

Bayesian estimation of time-trees:

A journey through a strange land

Luiz Max Carvalho [lmax.fgv@gmail.com]

School of Applied Mathematics

Getúlio Vargas Foundation, Rio de Janeiro.



Acknowledgments



Andrew Rambaut
UoE



Marc Suchard
UCLA



Guy Baele
KU Leuven

Plan for today

Problem

What are trees and why are interested in them?

Parameter space

What does the space we are trying to explore look like?

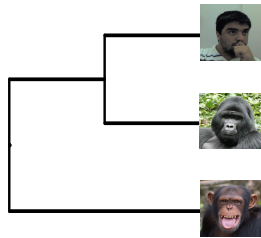
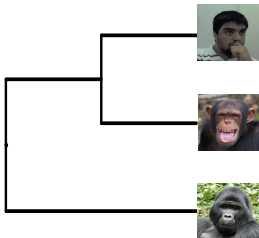
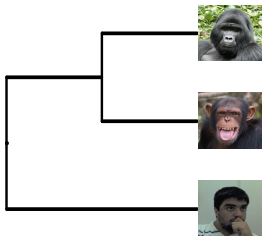
MCMC in tree space

A journey through a strange land

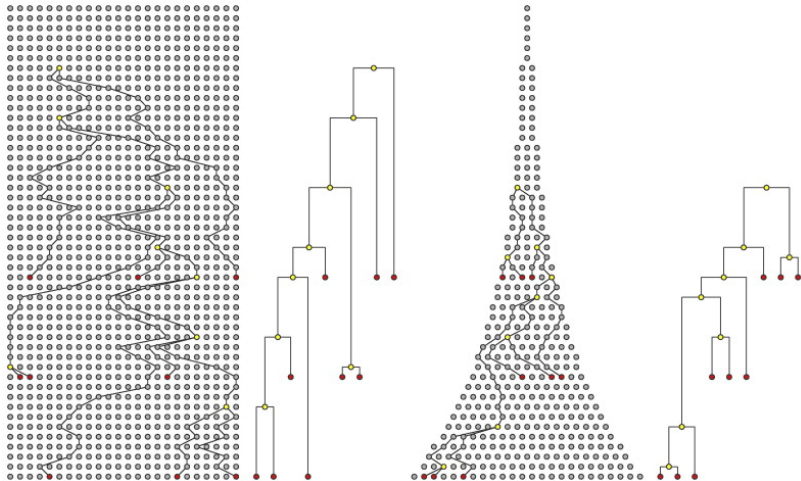
Preliminary results and perspectives

Performance analyses and open problems.

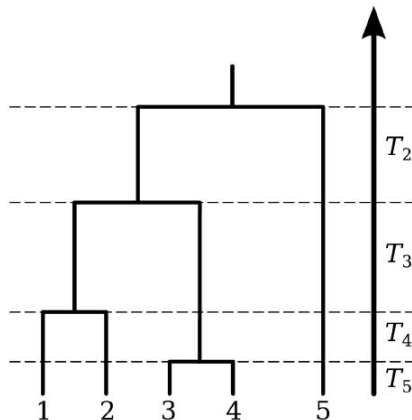
Trees are hypotheses



Trees and the coalescent



Central object: time-calibrated trees



Let T_n denote the time for n lineages to *coalesce*, i.e., merge into one ancestral lineage, in a population of size N_e . Then:

$$Pr(T_n = t) = \lambda_n e^{-\lambda_n t}$$

$$\lambda_n = \binom{n}{2} \frac{1}{N_e} = \binom{n}{2} \frac{1}{N_e \tau}$$

where N_e is the effective population size and τ is the generation time. Let T_{mrca} denote the age of the most recent common ancestor:

$$\begin{aligned} \mathbb{E}[T_{\text{mrca}}] &= \mathbb{E}[T_n] + \mathbb{E}[T_{n-1}] + \dots + \mathbb{E}[T_2] \\ &= 1/\lambda_n + 1/\lambda_{n-1} + \dots + 1/\lambda_2 \\ &= 2N_e \left(1 - \frac{1}{n}\right) \end{aligned}$$

Figure: Figure 4 from [Volz et al. \(2013\)](#).

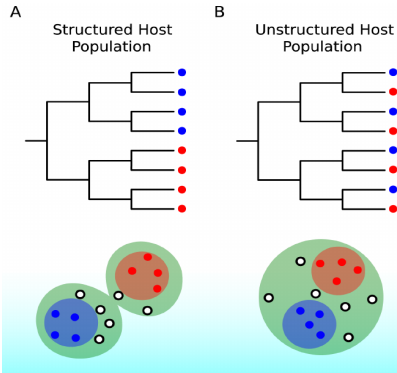
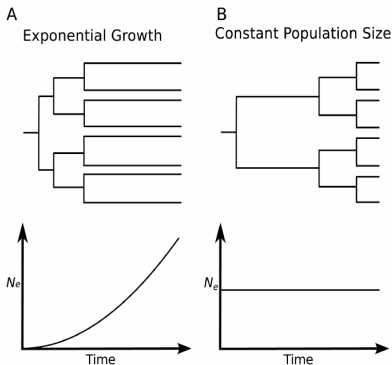
Motivation

Phylogenetics of fast-evolving viruses

Inferring spatial and temporal dynamics from genomic data:

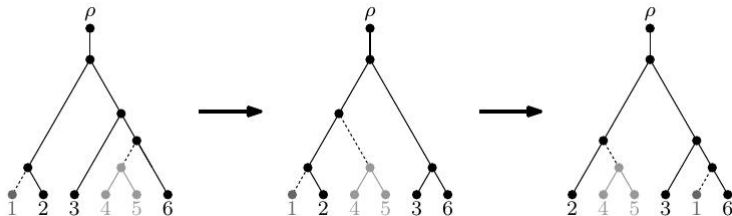
Phylogenies*!

* plus complicated models



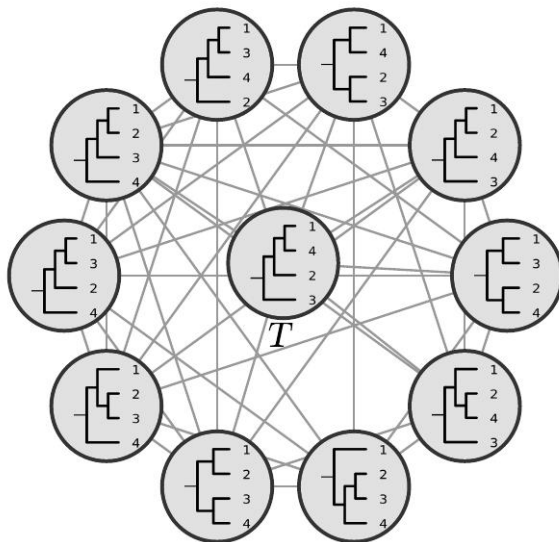
Discrete tree space: tree surgery

Subtree prune-and-regraft (SPR):



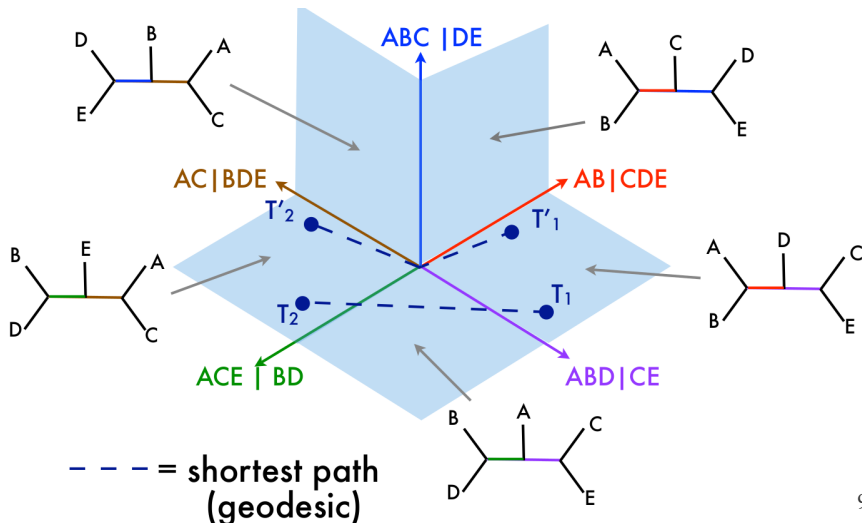
Discrete tree space: SPR graph

For curvature results, see [Whidden & Matsen\(2017\)](#).

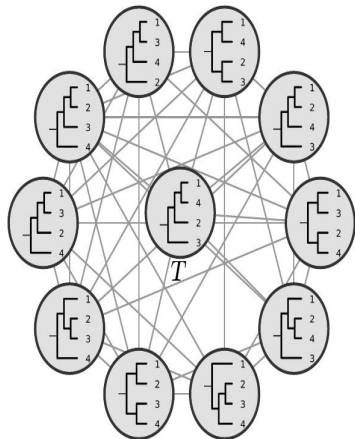
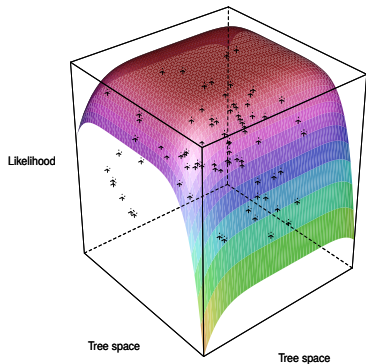


Continuous tree space: BHV

Billera, Holmes & Vogtmann (2001).



Tree space: a strange land



$$p(t, \mathbf{b}, \boldsymbol{\omega} | D) = \frac{f(D|t, \mathbf{b}, \boldsymbol{\omega})\pi(t, \mathbf{b}, \boldsymbol{\omega})}{\sum_{t_i \in T_n} \int_B \int_{\Omega} f(D|t_i, \mathbf{b}_i, \boldsymbol{\omega})\pi(t_i, \mathbf{b}_i, \boldsymbol{\omega})d\boldsymbol{\omega}d\mathbf{b}_i} \quad (1)$$

- ⊙ D : observed sequence (DNA) data;
- ⊙ T_n : set of all binary ranked trees ($\mathbb{G}^{(2n-3)!!}$);
- ⊙ \mathbf{b}_k : set of branch lengths of $t_k \in T_n$ (\mathbb{R}_+^{2n-2} , kind of) ;
- ⊙ $\boldsymbol{\omega}$: set of parameters of interest such as substitution model parameters, migration rates, heritability coefficients, etc.

(Adaptive) Metropolis-Hastings for trees

General MH setup.

Let $\tau = (t, \mathbf{b})$ denote a tree with topology t and branch lengths \mathbf{b} . For two trees τ and τ' , denote the transition kernel by $q_\gamma(\tau|\tau') := \Pr(\tau' \rightarrow \tau|\gamma)$.

Accepting with probability

$$A_\gamma(\tau|\tau') = \min \left(1, \frac{p(\tau', \omega|D)q_\gamma(\tau|\tau')}{p(\tau, \omega|D)q_\gamma(\tau'|\tau)} \right)$$

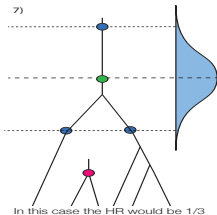
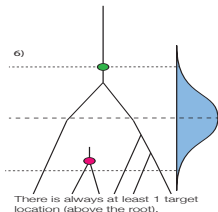
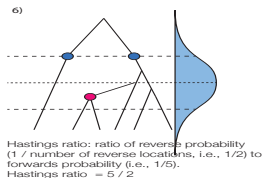
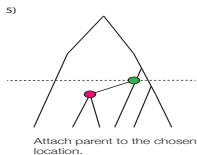
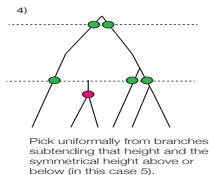
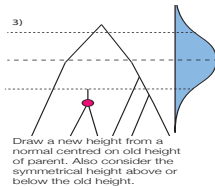
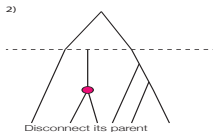
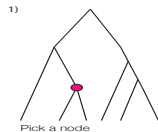
leads to the desired target.

Note: Here $\gamma > 0$ is a so-called tuning parameter.

Height-constrained kernels: SubTreeLeap (STL)

1. Excluding the root, pick a node i in τ uniformly at random, i.e., with probability $1/(2n - 3)$;
2. Draw a patristic distance δ from the distance kernel $k(\delta|\sigma)$;
3. Find the set of destination nodes \mathbf{D}_i^δ that are within distance δ **and** whose heights are not less than $h(i) - \delta$;
If $\mathbf{D}_i^\delta = :$
 - prune p_i and regraft it at height $h_b = h(p_i) - \delta$ or $h_a = h(p_i) + \delta$ with probability $1/2$, creating a new tree τ' , else
 - pick a node $j \in \mathbf{D}_i^\delta$ with probability $Pr(i \rightarrow j) = 1/|\mathbf{D}_i^\delta|$, prune the tree at p_i and regraft it at p_j , creating a new tree τ' ;

STL – illustration



- ⊙ Adaptive → more efficient (?);
- ⊙ Height-constrained → time-precedence constraints are respected;
- ⊙ Changes topology and branch lengths **simultaneously** → presumably more efficient;
- ⊙ Inherits cool properties from SPR.
 - We know a bunch of things about the SPR graph;
 - SPR graph admits a Hamiltonian ([Gordon et al., 2013](#)).

Carvalho (2019), Chapter 2.

Remark

Assume strictly positive branch lengths. Then SubTreeLeap induces an irreducible Markov chain on \mathbb{G} .

Sketch: Starting at $x \in \mathbb{G}$, notice there exists $\delta_y^\star > 0$ such that $P(x \rightarrow y \mid \delta_y^\star) > 0$ for any tree $y \in \mathbb{G}$ in the SPR neighbourhood of x .

Theorem

Assume the target satisfies $p(A) > 0$ for all $A \subset \Psi$. Then, SubTreeLeap induces an ergodic Markov chain on Ψ .

Sketch: Employ the remark to get to the case where $d_{\text{SPR}}(x, y) = 0$ and then show establish Harris recurrence.

Experimental setup

All MCMC implemented in the JAVA open-source software BEAST (<http://beast.community/>);

- ⊙ Default kernels:

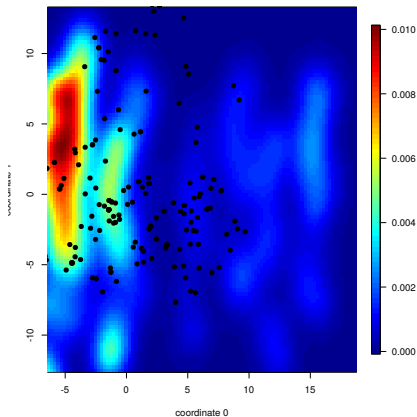
- SubTreeSlide – adaptive, rarely moves topology;
- Narrow exchange – non-adaptive, local moves;
- Wide exchange – non-adaptive, bold moves;
- NodeHeights – scale all node heights by a factor (within their bounds);

- ⊙ SubTreeLeap;

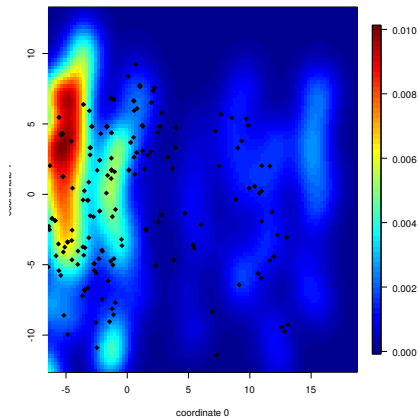
- Most results will be shown for 100 MCMC runs.

Traversing tree space – Topology

Default kernels

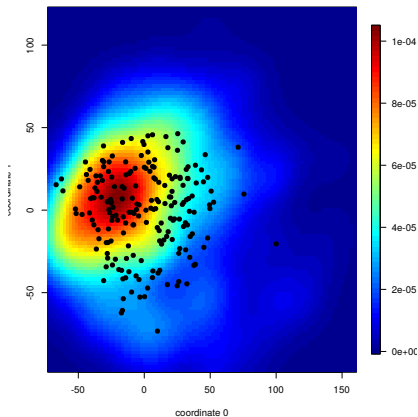


STL

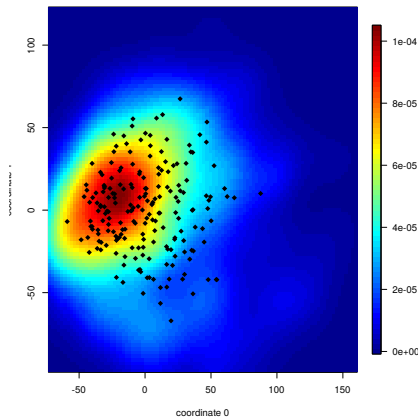


Traversing tree space – Topology + branch lengths

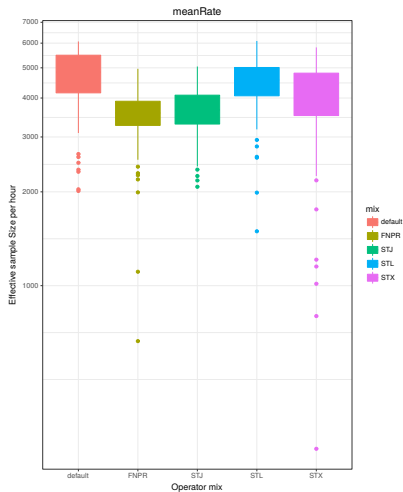
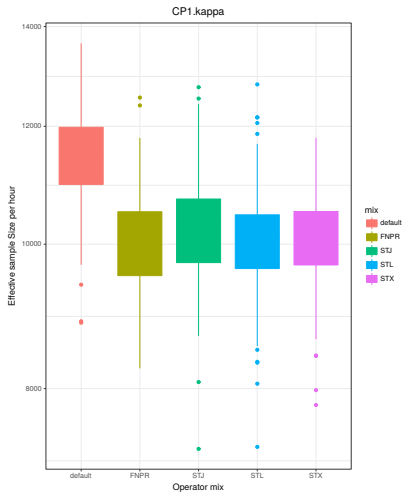
Default kernels



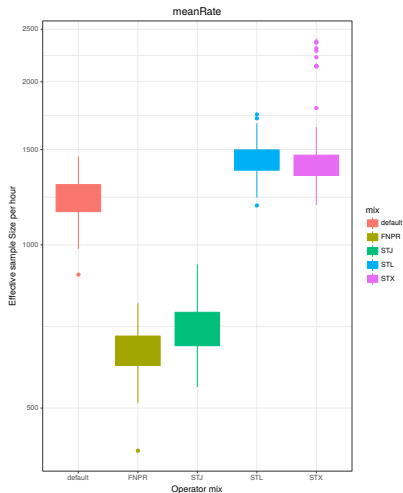
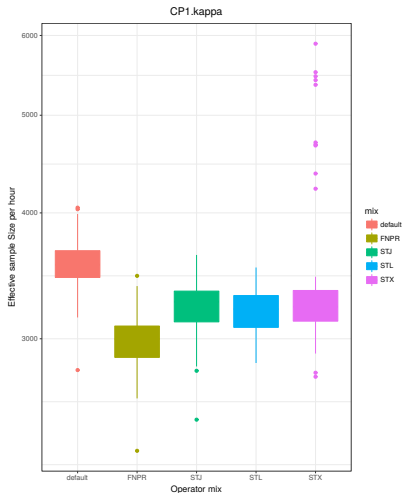
STL



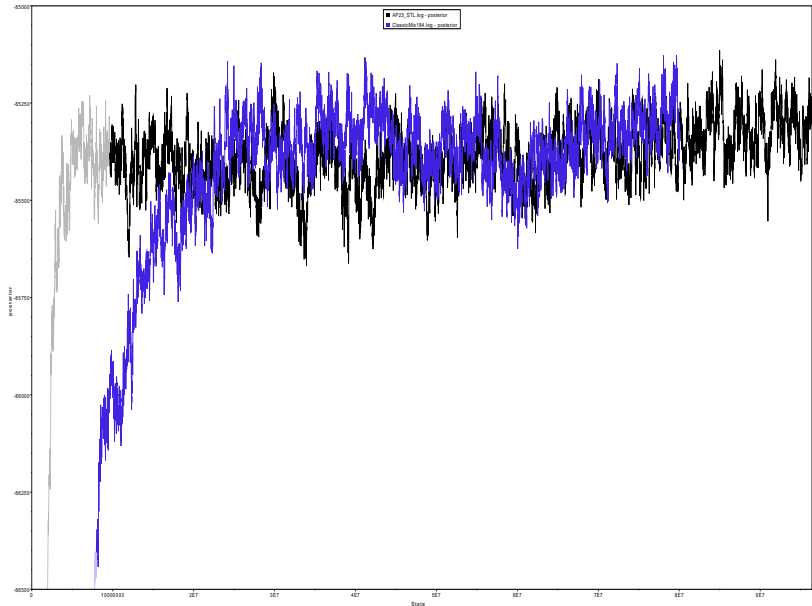
Dengue 4 *env* (17 taxa, 1485 NT sites)



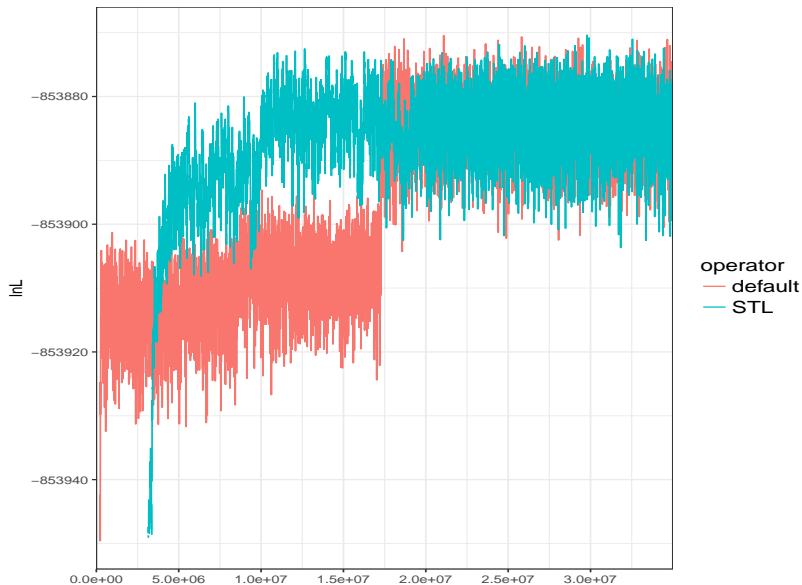
RSVA G protein (35 taxa, 629 NT sites)



Ebola virus full genome (1610 taxa (!), 18990 NT sites)



Metazoans (contemporaneous, 55 taxa, 30257 AA sites)



Open problems in MCMC for phylogenies

Open problems:

- How can we construct more efficient proposals? How to exploit structure? **Geometry!**
- How to quantify exploration of the target? **Tools!**
- **Optimal scaling: what's the optimal acceptance probability?**

Statistics in the space of phylogenetic trees

- ⊙ Central Limit Theorem(s) in BHV space: [Barden, Le & Owen \(2013\)](#);
- ⊙ “Statistics in the Billera-Holmes-Vogtmann space”: [Weyenberg \(2015\)](#);
- ⊙ Consistency of the MLE: [RoyChoudhury, Willis & Bunge \(2015\)](#);
- ⊙ How to turn tree space into an Euclidean space: [Barden & Le \(2017\)](#);
- ⊙ Quantifying uncertainty about phylogenies: [Willis & Bell \(2018\)](#);
- ⊙ Confidence sets for phylogenies: [Willis \(2018\)](#);
- ⊙ Probabilistic path Hamiltonian Monte Carlo for phylogenies: [Dinh et al. \(2017\)](#).

Searching trees is **hard**

Complicated and **HUGE** parameter space

¹this talk is available [online](#)

Take home¹

Searching trees is **hard**

Complicated and **HUGE** parameter space

Height-preserving tree rearrangements are **good**

Use the extra information provided by the tip dates

¹this talk is available [online](#)

Take home¹

Searching trees is **hard**

Complicated and **HUGE** parameter space

Height-preserving tree rearrangements are **good**

Use the extra information provided by the tip dates

Adaptive moves are more efficient

Avoid wasting computing power

¹this talk is available [online](#)

Take home¹

Searching trees is **hard**

Complicated and **HUGE** parameter space

Height-preserving tree rearrangements are **good**

Use the extra information provided by the tip dates

Adaptive moves are more efficient

Avoid wasting computing power

Much more work is needed

We should prepare for an era of plenty

¹this talk is available [online](#)

THE
END