The space of sampled ancestor trees @GSA2016

Alex Gavryushkin

ETH zürich

Joint work with
Alexei Drummond, the University of Auckland, NZ,
Erick Matsen and Chris Whidden,
Fred Hutch Cancer Research Center, Seattle, WA, USA

September 28, 2016

Motivation

- General statistics is at least 5 years ahead of phylostatistics.

Motivation

- General statistics is at least 5 years ahead of phylostatistics.
- The discrete component of tree space is *the* bottleneck for tree search algorithms.

Motivation

- General statistics is at least 5 years ahead of phylostatistics.
- The discrete component of tree space is *the* bottleneck for tree search algorithms.
- What's wrong with trees?

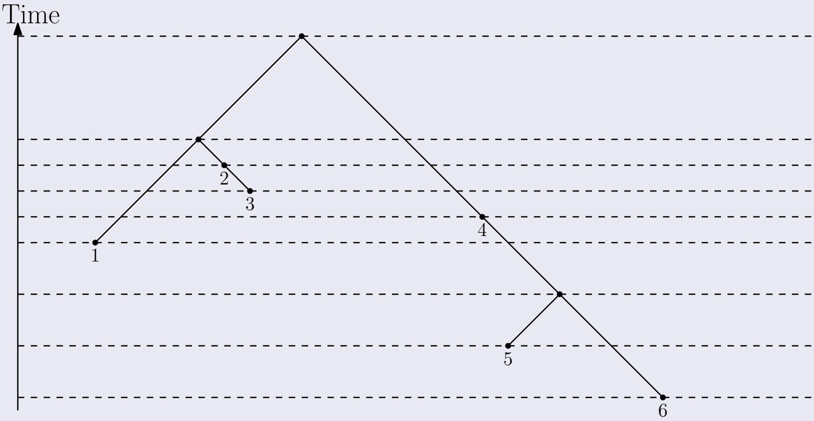
Motivation

- General statistics is at least 5 years ahead of phylostatistics.
- The discrete component of tree space is *the* bottleneck for tree search algorithms.
- What's wrong with trees?

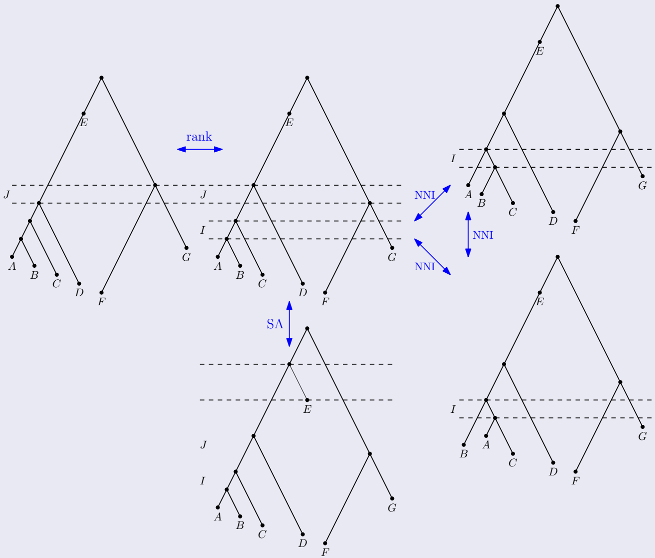
Same as above but with a mortarboard on

- MCMC algorithms
  - Improving efficiency = smart proposals
  - Point estimates AKA posterior summary
- Tree search methods in general
  - Semi-convergence
  - Valleys
  - Terraces

Sampled ancestor tree

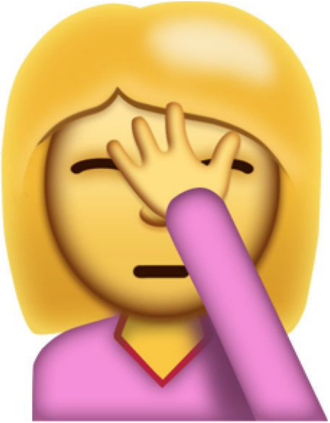


Sampled ancestor tree graph



Graph = Metric space

Graph = Metric space



What’s wrong with the tree space?

Answer

- Over 25 years to solve the complexity problem!

What's wrong with the tree space?

Answer

- Over 25 years to solve the complexity problem!
- Over 7 erroneous “solutions” published on the way!

What's wrong with the tree space?

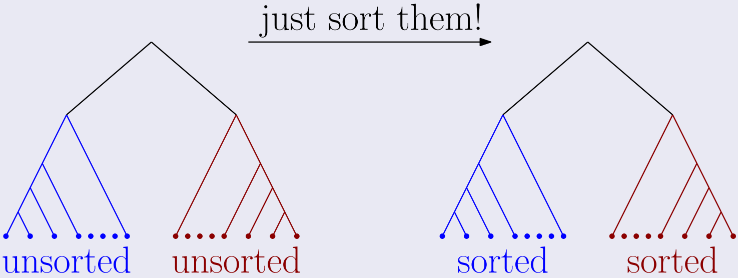
Answer

- Over 25 years to solve the complexity problem!
- Over 7 erroneous "solutions" published on the way!

I'm talking about the NNI graph here.

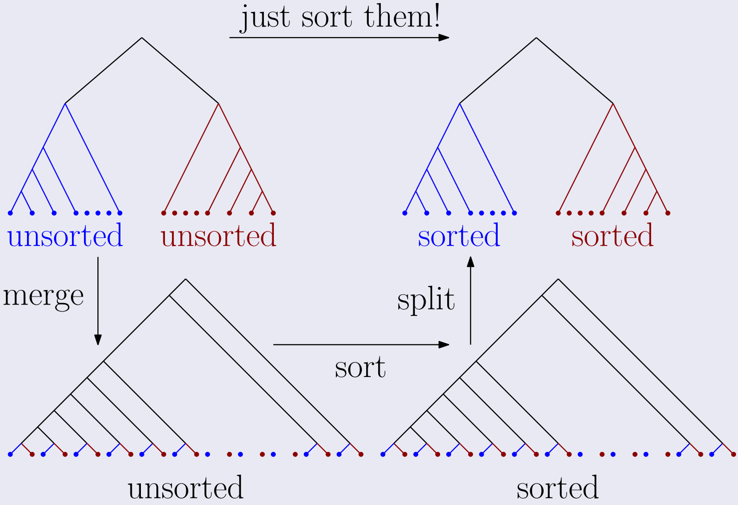
What is actually “wrong”

The Split Theorem



What is actually “wrong”

Merge and sort trick



just sort them!

unsorted unsorted

sorted sorted

merge

split

sort

unsorted

sorted

Good news!

Sampled ancestor trees (the SANNI graph) free from all these (G, Whidden, Matsen. *bioRxiv*, 2016)

- Split Theorem. Tick.
- Merge and sort trick. Tick.

Good news!

Sampled ancestor trees (the SANNI graph) free from all these (G, Whidden, Matsen. *bioRxiv*, 2016)

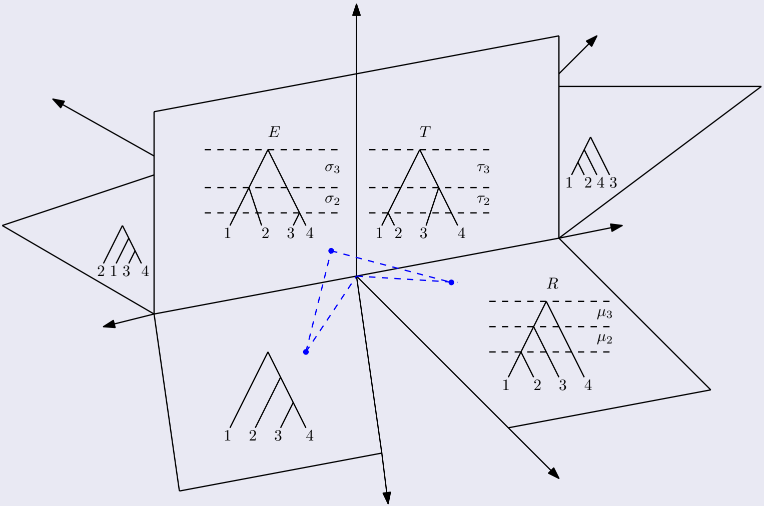
- Split Theorem. Tick.
- Merge and sort trick. Tick.

Even more good news

Efficient approximate algorithm for computing shortest SANNI-paths.

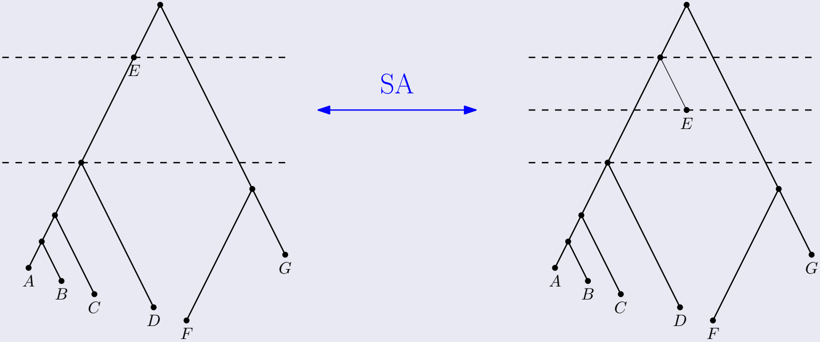
What about branch lengths?

G and Drummond. *JTB*, 2016



Looks like a problem

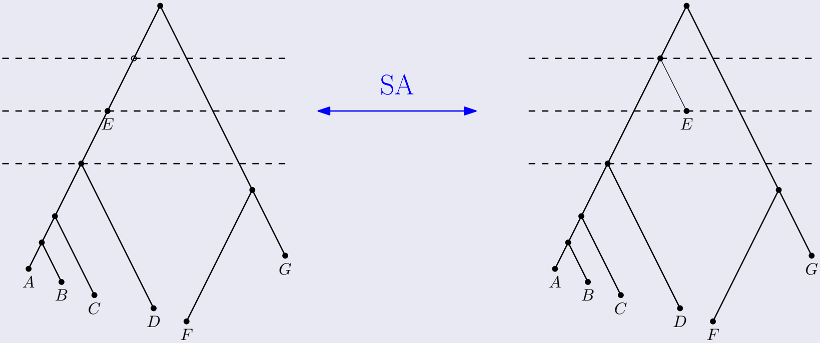
Trees have different dimensions



Branch lengths are fine too!

Stadler (*JTB*, 2010) is cheating* anyway...

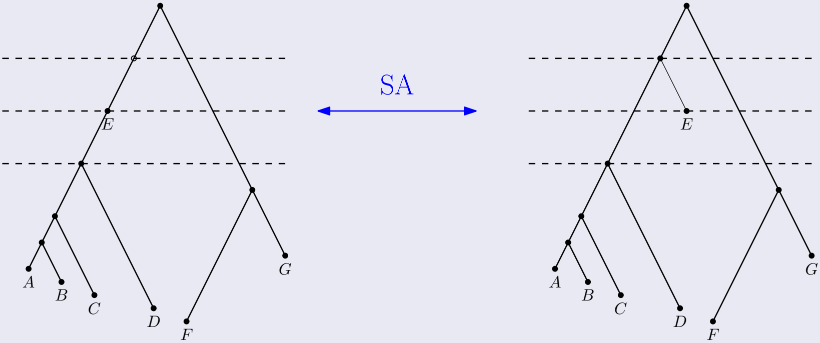
... so we can too: introduce "imaginary" nodes:



Branch lengths are fine too!

Stadler (*JTB*, 2010) is cheating* anyway...

... so we can too: introduce “imaginary” nodes:



*by putting non-zero probability mass onto facets of the space

What we've done

- Introduced the SANNI graph on ranked sampled ancestor trees (to the best of our knowledge)
- Sampled ancestor trees and classical phylogenetic trees have different geometric and algorithmic properties
- Often, geometric and algorithmic results for classical trees do not scale to sampled ancestor trees
- Natural and efficient data structures
- Connections to other areas of math

What we've done

- Introduced the SANNI graph on ranked sampled ancestor trees (to the best of our knowledge)
- Sampled ancestor trees and classical phylogenetic trees have different geometric and algorithmic properties
- Often, geometric and algorithmic results for classical trees do not scale to sampled ancestor trees
- Natural and efficient data structures
- Connections to other areas of math
- Failed to prove that SANNI is NP-hard

References

Li, Tromp, and Zhang
Some Notes on the Nearest Neighbour Interchange Distance.
Computing and Combinatorics, 343–351, 1996.

Dasgupta, He, Jiang, Li, Tromp, and Zhang
On Computing the Nearest Neighbor Interchange Distance
Discrete Mathematical Problems with Medical Applications, Vol. 55, 2000.

Alex Gavryushkin and Alexei Drummond
The space of ultrametric phylogenetic trees
Journal of Theoretical Biology, Vol. 402, 197–208, 2016

Alex Gavryushkin, Chris Whidden, and Frederick A. Matsen IV
Combinatorics of discrete time-trees: algorithmic insights and open problems
bioRxiv, 2016 ← available as a blog post by Matsen

`https://github.com/gavruskin/tauGeodesic`

`https://github.com/gavruskin/tTauCurvature`

Thank you for your attention!

Funding

CENTRE FOR
COMPUTATIONAL
EVOLUTION

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND