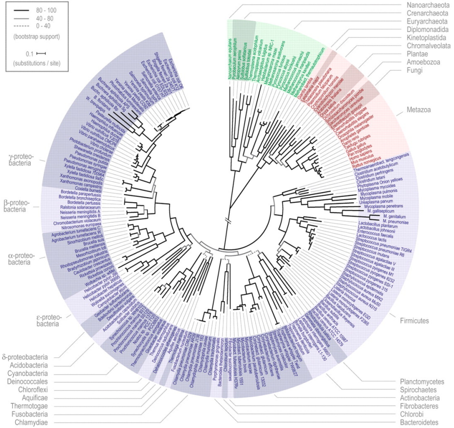


A subtree of the Tree of Life
Ciccarelli et al *Science* 2006

Introduction : speciation, reproductive isolation, phylogeny

Dobzhansky, Mayr...

- In systematics, species are groups of individuals which
  - can interbreed/mate/hybridize
  - cannot interbreed outside the group = are reproductively isolated (RI) from other such groups

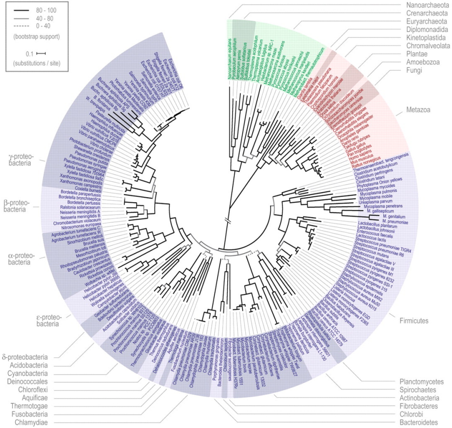


A subtree of the Tree of Life
Ciccarelli et al *Science* 2006

Introduction : speciation, reproductive isolation, phylogeny

Dobzhansky, Mayr...

- In systematics, species are groups of individuals which
  - can interbreed/mate/hybridize
  - cannot interbreed outside the group $=$ are reproductively isolated (RI) from other such groups
- **Q. How do new groups emerge from old groups ?**

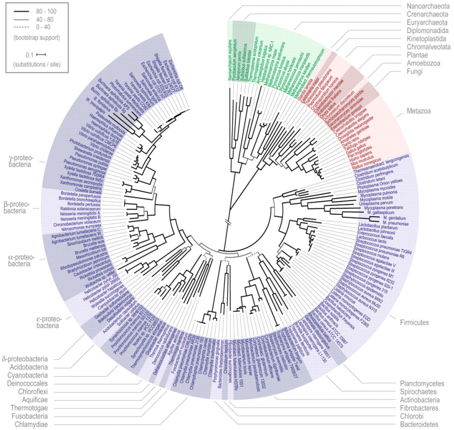


A subtree of the Tree of Life
Ciccarelli et al *Science* 2006

Introduction : speciation, reproductive isolation, phylogeny

Dobzhansky, Mayr...

- In systematics, species are groups of individuals which
    - can interbreed/mate/hybridize
    - cannot interbreed outside the group $=$ are reproductively isolated (RI) from other such groups
- **Q. How do new groups emerge from old groups ?**
- RI generates diversity

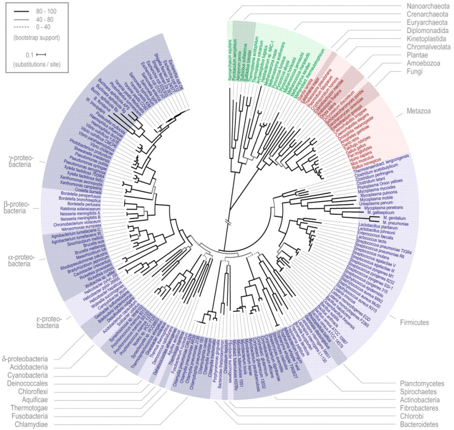


A subtree of the Tree of Life
Ciccarelli et al *Science* 2006

Introduction : speciation, reproductive isolation, phylogeny

Dobzhansky, Mayr...

- In systematics, species are groups of individuals which
  - can interbreed/mate/hybridize
  - cannot interbreed outside the group = are reproductively isolated (RI) from other such groups
- **Q. How do new groups emerge from old groups ?**
- RI generates diversity
- Phylogeny = Genealogy of species

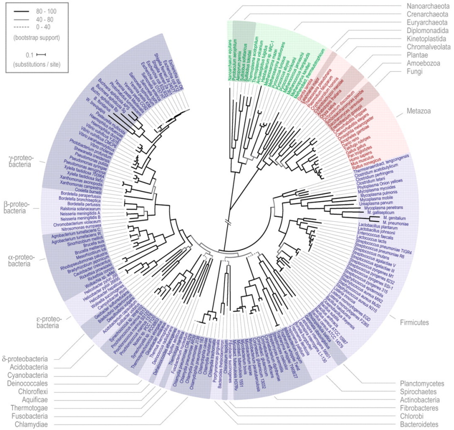


A subtree of the Tree of Life
Ciccarelli et al *Science* 2006

Introduction : speciation, reproductive isolation, phylogeny

Dobzhansky, Mayr...

- In systematics, species are groups of individuals which
  - can interbreed/mate/hybridize
  - cannot interbreed outside the group $=$ are reproductively isolated (RI) from other such groups
- **Q. How do new groups emerge from old groups ?**
- RI generates diversity
- Phylogeny $=$ Genealogy of species
  - Tree shape, edge lengths

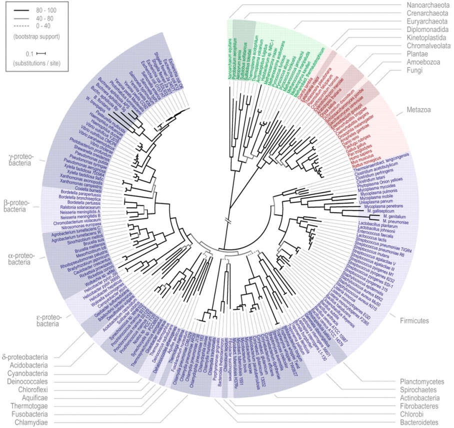


A subtree of the Tree of Life
Ciccarelli et al *Science* 2006

Introduction : speciation, reproductive isolation, phylogeny

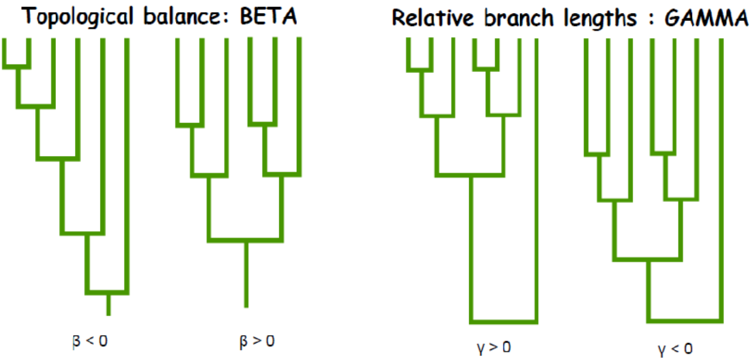
Dobzhansky, Mayr...

- In systematics, species are groups of individuals which
  - can interbreed/mate/hybridize
  - cannot interbreed outside the group = are reproductively isolated (RI) from other such groups
- **Q. How do new groups emerge from old groups ?**
- RI generates diversity
- Phylogeny = Genealogy of species
  - Tree shape, edge lengths
  - Can we learn from the phylogeny about the diversification process ?



A subtree of the Tree of Life
Ciccarelli et al *Science* 2006

Two popular examples of observable statistics



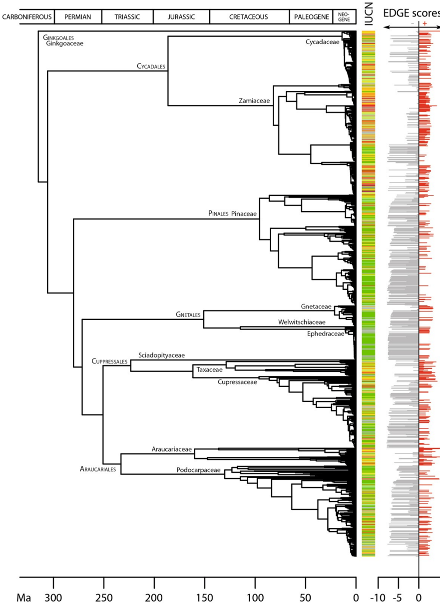
Picture by Marc Manceau

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

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

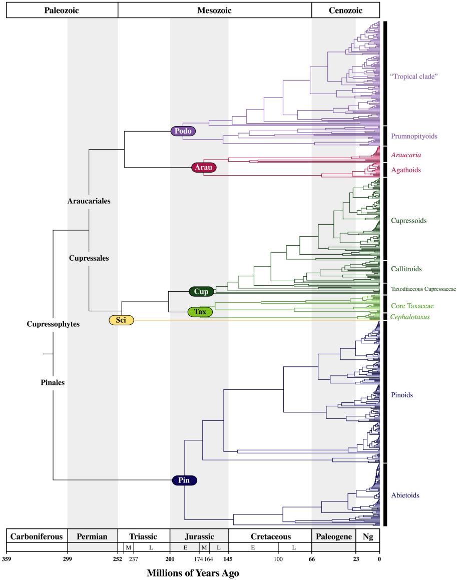
Phylogenetic tree of Gymnosperms

Ginkgoaceae



Forest et al *Sci Reports* 2018

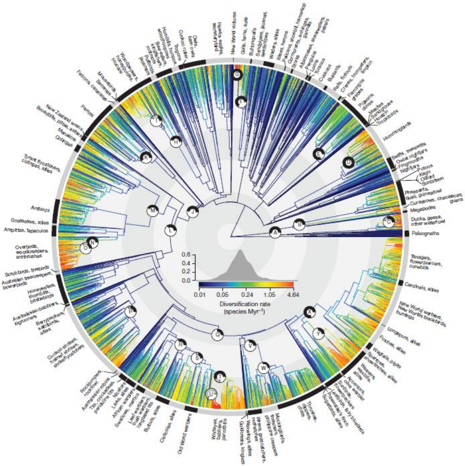
Phylogenetic tree of Conifers



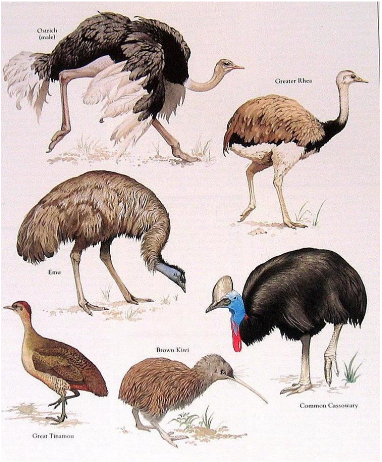
Leslie et al *Am J Botany* 2018

Sciadopityaceae

Phylogenetic tree of Birds

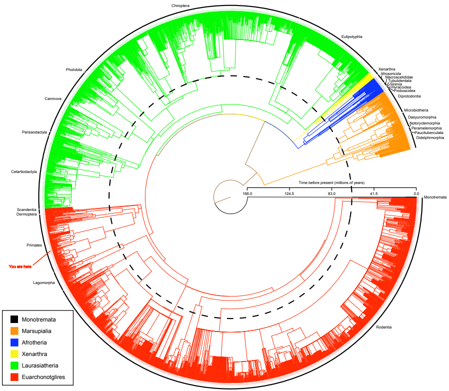


Jetz et al *Nature* 2012



Paleognathae

Phylogenetic tree of Mammals



Bininda-Emonds et al *Nature* 2007

Monotremata

Outline

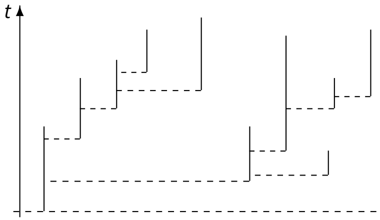
- Lineage-based models of diversification, coalescent point process
- Three works inferring diversification from phylogenies
- Three bottom-up models of speciation, progressive emergence of RI
- Applications
  - **Q1.** How does the interbreeding graph look like ?
  - **Q2.** (Why) are empirical phylogenies imbalanced ?

Outline

Birth-death model of diversification

Stanley, Savage, Raup, Simberloff, Gould, Nee, May...

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

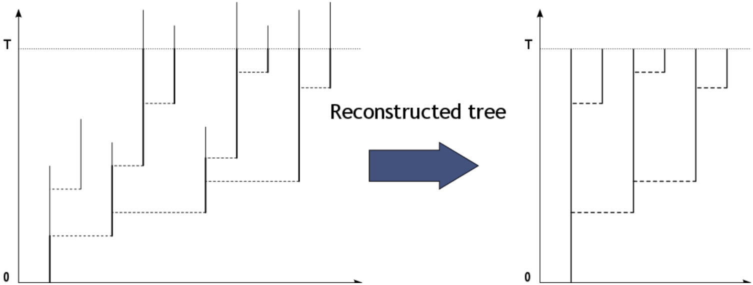


- **time** t
- **number** n of co-occurring particles
- a **non-heritable trait** a (e.g., age)
- a **heritable trait** i (e.g., body mass)
- Orientation =
  Daughter sprouts to the right

- Yule model : $b =$ constant, $d = 0$.
- No information on the process of speciation, but
- Plainly generates a phylogeny

Reconstructed tree

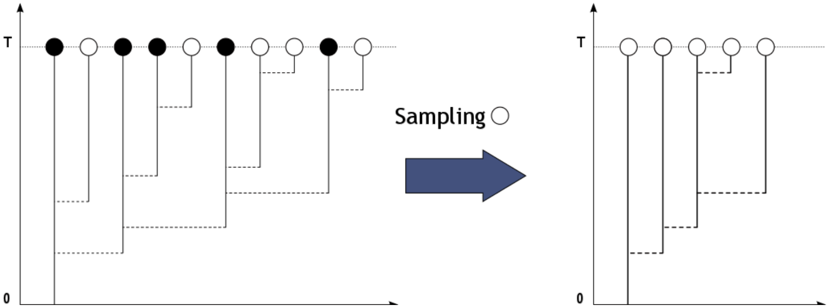
Nee, May & Harvey (1994), Lambert & Stadler (2013)...



- Q : What is the law of the reconstructed tree under the model ?
- 'Reconstructed tree' or 'reduced tree' at time T
  $=$ Tree spanned by species extant at T

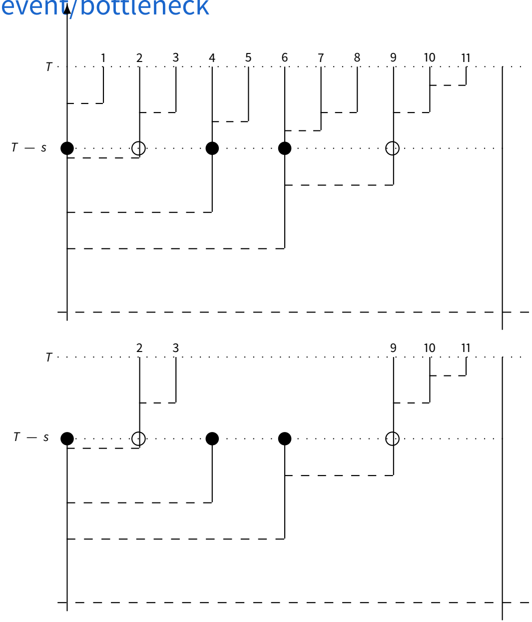
  ...or possibly by a sample of these extant species

Missing species



Each species is removed independently with the same probability.

Mass extinction event/bottleneck



Classifying lineage-based models

Lambert "The contour of splitting trees is a Lévy process" *Ann Probab* (2010)
Lambert & Stadler "Birth-death models and coalescent point processes : The shape and probability of reconstructed phylogenies" *TPB* (2013)

Proposition (Lambert 2010, Lambert & Stadler 2013)

Under the birth-death model with $b = b(t, n, a, i)$ and $d = d(t, n, a, i)$,

1. *Tree shape only. The reconstructed tree always has the same topology in distribution as the pure-birth Yule tree ($b =$ constant, $d = 0$), IFF $b = b(t, n)$ and $d = d(t, n, a)$.*

2. *Tree shape + edge lengths. The likelihood of the reconstructed tree always has an explicit product form IFF $b = b(t)$ and $d = d(t, a)$.*

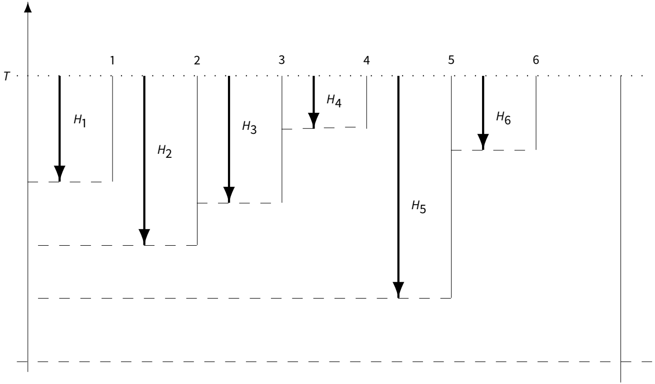
$\Longrightarrow$ *The reconstructed tree is a so-called* **coalescent point process...**

The Coalescent Point Process

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

Assume you are given the law of some random variable $H > 0$.

Coalescent Point Process (CPP) $=$ Oriented tree whose node depths $H_1, H_2, \ldots,$ form a sequence of **independent copies of** H **killed** at its first value larger than T.



- Super **fast simulation** of reconstructed tree
- Likelihood of reconstructed tree in explicit product form $\Rightarrow$ **Simple, efficient inference**

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

Lambert & Stadler "Birth-death models and coalescent point processes : The shape and probability of reconstructed phylogenies" *TPB* (2013)

Theorem (Lambert & Stadler 2013)

If $b = b(t)$ and $d = d(t, a)$, where t is time and a is any non-heritable trait, then the reconstructed (oriented) tree is a CPP with typical node depth H, where the function

$$F(t) := 1/P(H > t)$$

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 $F(0) = 1$, where $g(t, s) =$ density at time s of the extinction time of a species born at time t.

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

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

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

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

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

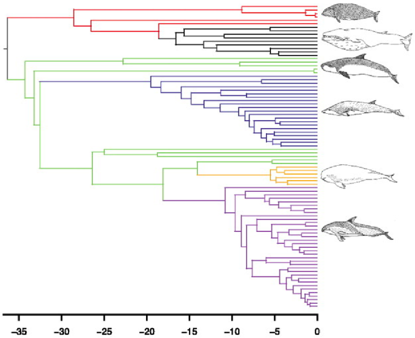
- Mass extinction event with survival probability p at time $T - s$

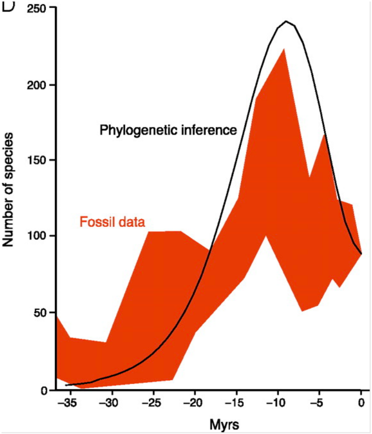
$$F_p(t) = \begin{cases} F(t) & \text{if } 0 \leq t \leq s \\ (1-p)F(s) + pF(t) & \text{if } s \leq t \leq T, \end{cases}$$

Outline

Appl.1 Diversification of Cetaceans : $b = b(t), d = d(t)$

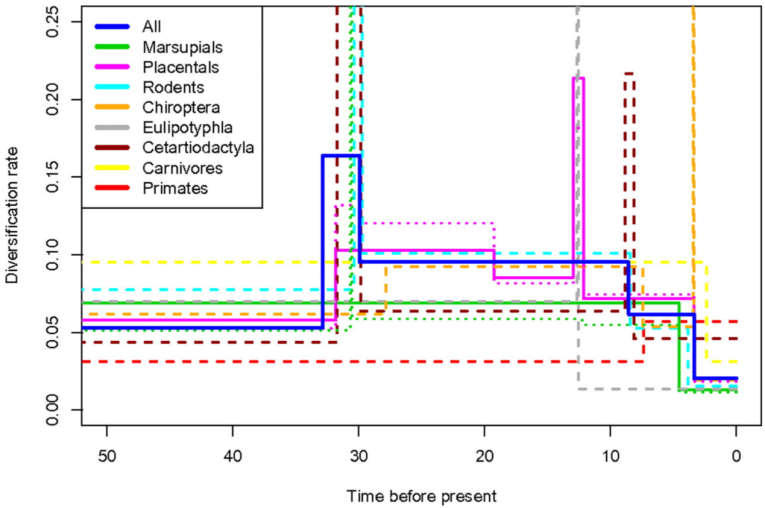
Morlon, Parsons & Plotkin "Reconciling molecular phylogenies with the fossil record" *PNAS* (2011)





Appl.2 Diversification of Mammals : $b = b(t), d = d(t)$

Stadler "Mammalian phylogeny reveals recent diversification rate shifts" *PNAS* (2011)



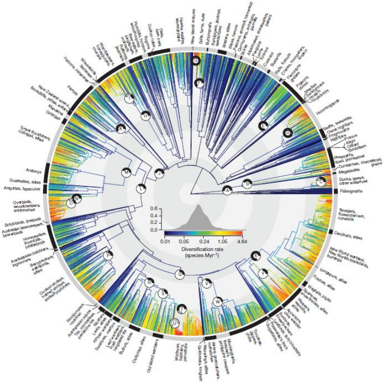
Appl.3 Do species age? b = constant, $d = d(a)$

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, k and s
- MLE on bird 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}



What’s next ?

Open the species box !

- Lineage-based models of macro-evolution

What's next ?

Open the species box !

- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)

What's next ?

Open the species box !

- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)
  - Age-dependent div (Alexander et al *Syst Biol* 2015, Hagen et al *Syst Biol* 2015)

What's next ?

Open the species box !

- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)
  - Age-dependent div (Alexander et al *Syst Biol* 2015, Hagen et al *Syst Biol* 2015)
  - Diversity-dependent diversification (Etienne et al *Proc B* 2012)

What's next ?

Open the species box !

- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)
  - Age-dependent div (Alexander et al *Syst Biol* 2015, Hagen et al *Syst Biol* 2015)
  - Diversity-dependent diversification (Etienne et al *Proc B* 2012)
  - Trait-dependent diversification : BiSSE, QuaSSE, HiSSE...
    (Maddison et al *Syst Biol* 2007, FitzJohn *MEE* 2012, Beaulieu & O'Meara *Syst Biol* 2016...)

What’s next ?

Open the species box !

- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)
  - Age-dependent div (Alexander et al *Syst Biol* 2015, Hagen et al *Syst Biol* 2015)
  - Diversity-dependent diversification (Etienne et al *Proc B* 2012)
  - Trait-dependent diversification : BiSSE, QuaSSE, HiSSE...
    (Maddison et al *Syst Biol* 2007, FitzJohn *MEE* 2012, Beaulieu & O’Meara *Syst Biol* 2016...)
- Four survey articles ! ! (Ricklefs *TREE* 2007, Pyron & Burbrink *TREE* 2013, Stadler *JEB* 2013, Morlon *Eco Lett* 2014)

Review TRENDS in Ecology and Evolution Vol.22 No.11

Estimating diversification rates from phylogenetic information

Robert E. Ricklefs

Review

Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses

R. Alexander Pyron[1] and Frank T. Burbrink[2,3]

JOURNAL OF Evolutionary Biology
doi: 10.1111/jeb.12139

REVIEW

Recovering speciation and extinction dynamics based on phylogenies

T. STADLER

ECOLOGY LETTERS
Ecology Letters, (2014) 17: 508–525 doi: 10.1111/ele.12251

REVIEW AND SYNTHESIS

Phylogenetic approaches for studying diversification

Hélène Morlon*

Abstract
Estimating rates of speciation and extinction, and understanding how and why they vary over

What's next ?

Open the species box !

- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)
  - Age-dependent div (Alexander et al *Syst Biol* 2015, Hagen et al *Syst Biol* 2015)
  - Diversity-dependent diversification (Etienne et al *Proc B* 2012)
  - Trait-dependent diversification : BiSSE, QuaSSE, HiSSE... (Maddison et al *Syst Biol* 2007, FitzJohn *MEE* 2012, Beaulieu & O'Meara *Syst Biol* 2016...)
- Four survey articles ! ! (Ricklefs *TREE* 2007, Pyron & Burbrink *TREE* 2013, Stadler *JEB* 2013, Morlon *Eco Lett* 2014)
- Species $\neq$ particles : lineage-based models do not inform us about the process of speciation

Review TRENDS in Ecology and Evolution Vol.22 No.11
Full text provided by www.sciencedirect.com ScienceDirect

Estimating diversification rates from phylogenetic information

Robert E. Ricklefs

Review Cell PRESS

Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses

R. Alexander Pyron[1] and Frank T. Burbrink[2,3]

JOURNAL OF Evolutionary Biology eseb
doi: 10.1111/jeb.12139
REVIEW

Recovering speciation and extinction dynamics based on phylogenies

T. STADLER

ECOLOGY LETTERS
Ecology Letters, (2014) 17: 508–525 doi: 10.1111/ele.12251
REVIEW AND SYNTHESIS

Phylogenetic approaches for studying diversification

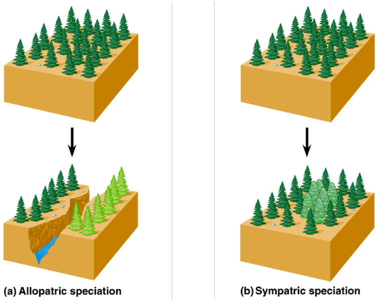
Hélène Morlon*

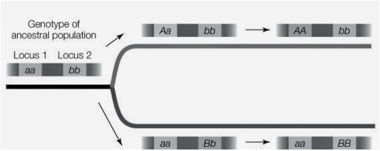
Abstract
Estimating rates of speciation and extinction, and understanding how and why they vary over

What's next ?

Open the species box !

- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)
  - Age-dependent div (Alexander et al *Syst Biol* 2015, Hagen et al *Syst Biol* 2015)
  - Diversity-dependent diversification (Etienne et al *Proc B* 2012)
  - Trait-dependent diversification : BiSSE, QuaSSE, HiSSE... (Maddison et al *Syst Biol* 2007, FitzJohn *MEE* 2012, Beaulieu & O'Meara *Syst Biol* 2016...)
- Four survey articles ! ! (Ricklefs *TREE* 2007, Pyron & Burbrink *TREE* 2013, Stadler *JEB* 2013, Morlon *Eco Lett* 2014)
- Species $\neq$ particles : lineage-based models do not inform us about the process of speciation
- Progressive emergence of reproductive isolation (RI) is ubiquitous

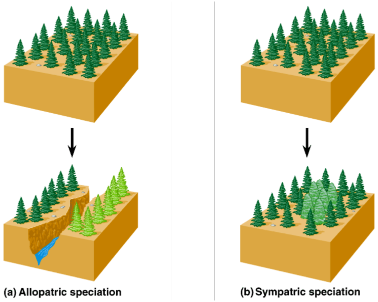




What's next ?

Open the species box !

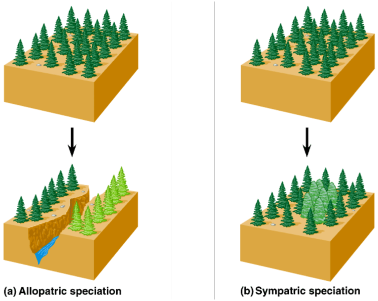
- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)
  - Age-dependent div (Alexander et al *Syst Biol* 2015, Hagen et al *Syst Biol* 2015)
  - Diversity-dependent diversification (Etienne et al *Proc B* 2012)
  - Trait-dependent diversification : BiSSE, QuaSSE, HiSSE...
    (Maddison et al *Syst Biol* 2007, FitzJohn *MEE* 2012, Beaulieu & O'Meara *Syst Biol* 2016...)
- Four survey articles ! ! (Ricklefs *TREE* 2007, Pyron & Burbrink *TREE* 2013, Stadler *JEB* 2013, Morlon *Eco Lett* 2014)
- Species $\neq$ particles : lineage-based models do not inform us about the process of speciation
- Progressive emergence of reproductive isolation (RI) is ubiquitous
  - RI as a by-product of local adaptation : allopatric speciation

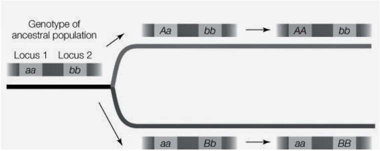


What's next ?

Open the species box !

- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)
  - Age-dependent div (Alexander et al *Syst Biol* 2015, Hagen et al *Syst Biol* 2015)
  - Diversity-dependent diversification (Etienne et al *Proc B* 2012)
  - Trait-dependent diversification : BiSSE, QuaSSE, HiSSE... (Maddison et al *Syst Biol* 2007, FitzJohn *MEE* 2012, Beaulieu & O'Meara *Syst Biol* 2016...)
- Four survey articles ! ! (Ricklefs *TREE* 2007, Pyron & Burbrink *TREE* 2013, Stadler *JEB* 2013, Morlon *Eco Lett* 2014)
- Species $\neq$ particles : lineage-based models do not inform us about the process of speciation
- Progressive emergence of reproductive isolation (RI) is ubiquitous
  - RI as a by-product of local adaptation : allopatric speciation
  - Bateson-Dobzhansky-Muller (BDM) incompatibilities : start with 2 monomorphic pop *aabb*, evolving as *AAbb* and *aaBB* resp., with $AAbb \times aaBB = AaBb$ unviable





What's next ?

Open the species box !

- Lineage-based models of macro-evolution
  - Time-dependent div (Morlon et al *PNAS* 2011, Stadler *PNAS* 2011)
  - Age-dependent div (Alexander et al *Syst Biol* 2015, Hagen et al *Syst Biol* 2015)
  - Diversity-dependent diversification (Etienne et al *Proc B* 2012)
  - Trait-dependent diversification : BiSSE, QuaSSE, HiSSE...
    (Maddison et al *Syst Biol* 2007, FitzJohn *MEE* 2012, Beaulieu & O'Meara *Syst Biol* 2016...)
- Four survey articles ! ! (Ricklefs *TREE* 2007, Pyron & Burbrink *TREE* 2013, Stadler *JEB* 2013, Morlon *Eco Lett* 2014)
- Species $\neq$ particles : lineage-based models do not inform us about the process of speciation
- Progressive emergence of reproductive isolation (RI) is ubiquitous
  - RI as a by-product of local adaptation : allopatric speciation
  - Bateson-Dobzhansky-Muller (BDM) incompatibilities : start with 2 monomorphic pop *aabb*, evolving as *AAbb* and *aaBB* resp., with $AAbb \times aaBB = AaBb$ unviable

$\Rightarrow$ Needs to open the species box

Defining species in individual-based models

- Species $\neq$ elementary particles

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can
  - Replicate $=$ colonization of a new habitat by founders from a seed pop

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can
  - Replicate $=$ colonization of a new habitat by founders from a seed pop
  - Die $=$ local extinction of a population

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can
  - Replicate $=$ colonization of a new habitat by founders from a seed pop
  - Die $=$ local extinction of a population
  - Mutate $=$ major genetic/phenotypic change, new stage in speciation process

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can
  - Replicate $=$ colonization of a new habitat by founders from a seed pop
  - Die $=$ local extinction of a population
  - Mutate $=$ major genetic/phenotypic change, new stage in speciation process
- Speciation $=$ consequence of genetic/phenotypic change/differentiation

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can
  - Replicate $=$ colonization of a new habitat by founders from a seed pop
  - Die $=$ local extinction of a population
  - Mutate $=$ major genetic/phenotypic change, new stage in speciation process
- Speciation $=$ consequence of genetic/phenotypic change/differentiation
- Compared to lineage-based models, we seek

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can
  - Replicate $=$ colonization of a new habitat by founders from a seed pop
  - Die $=$ local extinction of a population
  - Mutate $=$ major genetic/phenotypic change, new stage in speciation process
- Speciation $=$ consequence of genetic/phenotypic change/differentiation
- Compared to lineage-based models, we seek

  (A) A natural way of partitioning particles into species according to their geno/phenotypes

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can
  - Replicate $=$ colonization of a new habitat by founders from a seed pop
  - Die $=$ local extinction of a population
  - Mutate $=$ major genetic/phenotypic change, new stage in speciation process
- Speciation $=$ consequence of genetic/phenotypic change/differentiation
- Compared to lineage-based models, we seek
  - (A) A natural way of partitioning particles into species according to their geno/phenotypes
  - (B) A unique way of defining the species phylogeny consistently with the genealogy of particles

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can
  - Replicate $=$ colonization of a new habitat by founders from a seed pop
  - Die $=$ local extinction of a population
  - Mutate $=$ major genetic/phenotypic change, new stage in speciation process
- Speciation $=$ consequence of genetic/phenotypic change/differentiation
- Compared to lineage-based models, we seek
  - (A) A natural way of partitioning particles into species according to their geno/phenotypes
  - (B) A unique way of defining the species phylogeny consistently with the genealogy of particles
  - (C) A fast algorithm simulating the partition and the phylogeny

Defining species in individual-based models

- Species $\neq$ elementary particles
- Now elementary particles are (individuals or) populations that can
  - Replicate = colonization of a new habitat by founders from a seed pop
  - Die = local extinction of a population
  - Mutate = major genetic/phenotypic change, new stage in speciation process
- Speciation = consequence of genetic/phenotypic change/differentiation
- Compared to lineage-based models, we seek
  - (A) A natural way of partitioning particles into species according to their geno/phenotypes
  - (B) A unique way of defining the species phylogeny consistently with the genealogy of particles
  - (C) A fast algorithm simulating the partition and the phylogeny
  - (D) A statistical method for the inference of microscopic parameters of the process

Outline

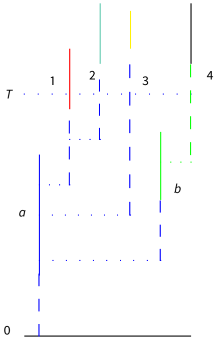
Model 1. Protracted Speciation

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

- Idea : Speciation takes time
- Species $=$ ensemble of pops, each pop gradually diverges from mother species
- Speciation is complete when a pop has accumulated k **mutations**
- Newborn particles are in stage **'incipient'** $= \in$ same species as mother population
- Arrive in stage **'good'** after k mutations $=$ new species (A)
- Each species is represented by one single particle
- Phylogeny $=$ tree (genealogy of particles) spanned by representative particles (B)

Model 1. Protracted Speciation – cont'd

Lambert, Morlon & Etienne "The reconstructed tree in the lineage-based model of protracted speciation" *J Math Biol* (2015)



- Here $k = 1$
- 4 extant populations at time T
- 3 extant species
- Species b is represented by Population 4
- Representative $=$ **leftmost particle** in natural tree orientation
- Species a is **represented** by Population 2.

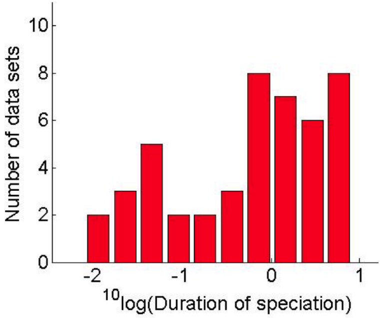
Proposition (Lambert, Morlon & Etienne 2015)

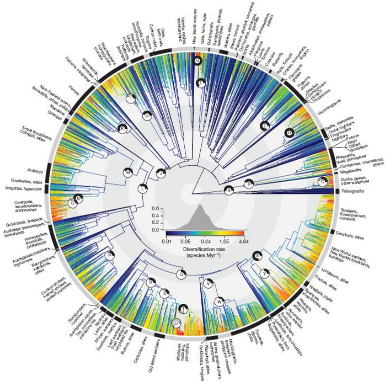
If the pop birth rate does not depend on speciation stage, then the tree spanned by ***representative*** *populations sampled at T is a* **coalescent point process** *with explicit node depth distribution (C, D).*

Model 1. Protracted Speciation – cont'd

Etienne, Morlon & Lambert "Estimating the duration of speciation from phylogenies" *Evolution* (2014)

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



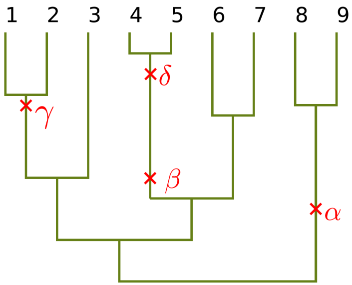


Jetz et al (2012)

Model 2. Speciation by Genetic Differentiation

Manceau & Lambert "The species problem from the modeler's point of view" *Bull Math Biol* (2019)

- No knowledge of mother species (ancestral state)
- Define species by one of the following two rules :
  - **Rule 1.** Particles separated by $\leq$ **q** mutations are in the **same** species.
  - **Rule 2.** Particles separated by $>$ **q** mutations are in **different** species.
- Partition into species (A) ? Species phylogeny (B) ?
- A subset of tips is monophyletic $=$ forms a subtree
- If species form monophyletic subsets
  $\Rightarrow$ $\exists!$ phylogeny obtained by collapsing each subset into one tip



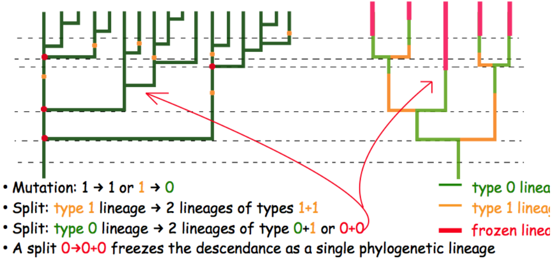
Proposition (Manceau & Lambert 2019)

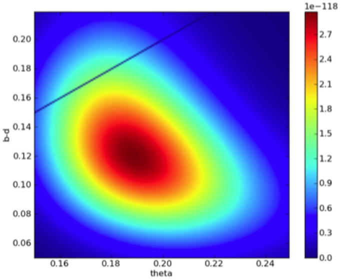
1. $\exists!$ *species partition that is the finest monophyletic partition satisfying Rule 1.*
2. $\exists!$ *species partition that is the coarsest monophyletic partition satisfying Rule 2.*

Model 2. Speciation by Genetic Differentiation – cont'd

Manceau, Lambert & Morlon "Phylogenies support out-of-equilibrium models of biodiversity" *Ecology Letters* (2015)

- Start an ind-based birth-death b, d process, Poisson mutations at rate θ, species and phylogeny defined from finest monophyletic partition (A, B) such that two clonal tips $\in$ same species (Rule 1, $q = 1$).
- The phylogeny can be generated by a 3-type time-inhom. branching process (C)
  - 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
- Likelihood computation by peeling algorithm (D), including the case of missing species
- Tests on simulations : precise ML estimates of θ and $b - d$





Model 3. The Split-and-Drift Evolving Graph

Bienvenu, Débarre & Lambert “The split-and-drift random graph, a null model for speciation” *SPA* (2019)

- SGD : draw an edge between particles separated by $\leq q$ differences (genealogy + mutations)
- Here : draw an edge between particles able to interbreed
- Minimal assumption : interbreeding evolves by
  - Plain replication : ‘Split’
  - Spontaneous divergence : ‘Drift’
- The interbreeding relationship is not transitive : e.g., ring species (see figure)
- Species = connected components of interbreeding graph (A)

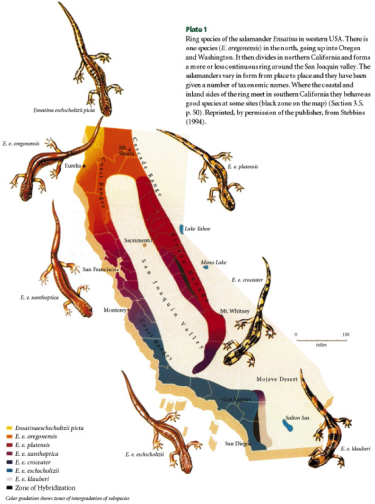
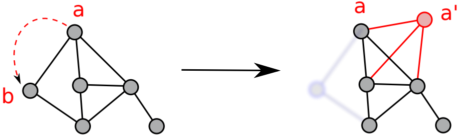


Illustration by Randy Schmieder. Reprinted from *Life on the Edge : A Guide To California’s Endangered Natural Resources* by Carl G. Thelander. Copyright 1994 by Ten Speed Press, Berkeley, CA

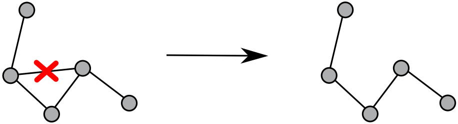
Split-and-Drift Evolving Graph

Bienvenu, Débarre & Lambert "The split-and-drift random graph, a null model for speciation" *SPA* (2019)

- n populations
- Extinction-recolonization as in Moran model

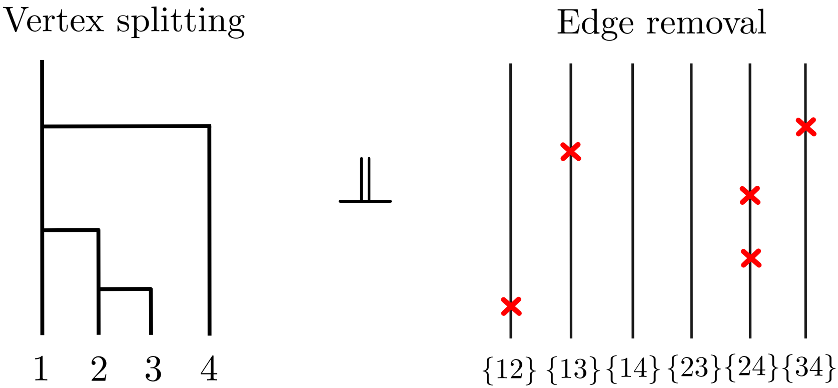


  - At rate 1/2 per oriented pair (a, b) : pop b goes extinct $+$ is replaced by a copy of pop a
  - The new pop a' inherits neighbors of mother pop a $+$ new edge mother-daughter
- Divergence : each edge disappears at rate r



- $G_{n,r} :=$ stationary state of this graph
- Simple two-parameter model
  - $n =$ metapopulation capacity
  - $r =$ rate of evolution of reproductive isolation

Backward-Forward Approach



⇒ Kingman coalescent (rate 1) + pairwise Poisson processes (rate r)

⇒ Super fast simulation of the graph at stationarity (C)

Degree and cliques : moments

Bienvenu, Débarre & Lambert "The split-and-drift random graph, a null model for speciation" *SPA* (2019)

- Fix k nodes in $G_{n,r}$
- By standard argument of competing clocks, the probability that these k nodes form a clique is

$$p_k(n,r) := \prod_{j=2}^{k} \frac{\binom{j}{2}}{\binom{j}{2} + r\binom{j}{2}} = \left(\frac{1}{1+r}\right)^{k-1}$$

- For $k = 2$ fixed nodes, probability of edge presence is

$$p_2(n,r) = \frac{1}{1+r}$$

- $D(n,r) :=$ Degree of a fixed node

$$\mathbb{E}(D(n,r)) = \frac{n-1}{1+r} \sim \frac{n}{r} \text{ as } n, r \to \infty$$

Degree and connected components : limiting behavior

Bienvenu, Débarre & Lambert "The split-and-drift random graph, a null model for speciation" *SPA* (2019)

Recall D = degree of a fixed node and $\#CC$ = number of connected components.

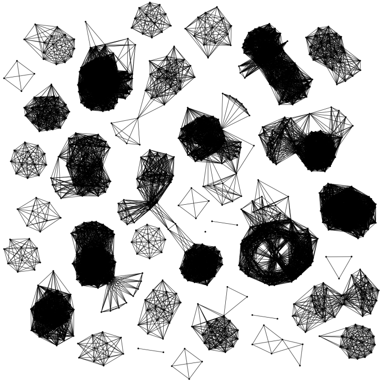
Theorem (Bienvenu, Débarre & L. 2019)

Assume that as $n \to \infty$, $r_n \to \infty$ and $r_n/n \to 0$. Then

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{D(n, r_n)}{n/r_n} > x\right) = \int_x^\infty 4ye^{-2y}dy$$

and

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{2} \leq \frac{\#CC(G_{n,r_n})}{r_n} \leq 2\right) = 1.$$

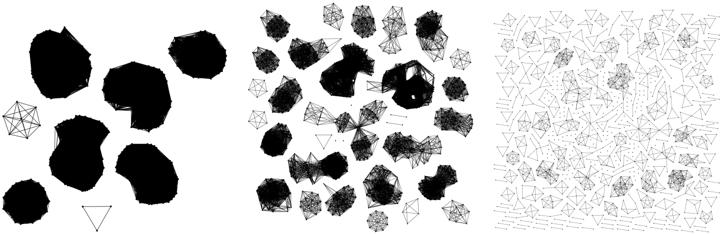


$n = 1000, r = 54$

Perspectives

Bienvenu, Débarre & Lambert "The split-and-drift random graph, a null model for speciation" *SPA* (2019)

- A highly tractable neutral model for the evolution of RI
- Convergence in distribution of $\#CC/r_n$?
- Distribution of sizes of connected components ?
- Convergence in the graphon sense ? (dense regime, r constant)
- Definition/simulation/law of the phylogeny (B,C) ?
- Inference (D) ?



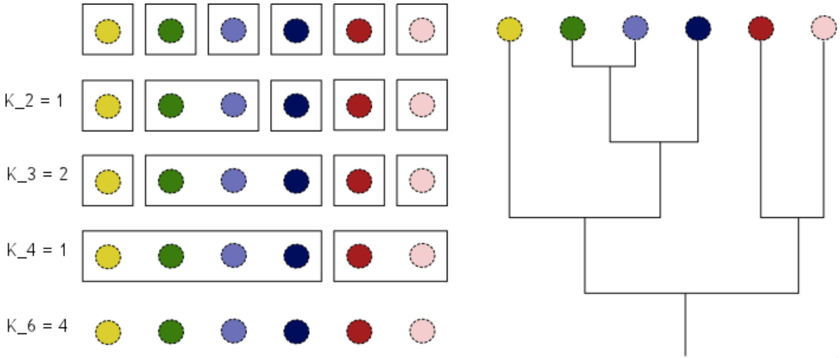
$n = 1000$. Left : $r = 5$, middle : $r = 41$, right : $r = 347$.

Outline

Aldous' Markov branching model on binary tree shapes

Aldous (1996, 2001)

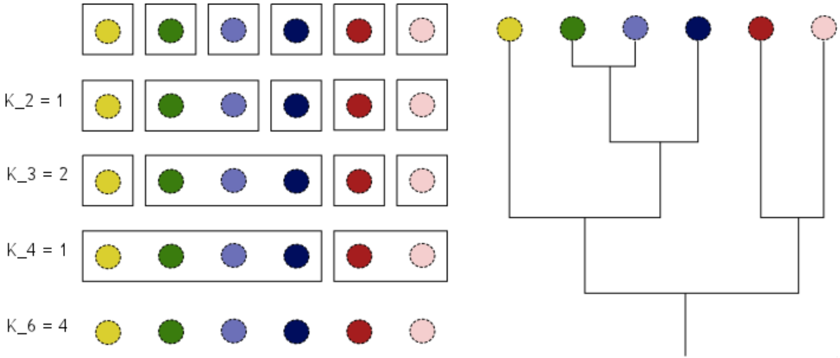
- Goal : generate a random, binary tree T_n with n exchangeable tips labelled by $\{1, \ldots, n\}$
- Assume given distributions q_n on $\{1, \ldots, n-1\}$, $n \geq 2$
- Recursively split each subset of n balls according to q_n (r.v.'s K_n below)



Aldous' Markov branching model on binary tree shapes

Aldous (1996, 2001)

- Goal : generate a random, binary tree T_n with n exchangeable tips labelled by $\{1, \ldots, n\}$
- Assume 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 stopped at a fixed time and Kingman coalescent

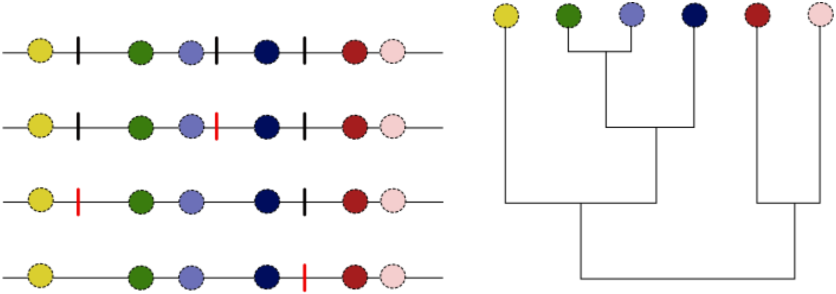
Sampling consistency

- Recall T_n is a random, binary tree with n exchangeable tips labelled by $\{1, \ldots, n\}$.
- 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

Construction

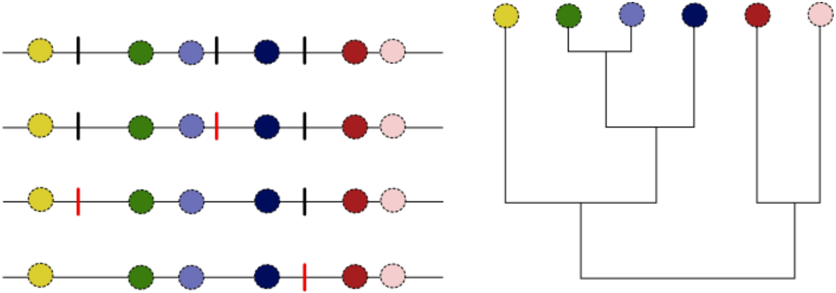
- Color dots are uniformly distributed in the interval
- Intervals are iteratively fragmented by r.v. with law μ



Aldous' Markov branching model

Construction

- Color dots are uniformly distributed in the interval
- Intervals are iteratively fragmented by r.v. with law μ



Theorem (Haas, Miermont, Pitman & Winkel 2008, Lambert 2016)

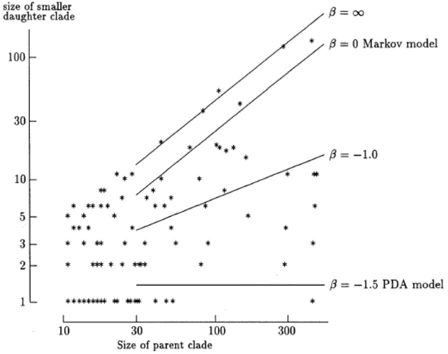
A MB tree model is sampling-consistent IFF it there is a symmetric measure μ on $[0,1]$ s.t.

$$q_n(i) = a_n(f)^{-1}\left\{ \binom{n}{i} \int_{(0,1)} x^i (1-x)^{n-i} \mu(dx) + n\mu(\{0\})1_{i=1} + n\mu(\{1\})1_{i=n-1} \right\}$$

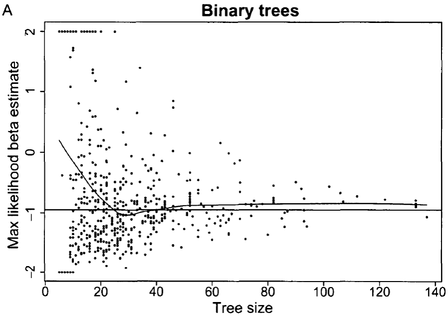
The β-splitting model

- The β-splitting model is for $\beta \in (-2, \infty)$: $\mu(dx) = cx^{\beta}(1-x)^{\beta}dx$
- Imbalance decreases with β

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



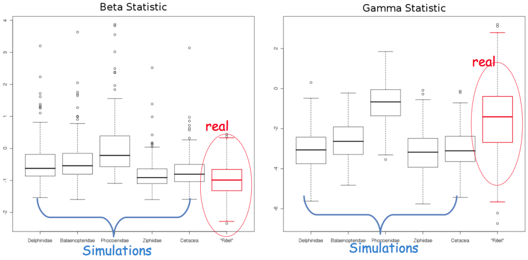
MLE of β (Blum & François 2006)



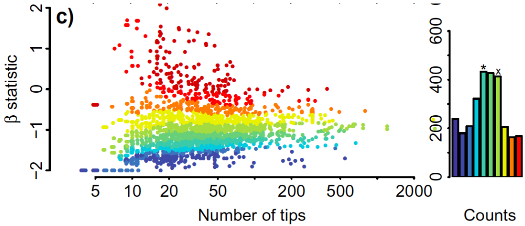
Aldous (2001) : "Why $\beta \approx -1$?" or "Are there mathematically simple/biologically plausible stochastic models for phylogenetic trees whose realizations mimic actual trees"

Why $\beta \approx -1$?

- Birth-death processes where $b = b(t, n)$ and $d = d(t, n, a)$ produce same tree shapes as $\beta = 0$
- Protracted speciation (Model 1) produces same tree shapes as $\beta = 0$
- SGD (Model 2) : Inference from Cetaceans generates realistic values of β and γ
- Age-dependent speciation rate $b = b(a) = ca^{\phi - 1}$ Hagen et al (2015)
    - Estimates of ϕ for 9243 empirical species trees from *TreeBase*
    - Estimates of ϕ lie in $(0, 1)$: speciation rate decreases with age
    - Distribution of β generated by ϕ estimates fits well



Manceau, Lambert & Morlon *Eco Lett* 2015



Hagen, Hartmann, Steel & Stadler *Syst Biol* 2015

Collaborators

Tanja STADLER (ETHZ) . . . . 

Helen ALEXANDER (Edinburgh) & Tanja STADLER (ETHZ) . . . .  

Rampal ETIENNE (Groningen) & Hélène MORLON (ENS Paris) . . . .  

Marc MANCEAU (ETHZ) . . . . 

Marc MANCEAU (ETHZ) & Hélène MORLON (ENS Paris) . . . .  

François BIENVENU (Oxford) & Florence DÉBARRE (CNRS) . . . .  

Outline

Books

- Semple, C. & Steel, M. (2003) *Phylogenetics.*
- Gascuel, O. (Ed) (2005) *Mathematics of Evolution and Phylogeny.*
- Evans, S. N. (2008). *Probability and real trees. — École d'été de probabilités de Saint-Flour XXXV-2005.*
- Dress, A., Huber, K., Koolen, J., Moulton, V. & Spillner, A. (2012) *Basic phylogenetic combinatorics.*
- Steel, M. (2016) *Phylogeny : Discrete and random processes in evolution.*

Monographs

- Lambert, “Population dynamics and random genealogies”, *Stoch Models* (2008)
- Le Gall, “Random trees and applications”, *Probability Surveys* (2005)
- Lambert, “Probabilistic models for the (sub)tree(s) of life”, *Braz J Probab Stat* (2017)
- Lambert, “Random ultrametric trees and applications”, *ESAIM P& S* (2018)

Papers

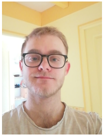
- Nee, May & Harvey, “The reconstructed evolutionary process”, *Phil Trans Roy Soc* (1994)
- Aldous, “Probability distributions on cladograms”, In *Random Discrete Structures* (1996)
- Aldous, “Stochastic models and descriptive statistics for phylogenetic trees...” *Stat Sci* (2001)
- Popovic, “Asymptotic genealogy of a critical branching process”, *Ann Appl Prob* (2004)
- Blum & François, “Which random processes describe the tree of life ? A large-scale study of phylogenetic tree imbalance”, *Syst Biol* (2006)
- Nee, “Birth-death models in macroevolution”, *Ann Rev Ecol Evol Syst* (2006)
- Haas, Miermont, Pitman & Winkel, “Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models”, *Ann Probab* (2008)
- Lambert, “The contour of splitting trees is a Lévy process”, *Ann Probab* (2010)
- Etienne & Rosindell, “Prolonging the past counteracts the pull of the present : Protracted speciation can explain...”, *Syst Biol* (2012)
- Lambert & Stadler, “Birth-death models and coalescent point processes : The shape and probability of reconstructed phylogenies” *Theoret Popul Biol* (2013)
- Etienne, Morlon & Lambert, “Estimating the duration of speciation from...”, *Evolution* (2014)
- Lambert, Morlon & Etienne, “The reconstructed tree in the lineage-based model of protracted speciation”, *J Math Biol* (2015)
- Alexander, Lambert & Stadler, "Quantifying age-dependent extinction from...", *Syst Biol* (2015)
- Manceau, Lambert & Morlon, “Phylogenies support out-of-equilibrium models of biodiversity", *Ecology Letters* (2015)
- Hagen, Hartmann, Steel & Stadler, “Age-dependent speciation can explain the shape of empirical phylogenies", *Syst Biol* (2015)
- Manceau & Lambert, “The species problem from the modeler’s point of view”, *Bull Math Biol* (2019)
- Bienvenu, Débarre & Lambert, “The split-and-drift random graph, a null model for speciation”, *Stoch Proc Appl* (2019)

SMILE : An interdisciplinary group in Paris

Below : SMILE members in May 2020

Jasmine 12/01

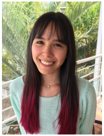
François 15/01



Guillaume T. 25/02

Pete 29/03

Emmanuel 18/04



Laura 21/05

Jean-Jil 24/05

Felix 05/09

Rob 16/09

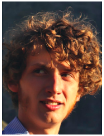
Julie 21/09

Elise 10/10

Léo 07/11

Guillaume A. 09/12

Amaury 16/12



Philibert 30/12

Alejandro

SMILE = **S**tochastic **M**odels for the **I**nference of **L**ife **E**volution

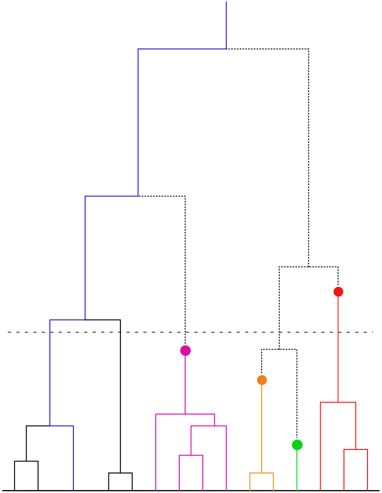
Degree Distribution : Proof (1)

Fix one node, say n, in $G_{n,r}$ and follow its lineage backward in time...

- Edge removal :
  - Each pair $\{i, n\}$ has a rate r Poisson process of edge removal
  - At the first atom backward in time, kill the lineage and color all its descending subtree
  - When $k+1$ lineages, the probability that the next event is a killing rather than a coalescence is

$$\frac{rk}{\binom{k+1}{2} + rk} = \frac{2r}{k+1+2r}$$

- Vertex splitting :
  - When $k+1$ lineages, the distinguished lineage is involved in the next coalescence event with probability $2/(k+1)$



Degree of blue node $D = 4$

Degree Distribution : Proof (2)

- Let (I_k, J_k) denote the numbers of uncolored/colored lineages when there are k lineages.

- $(I_k, J_k; k \geq 0)$ is a Markov chain starting from $(1, 0)$ with transition probabilities, writing $k = i + j$

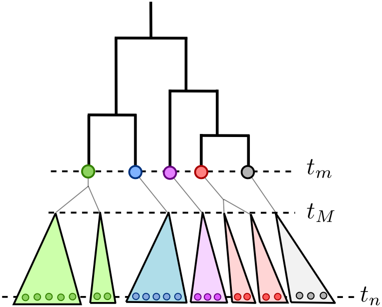
$$(i,j) \longrightarrow \begin{cases} (i+1, j) \text{ w. pr. } \dfrac{k+1}{k+1+2r}\,\dfrac{i+1}{k+1} & = \dfrac{i+1}{i+j+1+2r} \\[2ex] (i, j+1) \text{ w. pr. } \dfrac{k+1}{k+1+2r}\,\dfrac{j}{k+1} + \dfrac{2r}{k+1+2r} & = \dfrac{j+2r}{i+j+1+2r} \end{cases}$$

- $I_n - 1 =$ degree of distinguished node $+$ elementary calculations. □

Connected components

Bienvenu, Débarre & Lambert "The split-and-drift random graph, a null model for speciation" *SPA* (2019)

- Assume $1 \ll r_n \ll n$.
- Let $t_k :=$ time when the coalescent tree has k lineages
- Lower bound : Choose m s.t. the graph at time t_m is empty w.h.p.
  Result : $m \sim \frac{r_n}{2}$
- Upper bound : Choose M s.t. the descending subtrees of each of the M nodes of time t_M are connected w.h.p.
  Result : $M \sim 2r_n \log(n)$



Theorem

Assume that as $n \to \infty$, $r_n \to \infty$ and $r_n/n \to 0$. Then

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{r_n}{2} \leq \#CC(G_{n,r_n}) \leq 2r_n \log(n)\right) = 1.$$