Al buio non si trova

Biostatistics in the 21st century

Luiz Max Carvalho

lmax.fgv@gmail.com

Available from: https://github.com/maxbiostat/presentations/

FGV EMAp

Plan for today

Music

A little metaphor to get us in the mood

Problem I: using historical data to the fullest

Optimal Bayesian dynamic borrowing of information

Problem II: dealing with huge complex data

MCMC in tree space: a journey through a strange land

A football metaphor

Because why not include a second metaphor?

A (useful?) methaphor

Che gelida manina (La Boheme, Puccini, 1896)

Che gelida manina, se la lasci riscaldar.

What a frozen little hand, let me warm it for you.

Cercar che giova? Al buio non si trova.

What's the use of looking? We won't find it in the dark.

Ma per fortuna, è una notte di luna,

But luckily, it's a moonlit night,

E qui la luna... l'abbiamo vicina.

and the moon is near us here.

Roberto Alagna & Leontina Vaduva, Paris, 1995.

Le dirò con due parole, chi son

Personal

- Born and raised in Petrópolis-RJ;
- Eldest of three kids;
- Married and father of a daughter;
- Mais Querido supporter.

Academic

- BSc in Microbiology & Immunology (UFRJ, 2012);
- PhD Evolutionary Biology (Edinburgh, 2018);
- Post doctoral researcher at ENSP/Fiocruz (2019);
- Lecturer (Assistant Professor) at EMAp since Jan/2020.

E che faccio

Applications of Statistics/Mathematics

Applications in Epidemiology, (Molecular) Biology, Ecology, Psychology, Linguistics, etc.

Applied Statistics

Markov Chain Monte Carlo, Model combination and selection, Statistical Phylogenetics.

Problem I: efficiently utilising available information

Loads of historical data: how to build informative priors?

Let $\boldsymbol{y_0} = (y_{01}, \ldots, y_{0n_0})$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ be **historical** and **current** data, respectively.

Question: how do I build a prior that

- Uses information in $\boldsymbol{y_0}$ efficiently but also
- Does not lead to borrowing too much information when the data sets are not compatible?

Applications: clinical trials, quality control, policy-making.

I got the power

Normalised power prior[1]

$$\tilde{\pi}(\theta, a_0 \mid \boldsymbol{y}_0) = \frac{L(\boldsymbol{y}_0 \mid \theta)^{a_0} \pi(\theta \mid \eta) \pi_A(a_0 \mid \phi)}{c(a_0; \eta, \phi)}$$

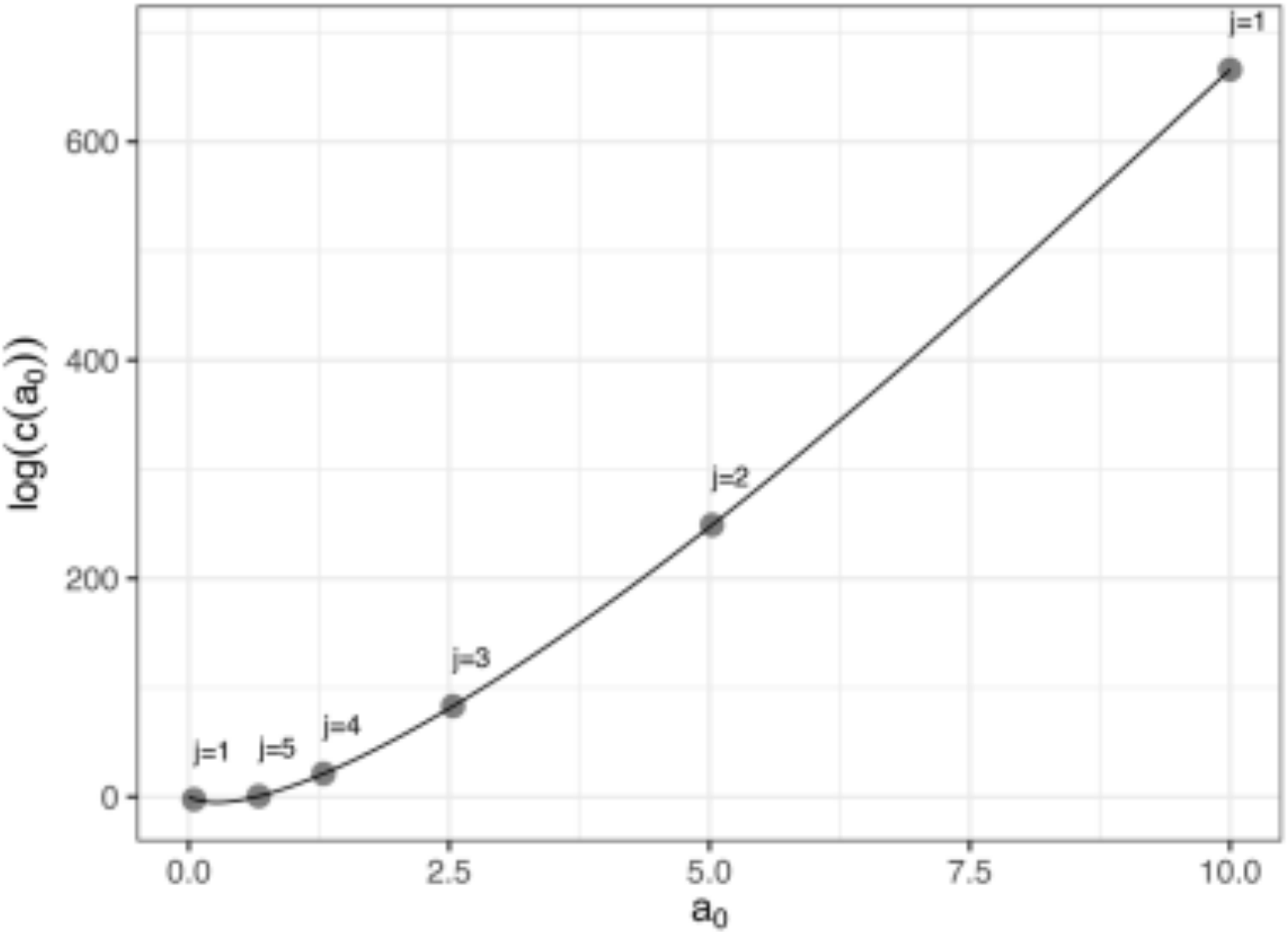
- How pick π_A such that prediction error (say) is minimised?
- How to **efficiently** compute

$$c(a_0; \eta, \phi) = \int_\Theta L(\boldsymbol{y}_0 \mid t)^{a_0} \pi(t \mid \eta) \, d\mu(t)$$

by leveraging its special properties as function of a_0?

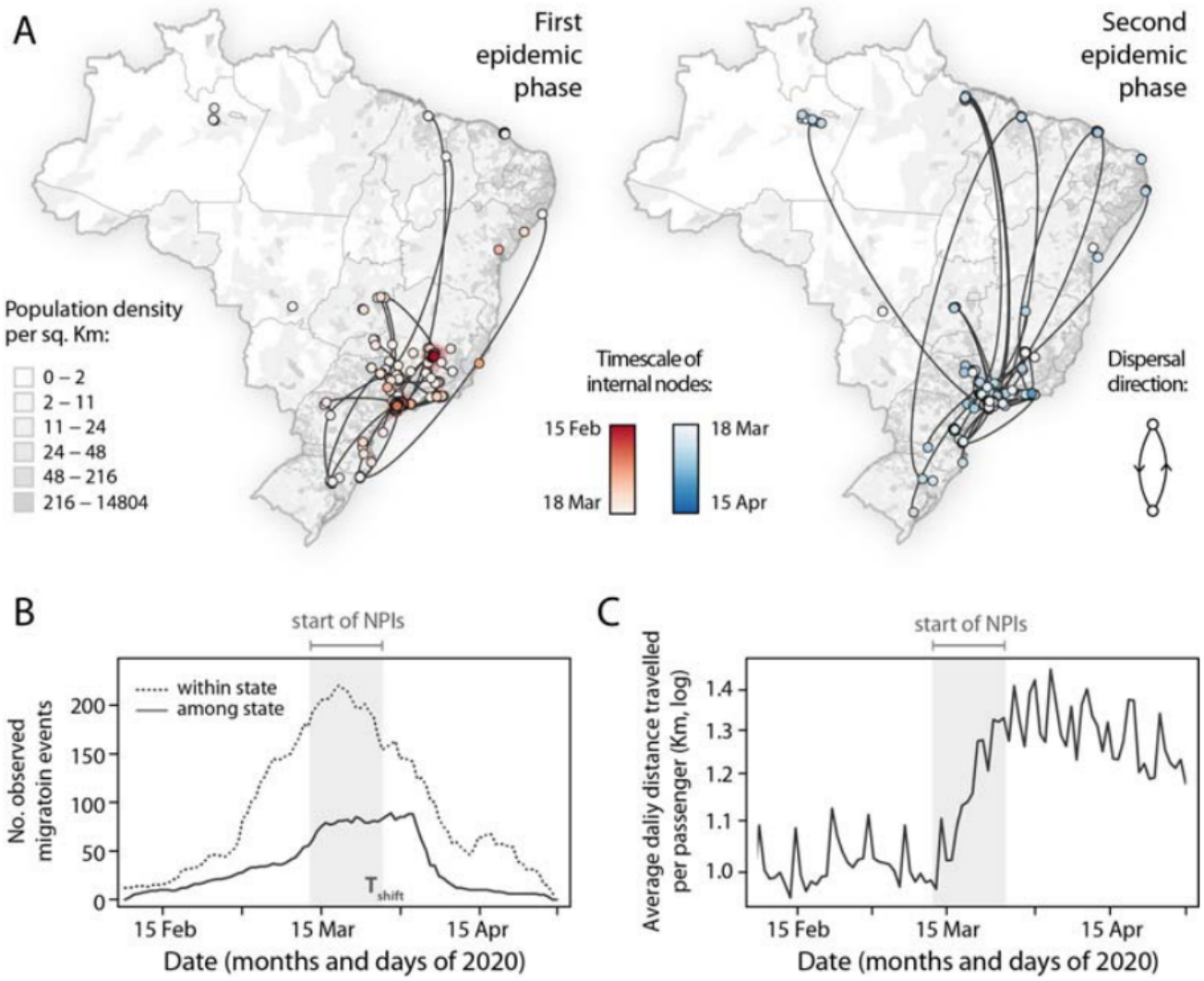
[1] https://doi.org/10.1002/sim.9124

Approximating the normalising constant



Problem II: dealing with huge complex data

Where did this virus come from?[2]



[2] https://doi.org/10.1126/science.abd2161

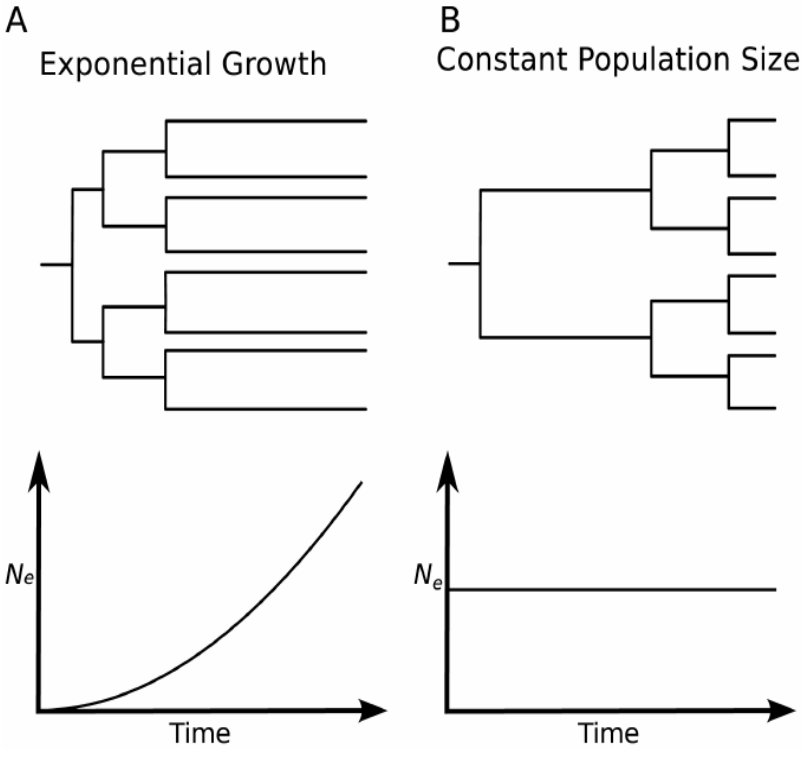
Motivation

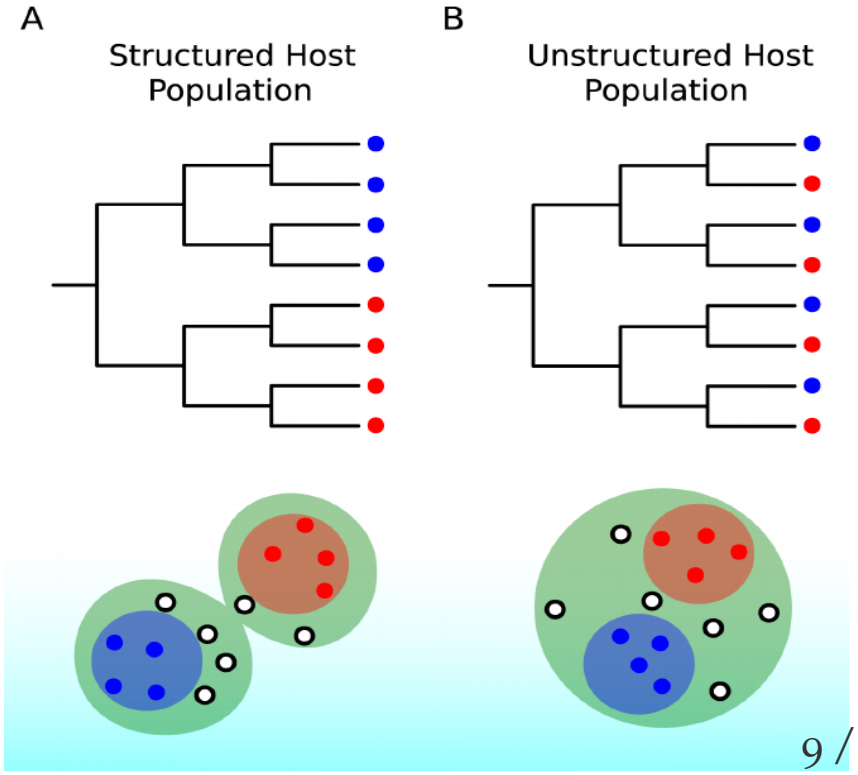
Phylodynamics of fast-evolving viruses

Inferring spatial and temporal dynamics from genomic data:

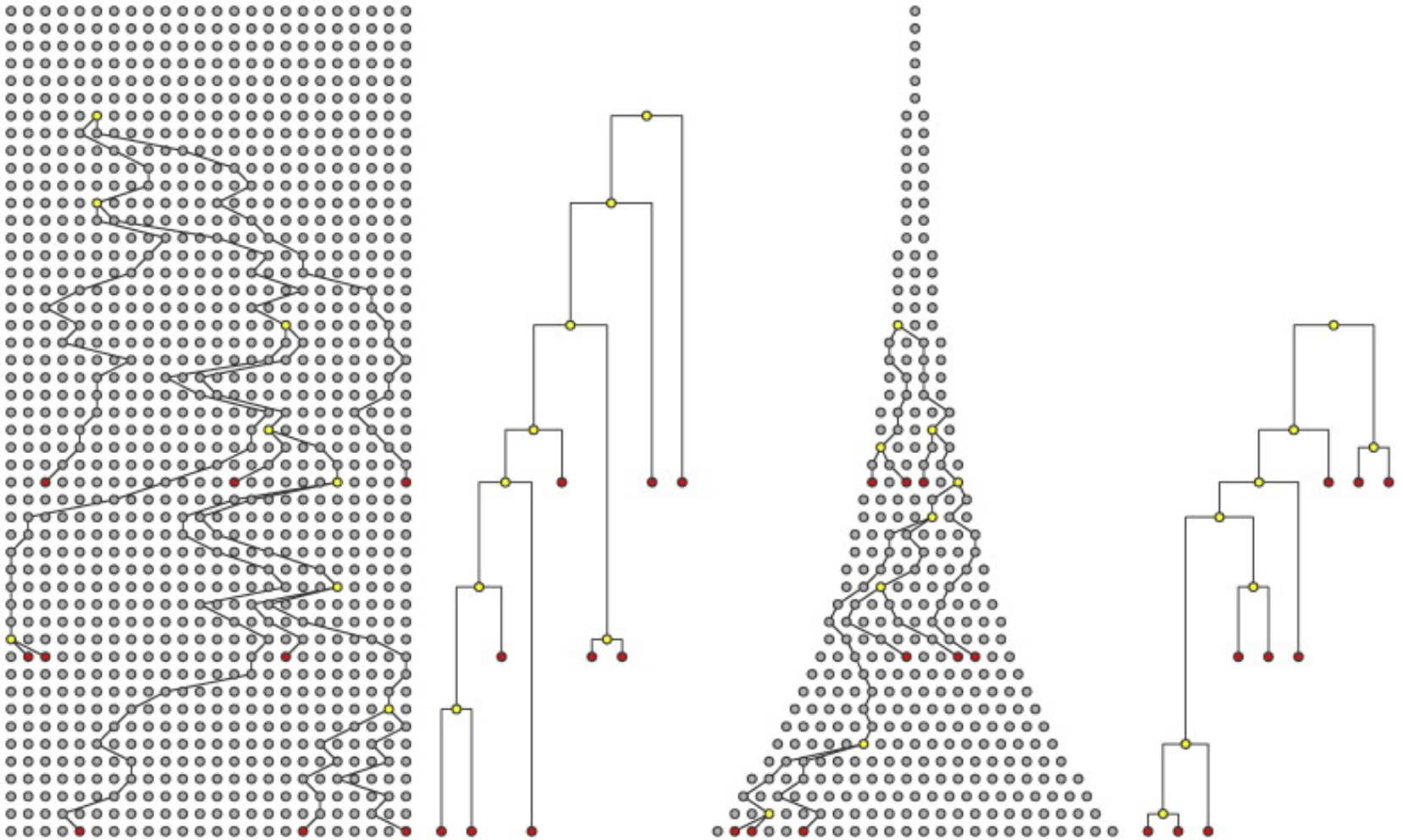
Phylogenies*!

* plus complicated models





Trees and the coalescent



Central object: time-calibrated trees

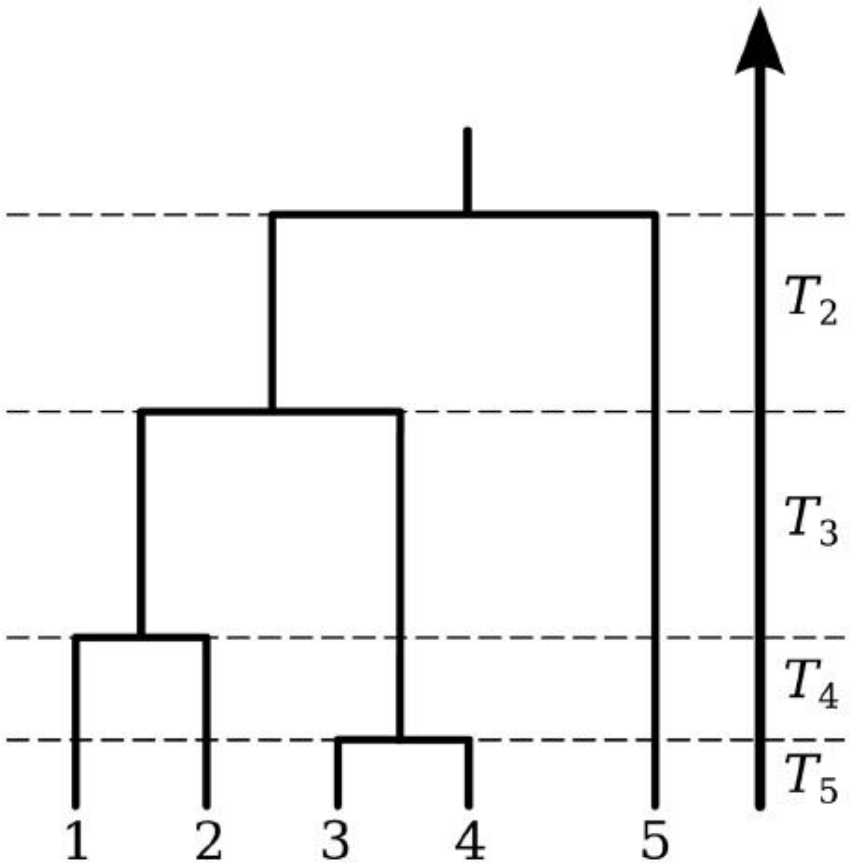


Figure: Figure 4 from Volz et al. (2013).

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:

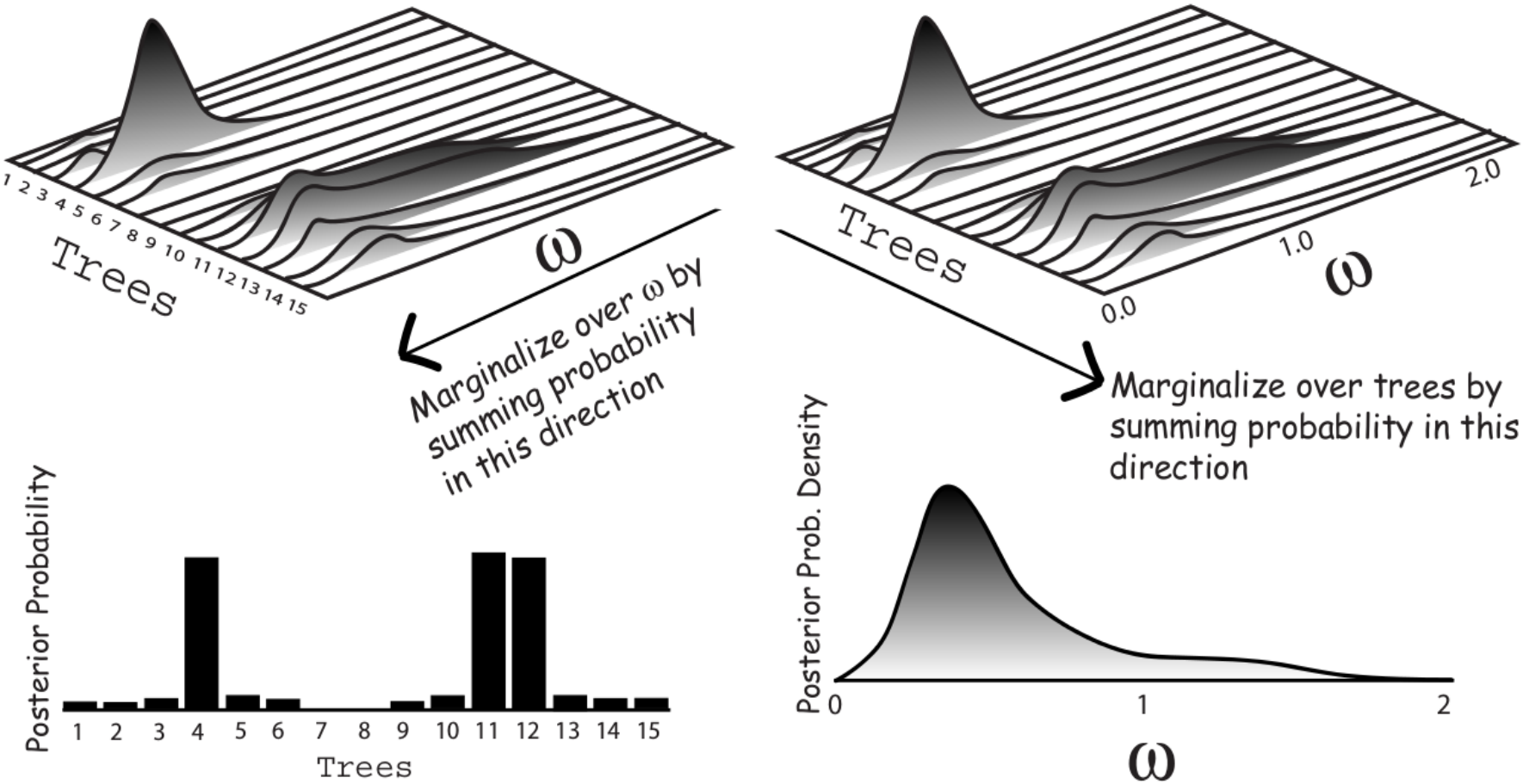
$$\begin{aligned} \mathbb{E}[T_{\text{mrca}}] &= \mathbb{E}[T_n] + \mathbb{E}[T_{n-1}] + \ldots + \mathbb{E}[T_2] \\ &= 1/\lambda_n + 1/\lambda_{n-1} + \ldots + 1/\lambda_2 \\ &= 2N_e \left(1 - \frac{1}{n}\right) \end{aligned}$$

Target

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

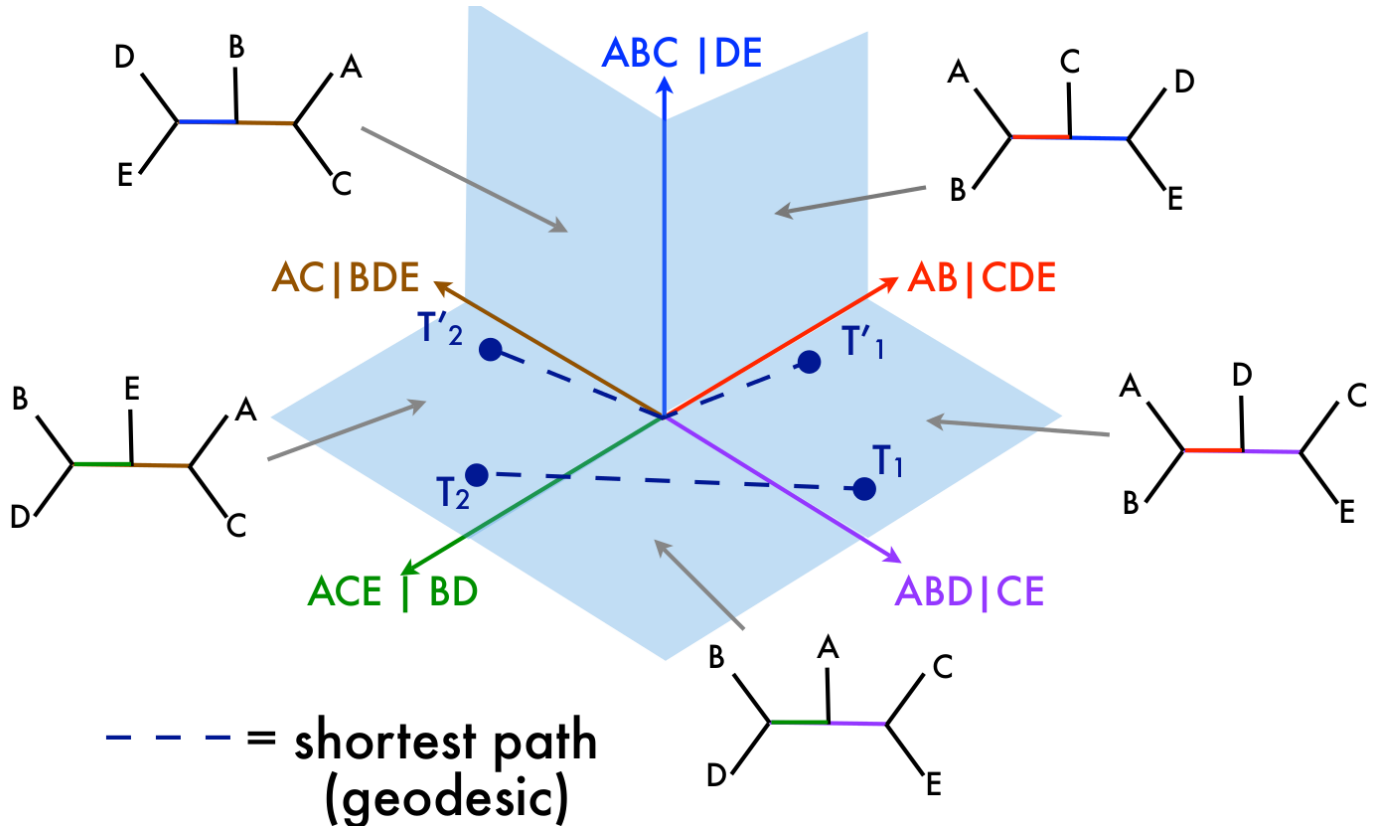
- D: observed sequence (DNA) data;
- $\boldsymbol{T}_n$: set of all binary ranked trees;
- $\boldsymbol{b}_k$: set of branch lengths of $t_k \in \boldsymbol{T}_n$ ($\mathbb{R}_+^{2n-2}$, kind of) ;
- $\boldsymbol{\omega}$: set of parameters of interest such as substitution model parameters, migration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

The end product



This place is weird..

Traversing cubic complexes efficiently[3]



Applications: Molecular Epidemiology, Evolutionary Biology.

[3] https://youtu.be/h9bWRQ6aeKA

(Adaptive) Metropolis-Hastings for trees

General MH setup.

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

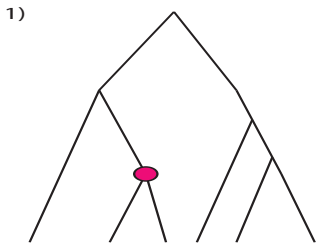
Accepting with probability

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

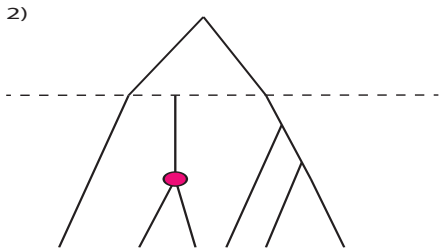
leads to the desired target.

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

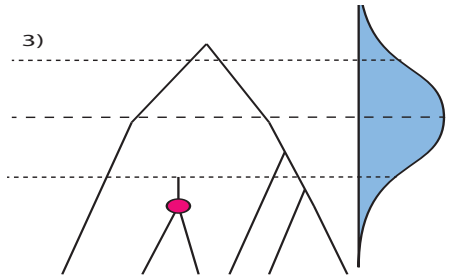
STL – illustration



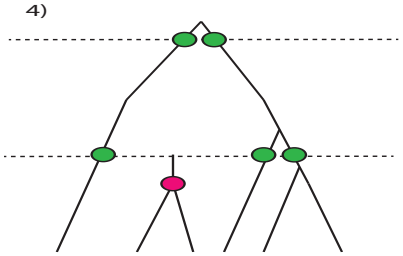
Pick a node



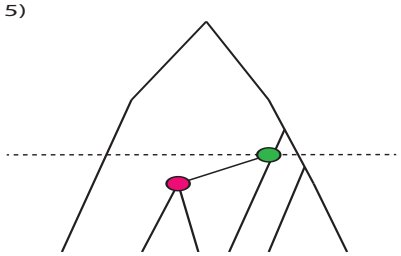
Disconnect its parent



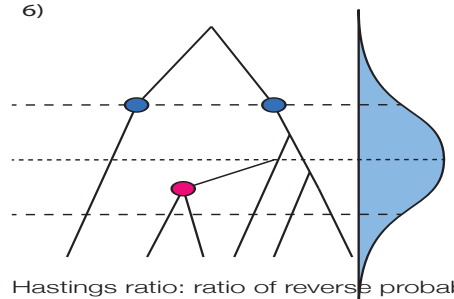
Draw a new height from a normal centred on old height of parent. Also consider the symmetrical height above or below the old height.



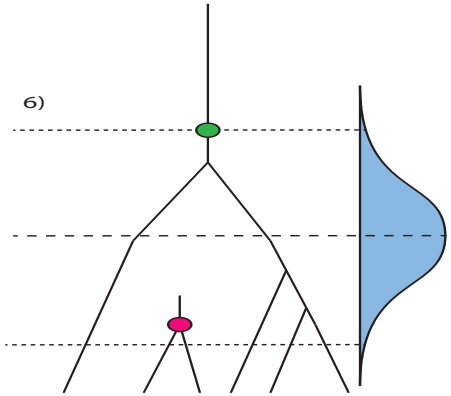
Pick uniformally from branches subtending that height and the symmetrical height above or below (in this case 5).



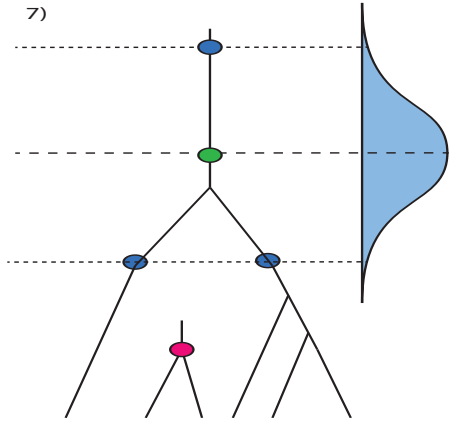
Attach parent to the chosen location.



Hastings ratio: ratio of reverse probability (1 / number of reverse locations, i.e., 1/2) to forwards probability (i.e., 1/5).
Hastings ratio = 5 / 2



There is always at least 1 target location (above the root).



In this case the HR would be 1/3

STL – ergodicity

Carvalho (2019), Chapter 2.

Remark

Assume strictly positive branch lengths. Then SubTreeLeap induces an irreducible Markov chain on $\boldsymbol{T}_n$.

Sketch: Starting at $x \in \boldsymbol{T}_n$, notice there exists $\delta_y^\star > 0$ such that $P\left(x \to y \mid \delta_y^\star\right) > 0$ for any tree $y \in \boldsymbol{T}_n$ in the SPR neighbourhood of x.

Theorem

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

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

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? (Custom) **Tools!**
- Optimal scaling: what's the **optimal** acceptance probability?

Another coat of golden paint?

"Why put another layer of gold paint on the Bentley when you are losing the engine?"
Zinedine Zidane, about Claude Makélélé leaving Real Madrid .

Take home

A light in the dark

Maths gives us methods with provable guarantees

Computational methods are key

Learn to program and learn Computational Statistics[4]

Maths works

Today we've employed: combinatorics, probability theory, basic calculus, pptimisation and classical **and** Bayesian Statistics.

We've got loads to do!

Biomedical statistics is where most of the cool data and problems are.

[4] Here's a place to start:
https://github.com/maxbiostat/Computational_Statistics

THE END